

Article

Short-Term Probabilistic Interval Forecasting of Electricity, Cooling, and Heating Loads Using an Improved Osprey Optimization Algorithm-Based TCN-BiLSTM-QR Model

Yiling Zhou ^{1,†}, Xiaohan Xia ^{2,†}, Minhui Wang ³, Lixin Xiao ², Hongrui Wang ^{4,*}  and Ru Zhang ^{5,*}

¹ School of Psychology and Mental Health, North China University of Science and Technology, Tangshan 063210, China; zhouyiling@stu.ncst.edu.cn

² College of Electrical Engineering, North China University of Science and Technology, Tangshan 063210, China; xiaxiaohan@stu.ncst.edu.cn (X.X.); xiaolixin@stu.ncst.edu.cn (L.X.)

³ Faculty of Business and Management, Beijing Normal-Hong Kong Baptist University, Zhuhai 519087, China; u430006193@mail.bnbu.edu.cn

⁴ College of Water Sciences, Beijing Normal University, Beijing 100875, China

⁵ College of Sciences, North China University of Science and Technology, Tangshan 063210, China

* Correspondence: henrywang@bnu.edu.cn (H.W.); zhangru@ncst.edu.cn (R.Z.)

† These authors contributed equally to this work.

Abstract

Integrated Energy Systems (IESs) are characterized by strong nonstationarity, seasonal heterogeneity, and stochastic fluctuations in electric, cooling, and heating loads, making accurate and reliable probabilistic forecasting challenging. To address this issue, this study proposes an Improved Osprey Optimization Algorithm-based TCN-BiLSTM-Quantile Regression (IOOA-TCN-BiLSTM-QR) model for short-term probabilistic forecasting of multi-energy loads. The model integrates Logistic chaotic mapping and Lévy flight to enhance the optimization capability of the Osprey Optimization Algorithm, while TCN and BiLSTM are employed to capture multiscale local features and long-term temporal dependencies, respectively. Quantile regression is further introduced to construct 90% prediction intervals and quantify forecasting uncertainty. Using year-round hourly operational data from a university IES, the proposed model achieves average PICP and PINAW values of 90.50% and 0.1513, respectively, indicating a favorable balance between interval coverage and width. For point forecasting, the model achieves test-set R^2 values of 0.9385, 0.9781, and 0.9731 for electric, cooling, and heating loads, respectively, with corresponding MAPE values of 2.61%, 2.67%, and 2.36%. Comparative and ablation experiments demonstrate the effectiveness of the integrated optimization and temporal feature extraction framework, while cross-dataset validation further confirms its applicability to different multi-energy load conditions. Overall, the proposed model provides accurate point forecasts and 90% prediction intervals with average coverage close to the nominal level, offering quantitative information for reserve capacity allocation, operational scheduling, and risk-aware decision-making in IESs.



Academic Editors: John Abela and Wanquan Liu

Received: 23 August 2026

Revised: 21 September 2026

Accepted: 23 September 2026

Published: 28 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: integrated energy system; electric-cooling-heating loads; probabilistic interval forecasting; IOOA-TCN-BiLSTM; feature contribution analysis

MSC: 68T07

1. Introduction

Traditional power systems rely heavily on electricity generation from fossil fuel combustion, which not only consumes large amounts of finite fossil energy resources but also produces substantial greenhouse gas emissions and pollutants, thereby exacerbating global climate change and ecological degradation and posing serious challenges to green and sustainable development [1]. Against the backdrop of continuously growing global energy demand and the ongoing advancement of the “Dual Carbon” goals, developing a new-type power system dominated by renewable energy has become an important pathway for reducing carbon emissions, improving energy utilization efficiency, and promoting energy structure transformation [2]. With the increasing penetration of renewable energy sources such as wind and photovoltaic power into new-type power systems, their inherent intermittency, volatility, and randomness have introduced new challenges to power supply-scheduled balance and stable system operation. Therefore, coordinated optimization and control, together with accurate load forecasting, are urgently required to enhance system scheduling and operational capabilities. Meanwhile, the Integrated Energy System (IES), through the coordinated complementarity and cascade utilization of heterogeneous energy sources such as electricity, gas, heating, and cooling, can further improve overall energy utilization efficiency [3] and has become an important development direction for the emerging energy system. Accurate integrated energy load forecasting is not only fundamental to maintaining the supply–demand balance of different energy carriers and ensuring safe and stable system operation, but also a critical prerequisite for the optimal scheduling, energy management, and operating cost control of multiple energy subsystems. Previous studies have shown that a 1% reduction in electricity load forecasting accuracy may increase system operating costs by up to USD 10 million [4], further highlighting the significant economic value of accurate load forecasting. However, complex multi-energy coupling relationships exist within integrated energy systems, while the large-scale integration of renewable energy further increases the randomness and uncertainty of load variations. Consequently, traditional forecasting methods based on a single energy carrier or individual variables may struggle to fully characterize the dynamic evolution of complex loads [5]. Therefore, how to effectively extract deep temporal patterns from multi-source load data, adequately characterize the uncertainty induced by renewable energy integration and multi-energy coupling, and achieve accurate and robust multi-load forecasting for integrated energy systems has become a critical issue that urgently needs to be addressed for the optimal operation of new-type power systems and integrated energy systems [6].

In conventional load forecasting studies, load demand is typically predicted independently for a single energy type. Existing approaches can be broadly categorized into two major groups: mathematical statistical methods and artificial intelligence-based methods. Mathematical statistical approaches mainly include regression analysis [7] and time-series methods [8]. In recent years, machine learning has been widely adopted because of its capability to model complex nonlinear relationships between explanatory variables and forecasting targets without requiring the construction of complicated mathematical models [9]. Common machine learning models include Random Forest (RF) [10], Support Vector Machine (SVM) [11], and Gated Recurrent Unit (GRU) [12]. Luo et al. [13] identified similar load days using AHC and integrated CNN-BiLSTM-Attention with XGBoost to construct a Stacking ensemble model, thereby improving short-term electricity load forecasting accuracy. Peng et al. [14] introduced a multi-scale spatiotemporal graph convolutional network and Transformer to fully exploit short-term spatiotemporal characteristics and long-term temporal features of different nodes within the power system topology, based on which optimal input features were constructed. Liu et al. [15] addressed the insufficient forecasting stability and generalization caused by complex and highly variable

load patterns in large public buildings by developing a hybrid forecasting model integrating MS-DTW, FIE, ITCN, and Transformer. Through load pattern clustering and feature importance evaluation, scenario-specific forecasting models were constructed, achieving better performance than benchmark models across different building types and forecasting horizons of 6, 24, and 48 h. In particular, the RMSE for the 6-h forecasting task was reduced by 39.2–47.1% compared with TCN-Transformer. However, these studies have primarily focused on forecasting a single type of load and therefore have limited ability to deeply exploit the complex coupling relationships among multiple loads in Integrated Energy Systems (IES), while the generalization capability of the developed models remains relatively limited [5,16]. With the ongoing energy transition, IESs have become an important pathway toward more efficient energy utilization and energy structure transformation [5]. By coupling multiple energy networks and enabling interconnection and complementarity among different energy carriers through energy conversion devices, IESs establish complex coupling relationships among different types of loads within the system [17]. Consequently, load forecasting research has gradually shifted from single-load forecasting toward multi-load forecasting. Tan et al. [18] proposed a multi-energy load forecasting method based on SCA, LPF, and MTL-LSTM to address the randomness and coupling characteristics of multi-energy loads in IESs. Specifically, SCA was employed to identify key features, LPF was used to characterize the interactions among different loads, and MTL-LSTM was adopted to extract dynamic coupling information among multiple load types, enabling coordinated forecasting of electricity, cooling, heating, and total loads, with average forecasting accuracies of 97.18% and 97.85%, respectively. Zhuang et al. [19] developed a multi-energy load forecasting method based on a Multi-Scale Spatiotemporal Graph Neural Network (MS-STGNN) to address the difficulties in characterizing complex correlations among multi-energy loads and the information loss caused by extracting features at a single temporal scale. The proposed framework exploited the relationships among multi-energy loads and auxiliary features through continuous graph learning and graph convolution modules, while feature attention and multi-scale convolution modules were employed to capture temporal characteristics at different time scales, thereby improving the forecasting performance and generalization capability for integrated energy loads. Furthermore, historical multivariate load data in IESs generally exhibit strong volatility and non-stationarity and are influenced by external factors such as seasons and holidays [20]. These characteristics further increase the difficulty of load forecasting and may lead to degraded forecasting accuracy.

Building on the aforementioned studies, existing research on multi-load forecasting in integrated energy systems has gradually evolved from single-load forecasting toward coordinated multi-energy forecasting, with methods such as multi-task learning, graph neural networks, and multi-scale feature extraction being employed to enhance the capability to capture correlations among different types of loads. However, most existing studies still focus primarily on point forecasting, which represents the future load level using a single predicted value and therefore has limited ability to characterize the uncertainties arising from renewable energy fluctuations, changes in user energy consumption behavior, and multi-energy coupling processes [5,6,20,21]. When actual loads deviate substantially from the predicted values due to unexpected factors, point forecasts cannot effectively provide the potential range of future load variations, thereby limiting their practical value in risk assessment, reserve capacity allocation, and optimal scheduling [22]. Therefore, while improving the accuracy of multi-energy load forecasting, it is also important to develop probabilistic interval forecasting approaches capable of characterizing forecasting uncertainty [23]. In recent years, interval forecasting has been increasingly applied to electricity load forecasting, providing decision-makers with the potential range of future

load variations by generating upper and lower prediction bounds at different confidence levels. Compared with conventional point forecasting, interval forecasting can provide a more comprehensive representation of load uncertainty and richer information for the safe operation and risk-aware decision-making of energy systems. Among existing approaches, Quantile Regression (QR) is particularly suitable for probabilistic interval forecasting of electricity loads because it does not rely on specific assumptions regarding the underlying data distribution and can directly learn the conditional distribution characteristics of loads at different quantiles [24]. By integrating QR with deep learning models, prediction results at different quantile levels can be obtained while extracting complex temporal characteristics of load sequences, thereby enabling the construction of load prediction intervals with different confidence levels. On the other hand, electricity, cooling, and heating loads in integrated energy systems exhibit distinct variation patterns and temporal scales, with load sequences simultaneously containing complex temporal characteristics such as local fluctuations, periodic variations, and long-term dependencies [25,26]. Conventional single-network architectures are often constrained by their receptive fields or their ability to model temporal dependencies during feature extraction. The Temporal Convolutional Network (TCN) employs causal and dilated convolutions to enlarge the receptive field while preserving temporal causality, enabling effective extraction of local and long-term temporal features at different scales [27]. The Bidirectional Long Short-Term Memory (BiLSTM) network, in contrast, learns contextual information from both forward and backward directions, thereby enhancing its ability to represent complex temporal dependencies [28]. Therefore, integrating TCN and BiLSTM can potentially exploit their complementary strengths in multi-scale feature extraction and temporal dependency modeling, providing more effective feature representations for multi-energy load forecasting. Furthermore, the forecasting performance of deep learning models depends strongly on the appropriate configuration of their network structures and hyperparameters, such as the number of convolutional layers, number of convolutional kernels, dilation factors, hidden-layer dimensions, and learning rate [29]. Determining these parameters through manual trial and error is not only inefficient but also susceptible to subjective experience, making it difficult to identify an optimal parameter combination for complex multi-energy load data. Therefore, introducing intelligent optimization algorithms with strong global search capabilities to optimize the hyperparameters of TCN-BiLSTM models can help improve forecasting performance and stability [30]. The Improved Osprey Optimization Algorithm (IOOA) enhances the global exploration and local exploitation capabilities of the population search process, providing a more effective optimization strategy for complex model parameter spaces and thereby reducing the risk of the model becoming trapped in local optima [31].

Based on the above considerations, this study proposes a short-term probabilistic interval forecasting method for electricity, cooling, and heating loads based on an Improved Osprey Optimization Algorithm (IOOA)-optimized TCN-BiLSTM-QR framework. First, operational data from a university integrated energy system are collected, and the electricity, cooling, and heating loads together with relevant influencing factors are preprocessed and organized to construct the input features. Second, a TCN-BiLSTM forecasting model is developed, where TCN is employed to extract multi-scale local temporal patterns from the load sequences, while BiLSTM further captures long-term temporal dependencies. Subsequently, IOOA is introduced to adaptively optimize the key hyperparameters of the forecasting model, and Quantile Regression (QR) is incorporated to construct load prediction intervals at different confidence levels, thereby quantifying forecasting uncertainty. Furthermore, the probabilistic interval forecasting characteristics of electricity, cooling, and heating loads are investigated across the four seasons and different representative time

periods, with particular emphasis on their seasonal and intraday variations. Multi-model comparisons and ablation experiments are then conducted to examine the effectiveness of the proposed model and the contributions of its core components. In addition, an independent load dataset is employed to evaluate the applicability and generalization capability of the proposed framework under different data scenarios. Finally, feature contribution analysis is performed to identify the relative influence of different input variables on electricity, cooling, and heating load forecasting, thereby improving the interpretability of the forecasting process. Through these investigations, this study establishes an integrated framework encompassing model development, point forecasting, probabilistic interval forecasting, seasonal and intraday analysis, applicability assessment, and feature contribution interpretation, providing methodological support for load forecasting and operational decision-making in integrated energy systems.

The contributions of this study are summarized as follows. First, an integrated IOOA-TCN-BiLSTM-QR framework is developed to simultaneously improve forecasting accuracy and uncertainty characterization for multiple energy loads. Second, the proposed probabilistic forecasting framework is systematically evaluated under different seasons and representative time periods, allowing the distinct seasonal and intraday characteristics of electricity, cooling, and heating loads to be explicitly examined. Third, comprehensive comparative and ablation experiments are conducted to quantify the effectiveness of the proposed optimization strategy and the individual contributions of the core model components. Fourth, an independent load dataset is introduced to further assess the applicability and generalization capability of the proposed framework under different data conditions. Finally, feature contribution analysis is performed to improve the interpretability of the forecasting results by identifying the relative influence of different input variables on the three types of energy loads. These contributions are important because they extend multi-energy load forecasting from deterministic point prediction toward more informative probabilistic forecasting, providing quantitative uncertainty information that can support reserve-capacity allocation, operational scheduling, and risk-aware decision-making in integrated energy systems.

2. Theoretical Foundation

2.1. Improved Osprey Optimization Algorithm (IOOA)

2.1.1. Traditional Osprey Optimization Algorithm (OOA)

The Osprey Optimization Algorithm (OOA) is a swarm intelligence optimization algorithm inspired by the predatory behavior of ospreys [32]. In this algorithm, the position of each osprey is encoded as a candidate solution to the optimization problem, and the solution space is iteratively explored by simulating the processes of prey searching, capturing, and transferring. The optimization process mainly consists of two stages: global exploration and local exploitation. First, ospreys adjust their search positions according to the locations of their prey, thereby expanding the search range and enhancing the global exploration capability of the population. Subsequently, by simulating the behavior of ospreys capturing prey and transferring it to a suitable location, the search space around the current promising solutions is further exploited, thereby improving the capability for fine-grained local search. Through the coordinated interaction between global exploration and local exploitation, OOA can simultaneously expand the search range and strengthen local optimization capability. It has demonstrated good search efficiency and convergence stability in complex nonlinear and high-dimensional optimization problems and is therefore suitable for complex parameter optimization and large-scale engineering optimization tasks [33].

In the Osprey Optimization Algorithm, the population consists of N ospreys, with the position of each osprey representing a candidate solution. The candidate solutions are represented by an $N \times m$ -dimensional matrix X , as shown in Equation (1). Each row of the matrix corresponds to the position of an individual osprey, representing a potential solution to the optimization problem.

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{bmatrix}_{N \times m} = \begin{bmatrix} x_{1,1} & \dots & x_{1,j} & \dots & x_{1,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & \dots & x_{i,j} & \dots & x_{i,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N,1} & \dots & x_{N,j} & \dots & x_{N,m} \end{bmatrix}_{N \times m} \quad (1)$$

where N represents the number of ospreys; m represents the number of dimensions in the problem; X_i represents the i -th osprey; and $x_{i,j}$ represents the j -th dimension, i.e., the problem variable.

The population is then initialized according to Equation (2):

$$x_{i,j} = b_j^l + r_{i,j} \cdot (b_j^u - b_j^l), i = 1, 2, \dots, N; j = 1, 2, \dots, m \quad (2)$$

where b_j^l and b_j^u are the upper and lower bounds of the j -th variable, respectively; $r_{i,j}$ is a random number in the interval $[0, 1]$.

The fitness of the objective function is determined according to Equation (3):

$$F = \begin{bmatrix} F_1 \\ \vdots \\ F_i \\ \vdots \\ F_N \end{bmatrix}_{N \times 1} = \begin{bmatrix} F(X_1) \\ \vdots \\ F(X_i) \\ \vdots \\ F(X_N) \end{bmatrix}_{N \times 1} \quad (3)$$

where F represents the objective function value; F_i represents the objective function value for the i -th osprey.

When constructing the OOA, the osprey positions corresponding to better objective function values in the search space are regarded as the locations of fish hidden underwater. The position of each osprey in the search space can be mathematically described by Equation (4):

$$F_{pi} = \{X_k \mid k \in \{1, 2, \dots, N\} \wedge F_k < F_i\} \cup \{X_{best}\} \quad (4)$$

where F_{pi} represents the set of positions of the i -th osprey; X_{best} represents the optimal osprey's position; F_k $x_{i,j}^{P1} = x_{i,j} + r_{i,j} \cdot (S_{i,j} - I_{i,j} \cdot x_{i,j})$ and F_i represent the target positions of the k -th and i -th ospreys, respectively.

The osprey randomly selects a target prey based on the positional information of prey within the population and simulates the predatory process of gradually approaching the selected target. During this process, the positions of individual ospreys are updated according to Equations (5) and (6), thereby generating new candidate solutions. The objective function value corresponding to the new position is then calculated and compared with the fitness of the current osprey position. If the new position yields a better objective function value, it replaces the original position; otherwise, the current solution is retained.

$$x_{i,j}^{P1} = x_{i,j} + r_{i,j} \cdot (S_{i,j} - I_{i,j} \cdot x_{i,j}) \quad (5)$$

$$x_{i,j}^{P1} = \begin{cases} x_{i,j}^{P1}, b_j^l \leq x_{i,j}^{P1} \leq b_j^u \\ b_j^l, x_{i,j}^{P1} < b_j^l \\ b_j^u, x_{i,j}^{P1} > b_j^u \end{cases} \quad (6)$$

where $x_{i,j}^{P1}$ represents the new position in the j -th dimension of the i -th osprey during the first stage; $S_{i,j}$ is a random number from the interval $[0, 1]$; $I_{i,j}$ is a random number from the set $\{1, 2\}$.

If a better new position is found, the osprey's position is updated according to Equation (7) [34]:

$$X_i = \begin{cases} X_i^{P1}, F_i^{P1} < F_i \\ X_i, Others \end{cases} \quad (7)$$

where F_i^{P1} represents the objective function value at the new location.

After capturing the prey, the osprey carries it to a suitable location for feeding. The second stage of the OOA simulates this predatory behavior to perform local exploitation around the current promising solutions. In this stage, each osprey adjusts its position within the search space according to a randomly generated feeding location. By progressively narrowing the search range, the algorithm further explores potentially promising solutions, thereby enhancing its local exploitation capability and convergence accuracy. Specifically, for each osprey, a candidate feeding position is randomly generated within the search space, and the corresponding objective function value is calculated. This value is then compared with that of the osprey's current position. If the candidate position yields better fitness, it is adopted to update the osprey's position; otherwise, the original position is retained. The position update process is expressed as follows:

$$x_{i,j}^{P2} = x_{i,j} + \frac{b_j^l + r \cdot (b_j^u - b_j^l)}{t}, i = 1, 2, \dots, m; t = 1, 2, \dots, T \quad (8)$$

$$x_{i,j}^{P2} = \begin{cases} x_{i,j}^{P2}, b_j^l \leq x_{i,j}^{P2} \leq b_j^u \\ b_j^l, x_{i,j}^{P2} < b_j^l \\ b_j^u, x_{i,j}^{P2} > b_j^u \end{cases} \quad (9)$$

where $x_{i,j}^{P2}$ represents the new position in the j -th dimension of the i -th osprey during the second phase; r is a random number within the interval $[0, 1]$; t is the current iteration number; and T is the maximum number of iterations.

If a better position is found, the previous osprey position is replaced according to Equation (10) [34]:

$$X_i = \begin{cases} X_i^{P2}, F_i^{P2} < F_i \\ X_i, Others \end{cases} \quad (10)$$

where F_i^{P2} represents the corresponding fitness value.

2.1.2. Improvement Methods

To address the limitations of the traditional Osprey Optimization Algorithm (OOA) in complex optimization problems, including uneven initial population distribution, insufficient global search capability, slow convergence, and a high tendency to become trapped in local optima [35], this study improves OOA from two perspectives: enhancing population diversity and expanding the search range [34]. First, a Logistic chaotic map is introduced to generate the initial population with favorable ergodicity and randomness [36], thereby improving the quality of population initialization and enhancing the global exploration capability during the early stage of optimization. Second, a Lévy flight strategy is incorporated to construct a stochastic search mechanism characterized by a combination of

long- and short-step movements, enabling some individuals to escape from the current local region and explore a broader search space, thereby reducing the probability of being trapped in local optima [37]. Through the synergistic effect of these two strategies, the population diversity, global exploration capability, local exploitation capability, and convergence performance of OOA are further enhanced, providing a more stable and efficient optimization foundation for the subsequent hyperparameter optimization of the proposed model. The specific improvement procedures are described as follows:

(1) Logistic Chaotic Map

The Logistic chaotic map is incorporated into the Osprey Optimization Algorithm (OOA) in this study to generate a chaotic sequence using the Logistic Equation:

$$r_{(i+1,j)} = c \cdot r_{i,j} \cdot (1 - r_{i,j}) \quad (11)$$

The parameter c is fixed to 4.0 in all experiments, which enables fully chaotic and uniform ergodic distribution of the initialized population. As c gradually approaches 4, the Logistic map exhibits increasingly pronounced chaotic characteristics. When $c = 4$, the system reaches a fully chaotic state, generating a sequence with favorable ergodicity and a relatively uniform distribution. Based on the aforementioned chaotic mapping mechanism, Equation (12) is employed to initialize the positions of the OOA population individuals, thereby improving the uniformity of the initial population distribution in the search space and enhancing population diversity and global exploration capability.

$$X_{i,j} = b^l + (b^u - b^l) \cdot r_{i,j} \quad (12)$$

where $X_{i,j}$ represents the position of the i -th osprey along the j -th dimension, and $r_{i,j}$ denotes the normalized random number sequence generated by the logistic chaotic map.

The Logistic chaotic sequence generated by Equation (12) is mapped into the search space of the optimization problem to achieve a more uniformly distributed initialization of the Osprey population. Owing to the favorable ergodicity, randomness, and nonlinear characteristics of the Logistic chaotic map, the resulting chaotic sequence enables the initial population to cover the entire search space more uniformly, reducing issues such as overlapping individual positions and uneven population distribution, thereby improving population diversity and the quality of the initial solutions. On this basis, the improved OOA can explore a broader search range during the early iterations, enhancing its global exploration capability and reducing the risk of premature convergence to local optima caused by an inappropriate initial population distribution. Meanwhile, a more uniform distribution of initial solutions helps reduce ineffective searches and improve the optimization efficiency and convergence performance of the algorithm, providing a favorable initial condition for subsequently obtaining the global optimum.

(2) Lévy Flight Strategy

During the global exploration phase of the OOA, the parameter r_1 primarily regulates the search step size and search range of the osprey individuals. However, a fixed or constrained search range may limit the exploration of the population, making it difficult for the algorithm to effectively escape from local optima and consequently weakening its global optimization capability. To address this issue, the Lévy flight strategy is introduced to improve the position update mechanism of the osprey individuals. Lévy flight is characterized by variable step lengths, combining short-distance searches with occasional long-distance jumps, which enables the search range of individuals to be expanded while preserving local search capability [38]. By incorporating Lévy flight into the global exploration process of OOA, the positional perturbations and diversity of search trajectories among population

individuals can be increased, allowing some individuals to escape from the current local region and explore new potentially promising regions. Consequently, the global exploration capability of the algorithm is enhanced, while the probability of becoming trapped in local optima is reduced. The mathematical formulation of the Lévy flight is given as follows [39]:

$$L(1) = \frac{u}{|v|^{\frac{1}{\beta}}} \quad (13)$$

$$\delta_u = \left[\frac{\Gamma(1 + \beta) \cdot \sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) \cdot \beta \cdot 2^{\left(\frac{\beta-1}{2}\right)}} \right]^{\frac{1}{\beta}}, u \sim (0, \delta_u^2); v \sim (0, \delta_v^2) \quad (14)$$

where Γ is the gamma function; $\beta = 1.5$; u and v follow Gaussian distributions with mean 0, respectively; δ_u^2 and δ_v^2 are the variances.

In the improved Osprey Optimization Algorithm, the Lévy flight strategy is employed to replace the conventional random number r_1 , and the corresponding update equation is given in Equation (15):

$$X_{i,j}^{t+1} = X_{i,j}^t + L(1) \times (T_{i,j} - I_{i,j} \times x_{i,j}) \quad (15)$$

where $T_{i,j}$ is a scaling factor used to control the step size of the Levy walk; $I_{i,j}$ is an indicator function used to determine whether to update the current dimension.

In the improved IOOA Lévy flight update rule, the scaling factor $T_{i,j}$ and indicator function $I_{i,j}$ are explicitly defined as follows:

The scale factor controls the step magnitude of Lévy mutation:

$$T_{i,j} = 0.01 \cdot \frac{t}{T_{\max}} \quad (16)$$

The binary indicator $I_{i,j}$ determines whether the Lévy flight update is activated for the j -th dimension of the i -th osprey:

$$I_{i,j} = \begin{cases} 1, & \text{rand}(\cdot) < 0.5 \\ 0, & \text{othersies} \end{cases} \quad (17)$$

Only dimensions with $I_{i,j}$ are updated via Lévy flight, which balances local fine search and global exploration.

After introducing the Lévy flight strategy, osprey individuals can exploit the heavy-tailed distribution to generate random search steps with different scales, effectively combining local exploration with long-distance jumps in the search space. When the population becomes trapped in a local optimum region, larger random steps can drive some individuals to escape the current search region and explore new potentially promising solutions, thereby expanding the coverage of the solution space and reducing the risk of premature convergence. In regions near promising solutions, smaller step sizes can be used for fine-grained searches, improving local exploitation capability and optimization accuracy. Therefore, the Lévy flight strategy can effectively enhance the search flexibility, population diversity, and global optimization capability of OOA, thereby improving its convergence performance in complex high-dimensional optimization problems.

In summary, the traditional OOA is improved by incorporating the Logistic chaotic map and Lévy flight strategy. Specifically, the Logistic chaotic map is employed to optimize the population initialization process. By exploiting its favorable ergodicity and uniformity, the distribution of initial individuals in the search space is improved, thereby enhancing population diversity and global exploration capability. The Lévy flight strategy, on the

other hand, is applied during the iterative search process to generate multi-scale random step sizes through a heavy-tailed distribution, enhancing the ability of the population to escape local optima while balancing global exploration and local exploitation. These two strategies improve the optimization performance of OOA from the perspectives of initial population quality and search-step mechanism, respectively, thereby improving the convergence speed, solution accuracy, and stability of the algorithm. The improved algorithm is referred to as the Improved Osprey Optimization Algorithm (IOOA), and its specific optimization procedure is shown in Figure 1.

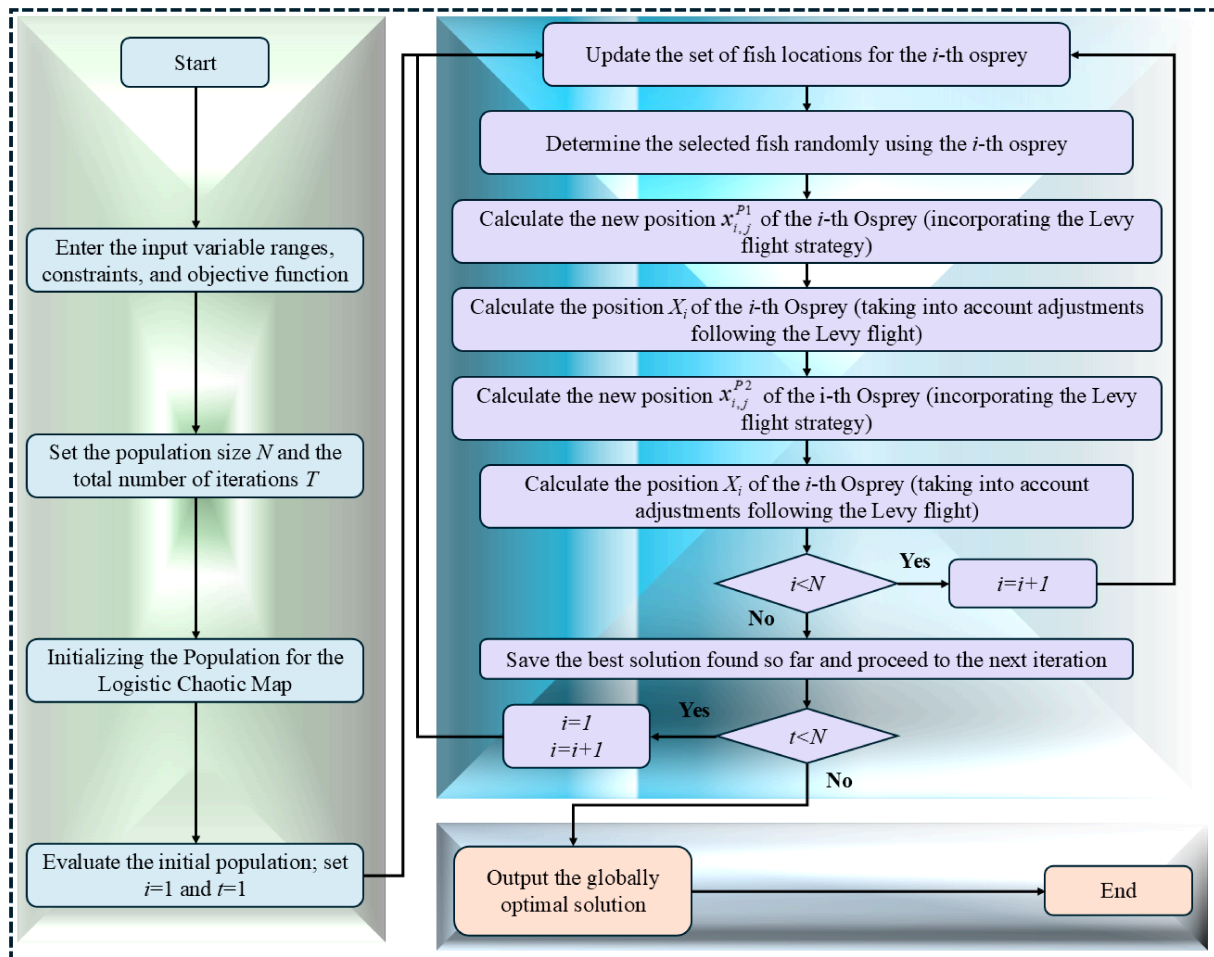


Figure 1. IOOA Flowchart.

2.2. Temporal Convolutional Network (TCN) Algorithm

The Temporal Convolutional Network (TCN) is a deep convolutional neural network specifically designed for time-series data. Its core advantage lies in its ability to effectively extract temporal features at different time scales by employing causal convolutions and dilated convolutions [40]. Compared with conventional recurrent neural networks, TCN does not require recursive computation step by step along the temporal dimension, resulting in higher parallel computing efficiency. Meanwhile, by enlarging the convolutional receptive field, TCN can capture long-range historical dependencies over extended time horizons. Therefore, TCN is particularly suitable for short-term electricity load forecasting tasks characterized by strong temporal correlations and periodic patterns.

As illustrated in Figure 2, TCN primarily consists of multiple cascaded residual blocks. Each residual block mainly comprises dilated causal convolution, Weight Normalization, the ReLU activation function, and Dropout, while a residual connection is employed to fuse the input of the module with the features obtained after nonlinear transformation.

When the input and output dimensions are inconsistent, a 1×1 convolution is applied to transform the input dimensions, enabling element-wise addition with the residual branch. This architecture not only enhances feature extraction capability but also alleviates potential gradient vanishing and network degradation problems during the training of deep networks [41].

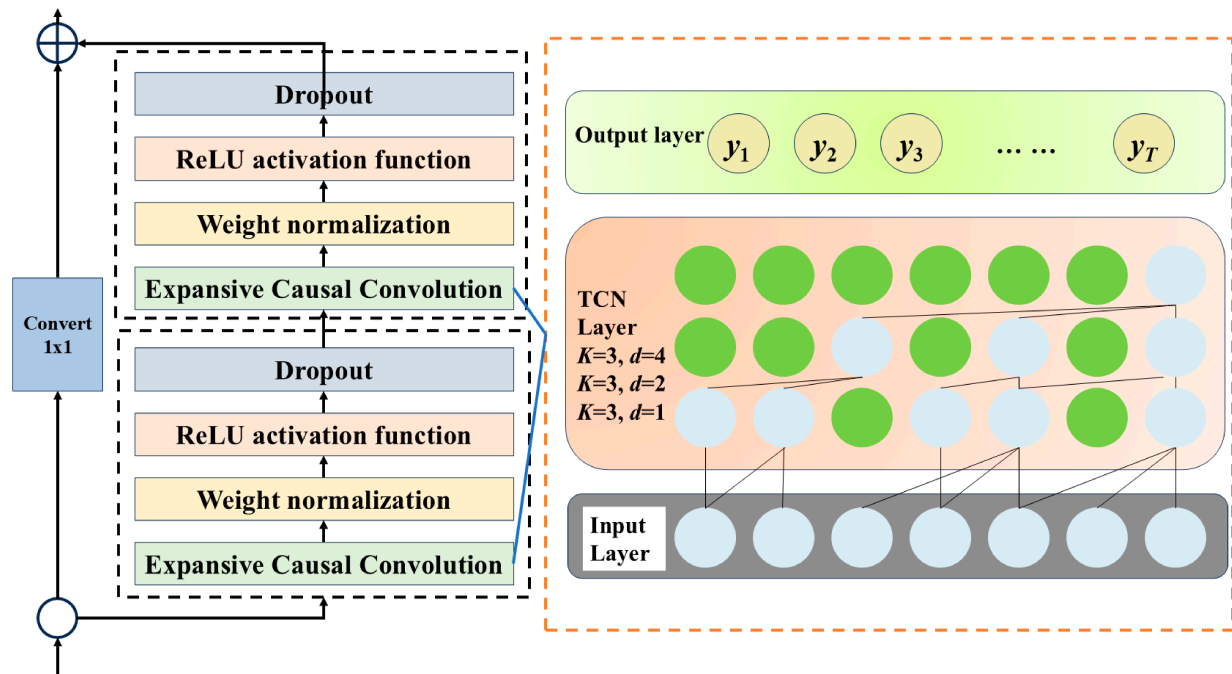


Figure 2. TCN architecture.

(1) Causal Convolution

Short-term electricity load forecasting follows a strict temporal ordering. When predicting the load at time t , only data from the current and previous time steps can be utilized, while future information must not be incorporated. Therefore, TCN employs causal convolution to preserve the temporal causality of the model. For an input sequence x , the one-dimensional causal convolution can be expressed as follows [42]:

$$y_t = \sum_{i=1}^{K-1} f(i)x_{t-i} \quad (18)$$

where y_t represents the convolution output at time t ; K represents the size of the convolution kernel; $f(i)$ represents the convolution kernel parameters; and x_{t-1} represents the input data at the current and past time steps.

As shown in Equation (16), the output x_{t-1} depends only on data up to and including time step t ; therefore, it avoids the issue of future information leakage that may arise in traditional convolutional networks when used for time series forecasting. For electricity load forecasting, this architecture ensures consistency in the use of temporal information between the model training process and the actual forecasting process.

(2) Dilated Causal Convolution

When only standard causal convolution is used, the network needs to stack a large number of convolutional layers to obtain a sufficiently large temporal receptive field, which not only increases the number of model parameters but may also reduce training efficiency. To address this issue, TCN introduces dilated convolution on the basis of causal

convolution, expanding the receptive field by inserting gaps between the elements of the convolution kernel. The dilated causal convolution can be expressed as follows [42]:

$$y_t = \sum_{i=1}^{K-1} f(i)x_{t-di} \quad (19)$$

where d is the expansion factor. When $d = 1$, expanded causal convolution is equivalent to standard causal convolution; as d increases, the convolution kernel can capture historical information over a longer range.

As shown in Figure 2, the TCN adopts a convolutional kernel size of $K = 3$, with dilation factors sequentially set to $d = 1$, $d = 2$, and $d = 4$. As the network becomes deeper, the dilation factor gradually increases, enabling the model to extract load variation features at different temporal scales. When the dilation factor increases exponentially, it can be expressed as follows:

$$d_l = 2^l \quad (20)$$

where l represents the number of layers in the TCN.

In this way, the receptive field can be rapidly expanded without significantly increasing the network depth, enabling the model to simultaneously capture short-term local fluctuations and longer-term variation trends. For a TCN with a convolutional kernel size of K and L network layers, the receptive field can be expressed as follows:

$$R = 1 + (K - 1) \sum_{l=0}^{L-1} d_l \quad (21)$$

When $K = 3$ and the dilation factors are sequentially set to 1, 2, and 4, the network can cover a longer historical time window, thereby enhancing the model's ability to capture long-term temporal dependencies in electricity load.

(3) Weight Normalization and ReLU Activation [43]

To improve the stability of the TCN training process, Weight Normalization is introduced after the dilated causal convolution, as illustrated in the figure. Weight Normalization reparameterizes the network weights to reduce the influence of differences in parameter scales across different network layers on the training process, enabling more stable parameter updates. After Weight Normalization, the ReLU activation function is employed to introduce nonlinear mapping, which can be expressed as follows:

$$\text{ReLU}(x) = \max(0, x) \quad (22)$$

ReLU preserves positively activated features while suppressing negative responses. Moreover, its simple computational structure and efficient gradient propagation facilitate the extraction of nonlinear variation patterns from electricity load sequences by the TCN.

(4) Dropout and Residual Connections

To reduce the risk of overfitting during model training, a Dropout layer is introduced after the activation function in the TCN. By randomly dropping a proportion of neurons, Dropout reduces the network's excessive dependence on specific features, thereby improving the generalization capability of the model. In addition, as illustrated in the figure, the TCN employs residual connections to directly fuse the input features with the output of the convolutional module. The basic formulation is given as follows:

$$o = F(x) + x \quad (23)$$

where x represents the input to the residual block; $F(x)$ represents the nonlinear features after expansion causal convolution, weight normalization, ReLU, and Dropout; and o represents the final output of the residual block.

When the input and output feature dimensions are inconsistent, a 1×1 convolution is employed for dimensional matching:

$$o = F(x) + W_s x \quad (24)$$

where W_s represents the parameters of a 1×1 convolution. The residual structure effectively improves gradient propagation in deep networks, enabling the network to maintain good training stability while continuously expanding its receptive field [44].

As shown in the Figure, the input time series first enters the TCN, where multi-layer dilated causal convolutions progressively extract load variation features at different temporal scales. The lower layers with smaller dilation factors primarily capture local variations between adjacent time steps, whereas the higher layers with progressively larger dilation factors further capture historical dependencies over longer time horizons. After convolution, Weight Normalization, ReLU activation, and Dropout, the features from different layers are fused through residual connections, ultimately producing high-dimensional temporal features that incorporate multi-scale temporal dependencies.

2.3. Bidirectional Long Short-Term Memory (BiLSTM)

The Bidirectional Long Short-Term Memory (BiLSTM) network is a bidirectional temporal feature extraction architecture developed on the basis of the conventional LSTM [45]. Traditional LSTM processes input sequences in chronological order and primarily learns features from historical information preceding the current time step. In contrast, BiLSTM consists of two independent LSTM networks that process the input sequence in the forward and backward directions, respectively, enabling the model to simultaneously learn the correlations between preceding and subsequent time steps and thereby obtain more comprehensive contextual features. In the BiLSTM architecture, the forward LSTM processes the input sequence in chronological order and extracts temporal features transmitted from past information to the current time step. The backward LSTM processes the input sequence in the reverse temporal order, extracting features associated with the current time step from subsequent information. The hidden states obtained from the two directions are then fused to form a comprehensive feature representation containing bidirectional temporal information. Through this bidirectional information interaction mechanism, BiLSTM effectively compensates for the limitation of unidirectional LSTM, which can only utilize historical information, and provides stronger feature representation and temporal dependency learning capabilities for complex time-series modeling tasks [46].

In the proposed model, BiLSTM is placed after the TCN feature extraction module. After TCN extracts multi-scale temporal features from the input data, BiLSTM further captures long-term dependencies and bidirectional correlations among different time steps. TCN effectively extracts local and multi-scale temporal features, while BiLSTM further strengthens the learning of global temporal information. The complementary characteristics of the two components improve the model's ability to characterize complex temporal variation patterns and enhance its prediction accuracy. The specific network architecture is illustrated in Figure 3. For one-hour-ahead forecasting at time (t) , the input window contains only observations available from $(t-L+1)$ to (t) , while the forecasting target is the load at $(t+1)$. Although BiLSTM processes the input sequence in both forward and backward directions, both directions operate exclusively within this historical window. The backward LSTM therefore does not access observations after the forecasting origin or the target value at $(t+1)$, and no future-target information is used in either training or testing.

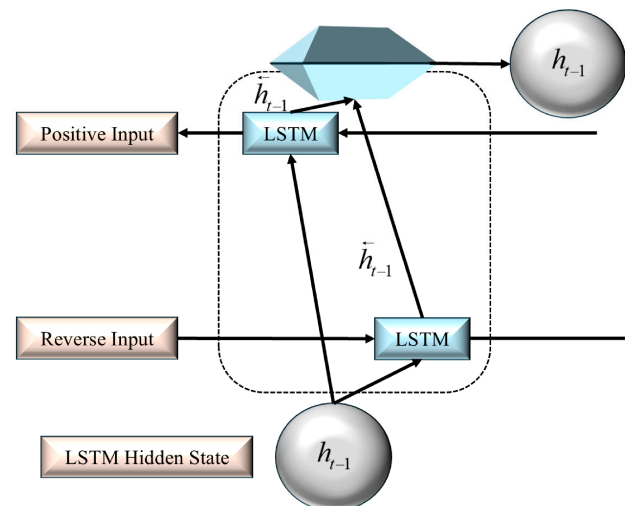


Figure 3. BiLSTM architecture.

2.4. Quantile Regression (QR)

The final layer of the proposed model is the Quantile Regression (QR) layer [47], which primarily maps the deep temporal features extracted by the TCN-BiLSTM network into electricity load prediction results at different conditional quantile levels. Electricity load is jointly influenced by various factors, including users' electricity consumption behavior, weather conditions, calendar types, seasonal variations, and random disturbances, and therefore exhibits substantial fluctuations and uncertainty. Conventional load forecasting methods mainly focus on point prediction, providing only a single predicted value for a given time step and thus failing to fully characterize the potential variation range and uncertainty of actual electricity load. Koenker et al. [21] introduced quantile regression in 1978. Without assuming that the data follow a specific probability distribution in advance, this method can effectively characterize the relationship between explanatory and response variables at different conditional quantiles. Compared with conventional mean regression, quantile regression describes the conditional distribution characteristics of electricity load across multiple quantile levels. It not only provides information on the central tendency of load variations but also reflects the fluctuation range of load predictions under different conditions. Therefore, incorporating quantile regression into the electricity load forecasting model enables a more comprehensive characterization of the uncertainty associated with the load forecasting process.

In this study, assuming that the independent variable is $X = [x_1, x_2, \dots, x_n]$ and the dependent variable is $Y = [y_1, y_2, \dots, y_n]$, the linear quantile regression model is expressed as follows:

$$Q_{y_i}(\tau|x_i) = \beta(\tau)x_i, i = 1, 2, \dots, n \quad (25)$$

where $Q_{y_i}(\tau|x_i)$ is the τ -th conditional quantile of the dependent variable y_i , lying in the interval $(0, 1)$; x_i is an $(m + 1)$ -dimensional vector; $\beta(\tau)$ is the vector of regression coefficients, $\beta(\tau) = [\beta_0(\tau), \beta_1(\tau), \dots, \beta_m(\tau)]$.

According to the above Equation, when the training data of the model are known, determining the regression coefficient vector $\beta(\tau)$ at different quantile levels can be transformed into an optimization problem that minimizes the loss function L [48], which can be expressed as follows:

$$\hat{\beta}(\tau) = \arg \min_{\beta} L = \arg \min_{\beta} \sum_i^n \gamma_{\tau} [y_i - x_i^T \beta(\tau)] \quad (26)$$

where $\gamma_\tau(s)$ denotes the pinball (quantile) loss function defined as follows:

$$\gamma_\tau(s) = \begin{cases} \tau s, & s \geq 0 \\ (\tau - 1)s, & s < 0 \end{cases} \quad (27)$$

where $s = y_i - x_i^T \beta(\tau)$ is the residual for sample i .

A prequential 168 h rolling conformal calibration is applied separately to each target's raw 90% prediction interval. For an observed sample i , y_i denotes the actual load, while the superscript raw and the subscripts L and U denote the uncalibrated lower and upper quantile forecasts, respectively.

The nonconformity score and rolling correction used at forecast time t are defined as follows:

$$\begin{cases} s_i = \max\{\hat{y}_{i,L}^{raw} - y_i, y_i - \hat{y}_{i,U}^{raw}, 0\} \\ \hat{q}_i = Q_{0.90}^{higher}(\{s_i : i \in H_t\}) \end{cases} \quad (28)$$

where H_t denotes the index set containing at most the latest 168 nonconformity scores available before forecast time t . It is initialized with the last 168 validation scores and is then updated only after each test observation becomes available. The quantile operator in Equation (28) denotes the empirical 0.90 quantile evaluated using the higher selection rule. The calibrated bounds and median are

$$\begin{cases} \hat{y}_{t,L}^{cal} = \hat{y}_{t,L}^{raw} - \hat{q}_t \\ \hat{y}_{t,U}^{cal} = \hat{y}_{t,U}^{raw} + \hat{q}_t \\ \hat{y}_{t,0.50}^{cal} = \hat{y}_{t,0.50}^{raw} \end{cases} \quad (29)$$

Because the rolling correction is nonnegative, Equation (29) leaves the median forecast unchanged and expands the raw lower and upper bounds by the same additive amount.

In this study, the TCN-BiLSTM network is employed to extract deep temporal features from electricity load data, while the quantile regression layer simultaneously outputs prediction results at different quantile levels (0.05 and 0.95), thereby constructing prediction intervals for electricity load. Compared with conventional point forecasting methods, this approach not only provides point estimates of electricity load but also characterizes the potential variation range of the predictions through the lower and upper quantiles, effectively capturing the uncertainty inherent in electricity load forecasting. Consequently, the proposed method enhances the model's ability to characterize load fluctuations while improving the reliability and practical applicability of the forecasting results.

2.5. Framework of the IOOA-TCN-BiLSTM-QR-Based Short-Term Probabilistic Interval Forecasting Model for Electricity, Cooling, and Heating Loads

To achieve high-accuracy short-term probabilistic interval forecasting of electricity, cooling, and heating loads in integrated energy systems, an IOOA-based TCN-BiLSTM-QR forecasting framework was developed. The framework fully exploits the capability of TCN to extract local and multi-scale temporal features, the ability of BiLSTM to model long-term temporal dependencies, and the capability of quantile regression (QR) to characterize predictive uncertainty. Meanwhile, IOOA is employed to adaptively optimize the key hyperparameters of the TCN-BiLSTM-QR model. The specific forecasting procedure is as follows:

Step 1: Data preprocessing and sample construction.

First, historical electricity, cooling, and heating load data from the integrated energy system are preprocessed, including outlier treatment and data normalization, to reduce the influence of differences in units and numerical scales among different load types on model training. Subsequently, supervised learning samples are constructed using a sliding time

window, where continuous historical multi-load information is used as the model input and the electricity, cooling, and heating loads at the future time step are used as prediction targets, thereby establishing a multivariate short-term load forecasting dataset.

Step 2: IOOA initialization.

The population size, maximum number of iterations, and search ranges of the parameters to be optimized are specified for the Improved Osprey Optimization Algorithm (IOOA). An initial osprey population is then randomly generated. Each osprey represents a candidate set of TCN-BiLSTM-QR hyperparameters, with its position corresponding to a specific parameter combination and serving as a candidate solution for subsequent optimization.

Step 3: Multi-scale temporal feature extraction using TCN.

The input electricity–cooling–heating load time series are fed into the TCN, where causal and dilated convolutions are employed to progressively extract temporal features from historical load information. Different dilation rates are used to expand the receptive field, enabling the network to simultaneously capture local fluctuations, periodic variations, and dynamic characteristics at different temporal scales. This enhances the model’s ability to represent complex load variation patterns.

Step 4: Deep temporal feature learning using BiLSTM.

The high-dimensional temporal features extracted by the TCN are subsequently fed into the BiLSTM network. Forward and backward recurrent structures are employed to encode the load sequences from both temporal directions, further capturing the underlying long-term temporal dependencies among electricity, cooling, and heating loads. The cascaded TCN-BiLSTM architecture forms a hierarchical feature-learning process of “local feature extraction–long-term dependency modeling,” thereby improving the model’s adaptability to complex nonlinear load variations.

Step 5: Probabilistic interval forecasting using QR.

A quantile regression (QR) layer is introduced based on the features extracted by BiLSTM. The conditional predictions corresponding to different quantiles are learned using the pinball loss. For a given nominal confidence level of $1 - \alpha$, the lower quantile $\alpha/2$ and upper quantile $1 - \alpha/2$ are obtained to construct the corresponding prediction interval. Unlike conventional point forecasting, which provides only a single predicted value, QR further characterizes the potential fluctuation range of future electricity, cooling, and heating loads, thereby extending deterministic forecasting to probabilistic interval forecasting.

Step 6: Fitness evaluation.

For each candidate hyperparameter combination generated by IOOA, the TCN-BiLSTM-QR model is trained and evaluated on the validation set, and the corresponding fitness value is calculated based on the forecasting results. Prediction errors and probabilistic interval forecasting performance are incorporated into the model evaluation to quantitatively assess the quality of different hyperparameter configurations.

Step 7: Iterative optimization using IOOA.

According to the fitness values of the current osprey population, the improved osprey optimization strategy is employed to continuously update the population positions, gradually guiding candidate solutions toward promising regions of the search space. By enhancing global exploration and local exploitation capabilities, IOOA reduces the likelihood of premature convergence and continuously searches for the optimal hyperparameter configuration of the TCN-BiLSTM-QR model. The optimal parameter set is obtained when the maximum number of iterations is reached or the convergence criterion is satisfied.

Step 8: Optimal model determination and forecasting.

The optimal hyperparameter configuration obtained by IOOA is assigned to the TCN-BiLSTM-QR model, which is then retrained using the complete training dataset. During the testing stage, historical electricity, cooling, and heating load data are input into the trained

model to obtain point forecasts and prediction intervals at different confidence levels for the three load types.

Step 9: Forecasting performance evaluation.

Finally, model performance is evaluated from both point forecasting and probabilistic interval forecasting perspectives. R^2 , MAPE, MAE, and RMSE are employed to evaluate point forecasting accuracy, while PICP, PINAW, CWC, and MPICD are used to assess empirical interval coverage, interval width, the coverage-width trade-off, and center deviation. These metrics evaluate the specified 90% prediction interval and do not constitute an assessment of the full predictive distribution.

Overall, the framework of the proposed short-term probabilistic interval forecasting model for electricity, cooling, and heating loads is illustrated in Figure 4. The proposed framework integrates IOOA-based hyperparameter optimization, TCN-based multi-scale feature extraction, BiLSTM-based long-term dependency modeling, and QR-based probabilistic interval construction, thereby achieving a synergistic improvement in multi-energy load forecasting accuracy and predictive uncertainty quantification. It provides more reliable forecasting information for operational scheduling and energy management in integrated energy systems.

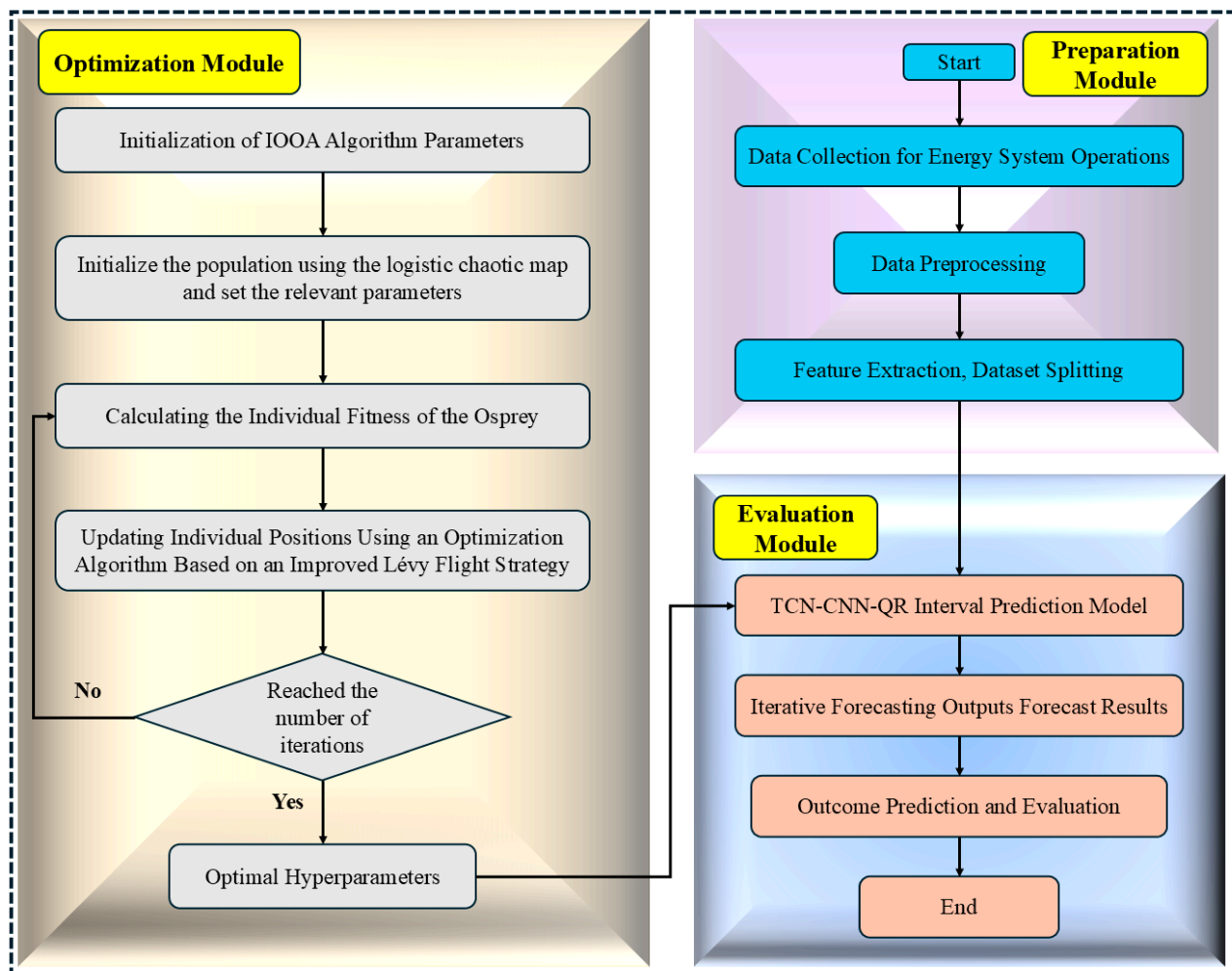


Figure 4. Framework of the probabilistic range prediction model for short-term electricity, cooling, and heating loads.

3. Simulation Experiments and Result Analysis

3.1. Data Collection and Preprocessing

To evaluate the proposed probabilistic forecasting framework under realistic multi-energy operating conditions, a university integrated energy system (IES) is selected as the primary case study. A university campus provides a representative application scenario for multi-energy load forecasting because its energy consumption is simultaneously affected by building occupancy, academic and administrative schedules, seasonal environmental conditions, and the coordinated operation of different energy subsystems. In particular, electricity, cooling, and heating demands exhibit distinct but interrelated temporal characteristics. Electricity consumption is closely associated with the daily operation of teaching, research, office, and supporting facilities, whereas cooling and heating demands are more strongly affected by outdoor environmental conditions and seasonal thermal requirements. These characteristics make the university IES an appropriate case for evaluating the capability of the proposed model to capture heterogeneous temporal patterns and forecasting uncertainty across multiple energy carriers.

The dataset consists of continuous operational observations from the university IES with a sampling interval of 1 h, covering 8760 consecutive hours over a complete year. The dataset contains 14 fields, including a timestamp, three energy-load variables, and ten meteorological and environmental variables. The three target variables are electricity load, cooling load, and heating load, measured in kW, ton, and MMBtu, respectively. The ten environmental variables include temperature, wind speed, wind direction, surface albedo, atmospheric pressure, precipitable water, solar zenith angle, relative humidity, dew point, and cloud type. These variables provide complementary information concerning atmospheric conditions, solar geometry, and environmental variations that may affect campus energy demand. The three energy loads exhibit distinct seasonal characteristics over the year. As shown in Figure 5, electricity load presents a pronounced high-load plateau during the summer period, reflecting increased electricity demand under intensive cooling and campus operation conditions. Cooling load generally increases during the warm season and decreases during cooler periods, while heating load remains relatively higher during the cold season. Therefore, the three load series exhibit different seasonal sensitivities while also retaining potential coupling relationships within the IES. Such characteristics provide a suitable test environment for examining whether the proposed IOOA-TCN-BiLSTM-QR model can simultaneously capture short-term fluctuations, long-term temporal dependencies, and uncertainty associated with seasonal load variations. For model development, the three historical load series and the available meteorological and environmental variables are jointly used as input information. In addition, temporal characteristics, including hour, day of the week, month, and weekend status, are extracted from the timestamp. Cyclical encoding based on sine and cosine transformations is applied to the hour, day-of-week, and month variables to preserve their periodic characteristics and avoid artificial discontinuities between adjacent temporal boundaries, such as 23:00 and 00:00 or December and January. Wind direction is similarly transformed into sine and cosine components to address the circular nature of angular measurements. To ensure that the forecasting evaluation reflects realistic future prediction conditions, the complete dataset is processed using the four-season rolling expanding-window forecasting strategy described in the subsequent sections. Data standardization is independently fitted within each training window to prevent information leakage from future observations. The complete hourly observations are retained for the analysis, while standardized representations are used only where required for model training and visualization. Overall, this university IES case provides a representative multi-energy forecasting scenario in which electricity, cooling, and heating loads are subject to different seasonal patterns, intraday operating

characteristics, and environmental influences. The combination of continuous year-round observations and multiple meteorological and environmental variables enables a comprehensive assessment of the proposed framework from the perspectives of point forecasting accuracy, probabilistic interval forecasting, seasonal adaptability, and feature contribution.

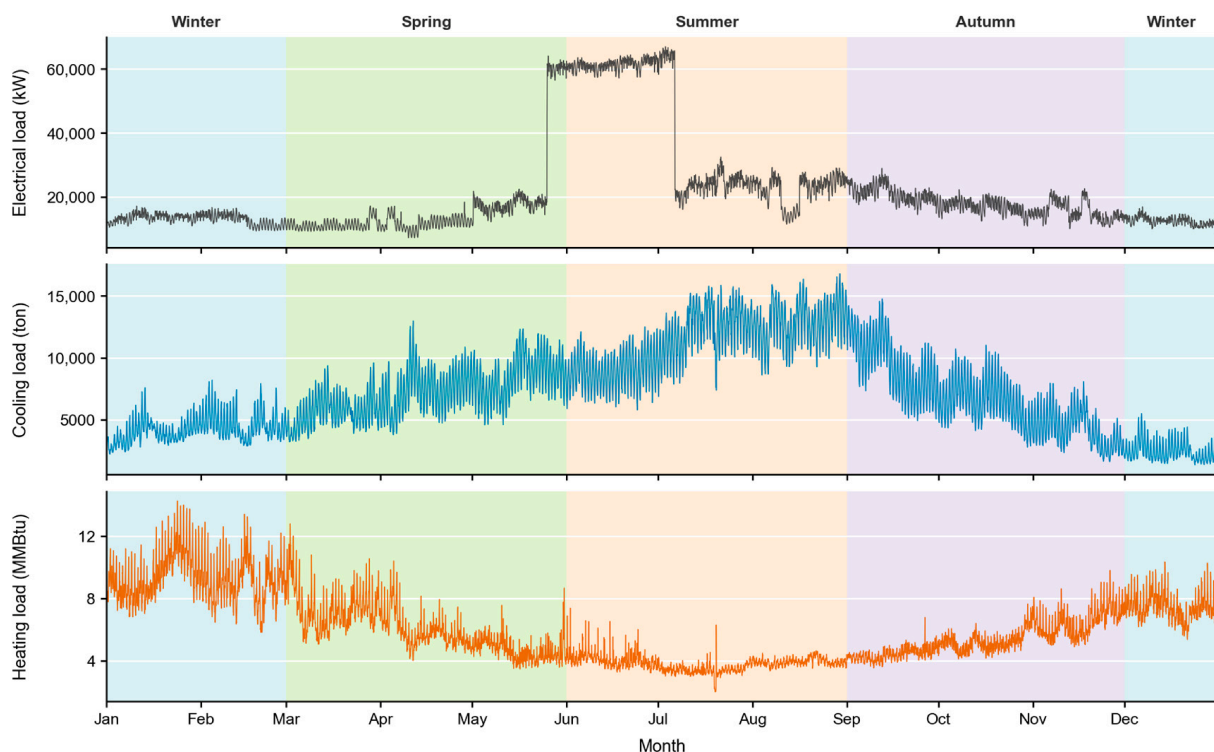


Figure 5. Historical Curves of Multiple Loads in an Integrated Energy System.

To further illustrate the full-year variations in the exogenous inputs, Figure 6 presents the different-scale features after conversion to full-year z-scores solely for visualization and groups them into three categories according to their physical meanings. This display transformation is not involved in model training; instead, data standardization for model input is independently fitted within the training window of each rolling fold. The original wind-direction angle is not plotted directly; rather, the sine and cosine components actually used by the model are displayed to avoid artificial discontinuities when the direction crosses 0° . All curves retain the complete hourly observations throughout the year. The z-score transformation is used only to eliminate scale differences among features with different units, thereby enabling their variations to be compared on the same axis.

A fixed division in which the final 20% of the annual data are used as the test set would cause the test period to be concentrated mainly in autumn and winter, making it difficult to effectively evaluate the model's forecasting and extrapolation capability under the high-load conditions occurring in summer. Therefore, a four-season rolling expanding-window evaluation strategy is adopted. For the spring fold, January–March data are used for training, April data for validation, and May data for testing. For the summer fold, January–May data are used for training, June data for validation, and July data for testing. For the autumn fold, January–August data are used for training, September data for validation, and October data for testing. For the winter fold, January–October data are used for training, November data for validation, and December data for testing. Each of the four test months contains 744 complete hourly observations, with May, July, October, and December representing the extrapolation tests under spring, summer, autumn, and winter operating conditions, respectively.

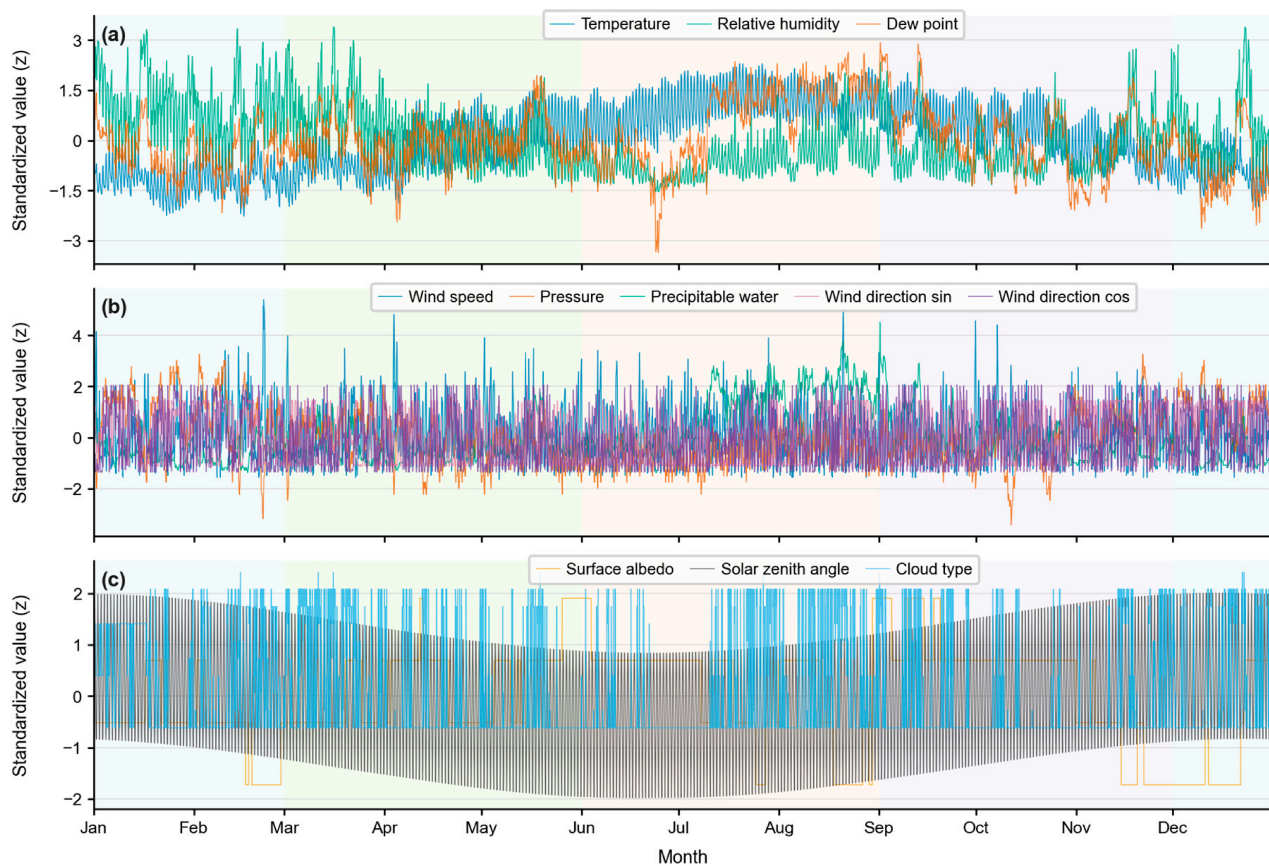


Figure 6. Year-round variations in input features. (a) Temperature, relative humidity, and dew point; (b) wind speed, atmospheric pressure, precipitable water, and sine/cosine-encoded wind direction; (c) surface albedo, solar zenith angle, and cloud type.

This evaluation scheme allows the same forecasting procedure to progressively expand the available historical information and refit or update the model at four different temporal origins. All four folds share the same model architecture and training rules rather than constructing four season-specific models. Within each fold, the training, validation, and test sets strictly follow chronological order. At each test time step, only historical information that has already been observed before the corresponding prediction time is accessible, while the target value and any future observations remain unavailable. Continuous input variables and the three load targets are standardized using method $z = \frac{x - \mu_{tr}}{\sigma_{tr}}$, where the parameters of μ_{tr} and σ_{tr} are estimated solely from the training data of each fold and subsequently applied to the corresponding validation and test sets, thereby effectively preventing data leakage.

A supervised learning scheme with a 24-h historical window is employed to forecast the electric, cooling, and heating loads for the subsequent 1 h. Samples are assigned to the corresponding fold according to their target timestamps. The resulting numbers of training sequences for the spring, summer, autumn, and winter folds are 2136, 3600, 5808, and 7272, respectively, while each fold contains 720 validation sequences and 744 test sequences. Validation and test samples are allowed to use adjacent historical observations that have already been observed before their respective prediction timestamps, including observations at the end of the preceding stage, but they do not use the target-time observation or any future information.

3.2. Experimental Setup

All experiments are conducted in a Python 3.11.9 environment, with Jupyter Notebook 7.2.2 employed to systematically organize the entire workflow, including data preprocessing, model construction, parameter training, performance evaluation, and result visualization, thereby ensuring the rigor, reproducibility, and traceability of the experiments. Data reading, temporal index processing, missing-value handling, and feature engineering are primarily implemented using Pandas 2.2.3 and NumPy 1.26.4, while dataset partitioning and input-sample construction are performed according to the characteristics of the time-series forecasting task. To mitigate the effects of differences in feature scales and numerical ranges on training stability, both input features and target variables are standardized using the preprocessing functions provided by Scikit-learn 1.5.2. The standardization parameters obtained during training are persistently stored using Joblib 1.4.2 and subsequently used for inverse transformation after model prediction, allowing the predicted results to be restored to their original physical units while facilitating experimental reproducibility and subsequent model deployment. The deep learning models are developed using the PyTorch 2.4.1 framework, with its automatic differentiation mechanism employed for network parameter updating and model optimization. During model training, the availability of the CUDA environment is automatically detected, and GPU-based parallel computation is preferentially adopted when available to improve computational efficiency; otherwise, the training process automatically falls back to the CPU. Experimental results and analytical figures are primarily generated using Matplotlib 3.9.2, with figure dimensions, fonts, axis labels, and resolution consistently configured before export.

3.3. Validation of the IOOA's Effectiveness

To validate the optimization performance of the Improved Osprey Optimization Algorithm (IOOA), this study selected the Particle Swarm Optimization (PSO) [49], Grey Wolf Optimization (GWO) [50], Whale Optimization Algorithm (WOA) [51], and the original Osprey Optimization Algorithm (OOA) as benchmark algorithms, all of which are commonly employed in short-term power load forecasting tasks. Six representative benchmark test functions widely used in the literature [52] were selected for comparative experiments, as presented in Table 1. The selected test functions comprise two categories: unimodal and multimodal functions. The unimodal functions (f_1 – f_3) are primarily employed to evaluate the convergence speed and optimization accuracy of the algorithms, whereas the multimodal functions (f_4 – f_6) are mainly used to assess their ability to escape local optima and their global search performance. By comprehensively evaluating the performance of the algorithms across different types of benchmark functions, the global exploration capability, local exploitation capability, and convergence stability of IOOA can be systematically assessed.

Table 1. Selected Test Functions.

Function	Dimensions	Scope	Target Value
$f_1(x) = \max_i \{ x_i , 1 \leq i \leq n\}$	30	$[-100, 100]$	0
$f_2(x) = \sum_{i=1}^n x_i + \prod_{i=1}^n x_i $	30	$[-10, 10]$	0
$f_3(x) = \sum_{i=1}^n x_i^2$	30	$[-100, 100]$	0
$f_4(x) = -20 \exp[-0.2(\frac{1}{n} \sum_{i=1}^n x_i^2)^{1/2}] - \exp[\frac{1}{n}(\sum_{i=1}^n \cos(2\pi x_i))] + 20 + e$	30	$[-32, 32]$	0
$f_5(x) = \sum_{i=1}^n [x_i^2 - 10 \cos(2\pi x_i) + 10]$	30	$[-5.12, 5.12]$	0
$f_6(x) = (1/4000) \sum_{i=1}^n x_i^2 - \prod_{i=1}^n \cos(x_i/\sqrt{i}) + 1$	30	$[-600, 600]$	0

To ensure the fairness and comparability of the experimental results, all algorithms were evaluated under the same experimental settings. The dimensionality of all benchmark

functions was uniformly set to 30, the population size was set to 50, and the maximum number of iterations was set to 500. Each algorithm was independently executed 10 times on each benchmark function, and the average convergence curves were calculated to reduce the influence of stochastic initialization and random search processes. The six benchmark functions were selected as a representative and complementary test suite rather than as an exhaustive representation of all possible optimization landscapes. Specifically, the functions include three unimodal functions (f_1 – f_3) and three multimodal functions (f_4 – f_6). The unimodal functions mainly evaluate convergence speed and optimization accuracy, whereas the multimodal functions introduce multiple local optima and therefore provide a means of assessing global exploration and local-optimum-escaping capability. Thus, the combination of these two categories allows the basic exploitation, exploration, convergence, and search stability of IOOA to be examined from complementary perspectives. By comparing the optimization results and convergence processes of PSO, GWO, WOA, OOA, and IOOA, the effectiveness of the proposed improvement strategies can be assessed under different representative search landscapes. The experimental results are presented in Figure 7. As shown in Figure 7, IOOA generally exhibits a relatively rapid decreasing trend in the convergence curves across the benchmark functions and approaches the theoretical optima within a comparatively small number of iterations. Compared with PSO, GWO, WOA, and the original OOA, IOOA demonstrates stronger exploration capability during the early search stage, which can be attributed to the improved diversity of the initial population introduced by the Logistic chaotic mapping. During the later search stage, the Lévy flight strategy provides variable search step lengths and occasional long-range movements, thereby improving the ability to escape local optima while maintaining effective local exploitation. These characteristics enable IOOA to continuously refine promising solutions and improve the final optimization accuracy. In particular, for the multimodal benchmark functions, IOOA maintains a relatively stable convergence process without pronounced stagnation or large fluctuations over extended iterations. This behavior indicates that the combination of Logistic chaotic mapping and Lévy flight contributes to a more effective balance between global exploration and local exploitation. Furthermore, the relatively consistent final convergence results obtained over the 10 independent runs indicate that the optimization performance of IOOA is not strongly dependent on a particular random initialization, thereby providing evidence of its robustness under stochastic search conditions.

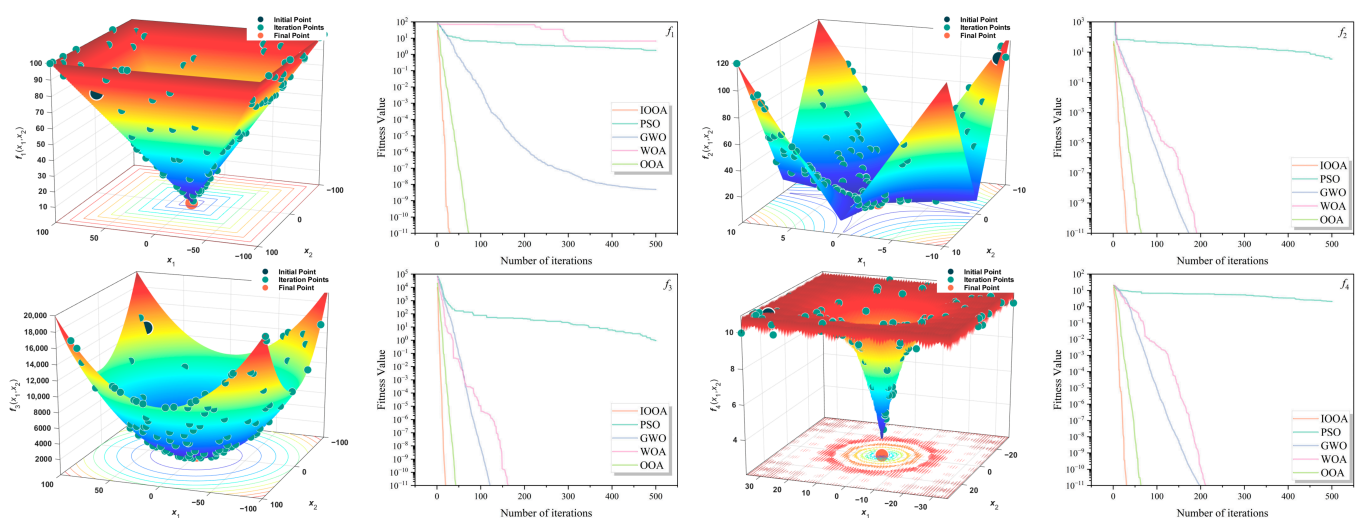


Figure 7. Cont.

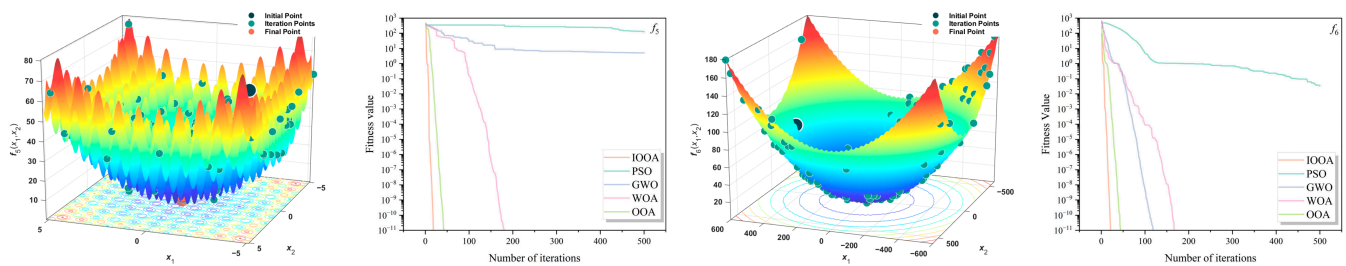


Figure 7. 3D images of the test functions f_1 – f_6 for the optimization algorithms and the fitness iteration curves.

It should be noted that the benchmark experiments are intended to provide algorithm-level evidence of the basic optimization capability of IOOA rather than to claim that six analytical functions can completely reproduce the high-dimensional and highly non-linear hyperparameter landscape of the forecasting model. The subsequent forecasting experiments therefore provide task-specific validation by directly employing IOOA to optimize the key hyperparameters of the TCN-BiLSTM-QR model. In this way, the benchmark experiments and the forecasting experiments constitute two complementary levels of evaluation: the former examines the fundamental search characteristics of IOOA under controlled and diverse optimization landscapes, while the latter verifies whether these characteristics can be effectively transferred to the nonlinear hyperparameter optimization problem encountered in the proposed forecasting framework. Overall, the benchmark results provide evidence that the improved IOOA possesses competitive convergence speed, optimization accuracy, global exploration capability, local-optimum-escaping ability, and search stability. These results support its subsequent application to the adaptive optimization of the key hyperparameters of the TCN-BiLSTM-QR model, while the forecasting experiments are used to further assess its effectiveness in the specific application scenario considered in this study.

3.4. Determination of Model Parameters and Evaluation Metrics

3.4.1. Determination of Model Parameters

To obtain a parameter combination that balances prediction accuracy, training stability, and computational complexity, IOOA was employed to jointly optimize eight key hyperparameters of the TCN-BiLSTM-QR model, covering the input window, TCN architecture, BiLSTM architecture, and training-related parameters. Each candidate solution was used to train the model exclusively on the training set within the corresponding rolling forecasting stage, and its fitness was evaluated on the validation set in the standardized target space. Within each rolling fold, the electricity, cooling, and heating targets were standardized independently using their respective means and standard deviations estimated only from the training partition; the same training-derived parameters were then applied unchanged to the validation and test partitions. For each validation sample, Pinball losses were calculated separately for each of the three standardized load targets at the quantile levels of 0.05, 0.50, and 0.95. The fitness function was defined as the unweighted arithmetic mean over all validation samples, three load types, and three quantile levels. Thus, each load type and each quantile contributed equally to the optimization objective, preventing the larger original numerical magnitude of the electricity load from dominating the losses of the cooling or heating load. Inverse transformation was performed only when reporting the forecasting metrics in their original physical units. A lower fitness value indicates a stronger capability of the candidate parameter combination to characterize the conditional quantiles. Discrete-valued hyperparameters were mapped to the nearest integer within their allowable ranges before fitness evaluation, while continuous-valued parameters were

retained at the corresponding precision. The test set was not involved in hyperparameter optimization. The initial values, search ranges, and final optimized values of the hyperparameters are presented in Table 2.

Table 2. Initial values, search ranges, and final values of the eight hyperparameters optimized by IOOA.

Hyperparameter	Initial Value (Search Range)	Final Value	Hyperparameter	Initial Value (Search Range)	Final Value
Input window length (h)	24 (12–72)	24	Number of BiLSTM layers	1 (1–3)	1
Number of TCN residual blocks	2 (1–4)	2	BiLSTM hidden units per direction	48 (32–96)	64
Number of convolutional channels	32 (16–64)	48	Dropout rate	0.20 (0.10–0.50)	0.15
TCN kernel size	3 (2–5)	3	Initial learning rate	0.001 (0.0001–0.005)	0.0003

Specifically, for a candidate hyperparameter vector h , the validation fitness was calculated as $F(h) = \frac{1}{N_v L Q} \sum_{n=1}^{N_v} \sum_{l=1}^L \sum_{q=1}^Q Q_{\tau_q} \left(y_{n,l}^* - \hat{y}_{n,l}^{*(\tau_q)}(h) \right)$, where $Q_{\tau}(u) = u[\tau - \mathbb{I}(u < 0)]$ is the pinball loss, N_v is the number of validation samples, $L = 3$ is the number of load types, $Q = 3$ is the number of quantile levels, $\tau_q \in \{0.05, 0.50, 0.95\}$, $y_{n,l}^*$ is the standardized validation target, and $\hat{y}_{n,l}^{*(\tau_q)}$ is its predicted conditional quantile in the same standardized space. PICP, PINAW, CWC, and MPICD were not included as additional weighted terms in $F(h)$. Instead, interval coverage and width were represented indirectly through the joint lower- and upper-quantile losses. Observations outside the predicted interval increase the corresponding tail loss, whereas interval bounds placed unnecessarily far from the observations also increase the loss. These interval metrics were calculated only for independent test-set evaluation and did not influence IOOA-based hyperparameter selection.

It is necessary to clarify the tuning fairness and computational budget settings of all comparative models. The proposed IOOA-TCN-BiLSTM-QR model employs an eight-dimensional adaptive hyperparameter search to determine the configuration of the integrated forecasting pipeline, with a maximum of 60 training epochs. In contrast, the baseline neural models, including BiLSTM, TCN, and TCN-BiLSTM, are trained using fixed hyperparameter configurations and a maximum of 20 training epochs. These settings reflect a comparison between complete forecasting pipelines rather than strictly controlled comparisons under identical hyperparameter-tuning budgets. The additional optimization effort of the proposed model should therefore be considered when interpreting the performance differences. Specifically, the observed improvements may result from the combined effects of the TCN-BiLSTM architecture, quantile regression framework, and the adaptive hyperparameter optimization provided by IOOA, while the larger tuning budget may also contribute to obtaining a more suitable model configuration. Therefore, the performance gains should not be interpreted as arising solely from the model architecture itself. Nevertheless, all comparative models were evaluated using the same dataset, data preprocessing procedure, chronological data-splitting strategy, forecasting targets, and evaluation metrics. The comparison consequently assesses the predictive performance of the complete forecasting pipelines under their respective configuration strategies, which is consistent with the practical objective of evaluating an integrated forecasting framework.

For the proposed model, IOOA ultimately determined the optimal configuration as follows: an input window of 24 h, two TCN residual blocks, 48 convolutional channels, a kernel size of 3, one BiLSTM layer with 64 hidden units per direction, a dropout rate of 0.15, and an initial learning rate of 3×10^{-4} . The batch size was fixed at 64 during model training.

The dilation rates of the two TCN residual blocks were set to 1 and 2, respectively. The model was trained using the Adam optimizer with a maximum of 60 training epochs. Early stopping was applied when the validation quantile loss failed to improve for 9 consecutive epochs, and the model weights corresponding to the minimum validation loss were restored. The quantile levels were set to 0.05, 0.50, and 0.95, where the 0.50 quantile was used for point forecasting, while the 0.05–0.95 quantiles were used to construct the 90% prediction interval. The eight-dimensional search space was defined according to the architectural and training hyperparameters listed in Table 2. The remaining training settings, including batch size, optimizer, dilation rates, quantile levels, and early-stopping patience, were fixed throughout the optimization and final training process.

Figure 8 shows that the training loss in all four rolling stages decreased rapidly during the initial epochs and subsequently entered a relatively stable convergence region. The spring, summer, autumn, and winter stages required 18, 27, 59, and 35 training epochs, respectively, with minimum validation quantile losses of 0.0311, 0.0200, 0.0113, and 0.0139, respectively. The spring stage was affected by the load regime transition in May, resulting in a relatively higher validation loss and earlier triggering of the early-stopping criterion. As the expanding-window training samples increased, the validation losses in the autumn and winter stages decreased further. The training and validation curves exhibited generally consistent trends, with only minor fluctuations in the later stages and no sustained divergence. These results indicate that the structural parameters determined by IOOA maintain good convergence stability across different seasonal training windows, while the early-stopping strategy effectively reduces unnecessary iterations and mitigates overfitting.

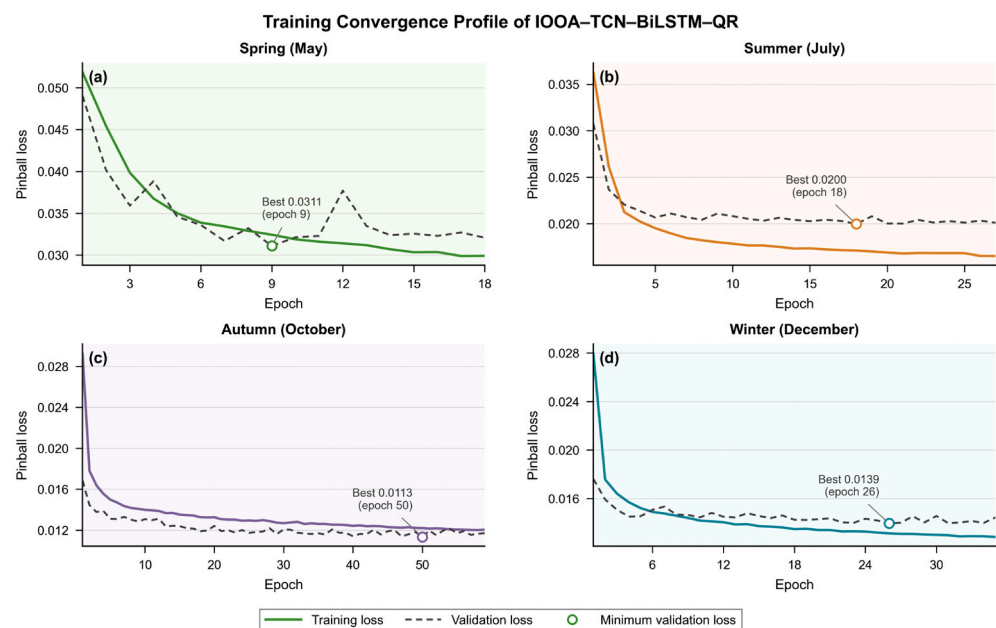


Figure 8. Training convergence profiles of the IOOA-TCN-BiLSTM-QR model for four seasons. (a) Spring (May): training and validation pinball loss, minimum validation loss 0.0311 at epoch 9. (b) Summer (July): training and validation pinball loss, minimum validation loss 0.0200 at epoch 18. (c) Autumn (October): training and validation pinball loss, minimum validation loss 0.0113 at epoch 50. (d) Winter (December): training and validation pinball loss, minimum validation loss 0.0139 at epoch 26. Solid lines: training loss; dashed lines: validation loss; hollow circles: minimum validation loss.

A one-factor-at-a-time sensitivity analysis was further conducted around the IOOA-selected configuration using the same four expanding-window seasonal validation folds, standardized pinball-loss objective, early-stopping rule, and fixed random seed of 100.

With the remaining settings held constant, input windows of 18 and 30 h, one and three TCN residual blocks, 40 and 56 convolutional channels, 48 and 80 BiLSTM hidden units, and initial learning rates of 0.0002 and 0.0005 were evaluated. The selected configuration yielded a four-fold mean validation pinball loss of 0.01908. The perturbed configurations ranged from 0.01887 to 0.01951, corresponding to relative changes of -1.13% to $+2.26\%$. The small, bidirectional variations show that forecasting performance is locally stable and does not depend on a narrowly tuned parameter point. Because several neighboring settings produced marginally lower or higher mean losses, the selected configuration should be interpreted as a robust and competitive solution on a shallow local performance plateau rather than as a unique numerical optimum. These diagnostic reruns were not used for secondary hyperparameter selection.

3.4.2. Determining Model Evaluation Metrics

(1) Interval Prediction Metrics

To comprehensively evaluate the probabilistic interval prediction performance of the Improved Osprey Optimization Algorithm-based TCN-BiLSTM-QR model for short-term electricity, cooling, and heating load forecasting, four evaluation metrics were selected from three aspects, namely the reliability, sharpness, and overall performance of the prediction intervals: Prediction Interval Coverage Probability (PICP), Prediction Interval Normalized Average Width (PINAW), Coverage Width-based Criterion (CWC), and Mean Prediction Interval Coverage Distance (MPICD). The corresponding formulations are given in Equations (30)–(33) [53].

$$\text{PICP} = \frac{1}{n} \sum_{i=1}^n I(y_i \in [\hat{y}_i^l, \hat{y}_i^u]) \quad (30)$$

$$\text{PINAW} = \frac{1}{n} \sum_{i=1}^n \frac{\hat{y}_i^u - \hat{y}_i^l}{y_{\max} - y_{\min}} \quad (31)$$

$$\text{CWC} = \text{PINAW}[1 + \gamma e^{-\eta[\text{PICP} - (1-\alpha)]}] \quad (32)$$

$$\text{MPICD} = \frac{1}{n} \sum_{i=1}^n \left(\left| \frac{\hat{y}_i^u - \hat{y}_i^l}{2} - y_i \right| \right) \quad (33)$$

where y_i is the true value of the i -th sample; $\hat{y}_i$ is the predicted value of the i -th sample; $\bar{y}$ is the average of the true values; n is the number of samples; y_i is the i th observation; $\hat{y}_i^l$ is the lower limit of the prediction interval; $\hat{y}_i^u$ is the upper limit of the prediction interval; $I(\cdot)$ is the indicator function, which takes the value of 1 when the condition inside the parentheses is established, otherwise it is 0; $y_{\min}$ and $y_{\max}$ denote the minimum and maximum values of the observed load, respectively. The parameter α represents the significance level associated with the prescribed confidence level. For instance, a 90% prediction confidence level corresponds to $\alpha = 0.1$. The adequacy of the prediction interval coverage is determined by comparing PICP with the nominal coverage probability $1 - \alpha$. Specifically, if $\text{PICP} \geq 1 - \alpha$, the obtained interval is regarded as satisfying the prescribed coverage requirement, and the penalty coefficient γ is assigned a value of zero. Under this condition, the CWC reduces directly to PINAW, indicating that no additional penalty is imposed on the interval width. Conversely, when $\text{PICP} < 1 - \alpha$, the prediction interval fails to achieve the desired coverage and is therefore considered insufficiently covered. In this case, γ is set to one, and the penalty parameter η is introduced to impose an additional penalty on the CWC. A larger η results in a stronger penalty for inadequate coverage, thereby enhancing the sensitivity of CWC to under-coverage and providing a more comprehensive assessment of the reliability and sharpness of the prediction interval.

In summary, PICP is primarily used to evaluate the coverage reliability of the prediction intervals, while PINAW assesses their compactness and sharpness. CWC measures the overall trade-off between interval coverage and width, whereas MPICD further quantifies the deviation between the center of the prediction interval and the actual load values. The coordinated use of these metrics enables a multidimensional assessment of the improvement achieved by the Improved Osprey Optimization Algorithm (IOOA) in the probabilistic interval prediction performance of the TCN-BiLSTM-QR model, thereby providing a more comprehensive basis for quantifying and evaluating prediction uncertainty in multi-load scenarios involving electricity, cooling, and heating demands.

(2) Point Prediction Metrics

For point prediction performance, the coefficient of determination (R^2), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) were employed to quantitatively evaluate the point prediction results [6,49,52]. These metrics characterize the predictive performance of the model from different perspectives, including goodness of fit, relative error magnitude, average deviation, and sensitivity to large errors. The corresponding formulations are given in Equations (34)–(37).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (34)$$

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n |(\hat{y}_i - y_i) / y_i| \quad (35)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (36)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (37)$$

The symbols in this Equation have the same meanings as in Section 3.4.1.

3.5. Analysis of Interval Forecast Results

To examine the capability of the proposed IOOA-TCN-BiLSTM-QR forecasting model to characterize uncertainty under different seasonal operating conditions and load fluctuation patterns, the four-season rolling expanding-window forecasting scheme established in Section 3.1 was adopted. A complete 744 h period in May, July, October, and December was selected as the representative test period for spring, summer, autumn, and winter, respectively. The prediction interval performance was further analyzed at the seasonal scale in terms of coverage, sharpness, and reliability. During each test period, the model simultaneously generated three quantiles, namely $q_{0.05}$, $q_{0.50}$, and $q_{0.95}$. Among them, $q_{0.50}$ was used as the median point prediction, while $q_{0.05}$ and $q_{0.95}$ were taken as the lower and upper bounds of the prediction interval, respectively, thereby constructing a prediction interval with a nominal confidence level of 90%. This setting enables the model to capture the expected variation trends of the loads while dynamically characterizing the prediction uncertainty induced by seasonal load fluctuations. Since the actual future load values are unavailable during real-time rolling forecasting, the dynamic adjustment of the prediction intervals was performed using only the most recent 168 h of observed prediction errors prior to the current forecasting time. No information from future periods was utilized, thereby effectively preventing data leakage from affecting the evaluation of interval width and coverage. Furthermore, the penalty coefficient η in the CWC metric was fixed at 50 to impose a stronger penalty on insufficient interval coverage, encouraging the model to maintain adequate coverage while preserving interval sharpness.

Figure 9 shows the dynamic variations in the prediction intervals for electricity, cooling, and heating loads across different seasons, providing an intuitive representation of the model's adaptability to different load levels and fluctuation intensities. Table 3 further provides a quantitative evaluation of the four-season interval prediction performance using PICP, PINAW, CWC, and MPICD, together with the average prediction interval width. Specifically, PICP measures the proportion of actual observations falling within the prediction interval, PINAW evaluates the normalized interval width, CWC jointly considers interval coverage and sharpness, and MPICD further characterizes the average deviation of the prediction interval from the actual observations. Together, these metrics assess the empirical coverage, normalized width, coverage-width trade-off, and center deviation of the specified 90% prediction interval.

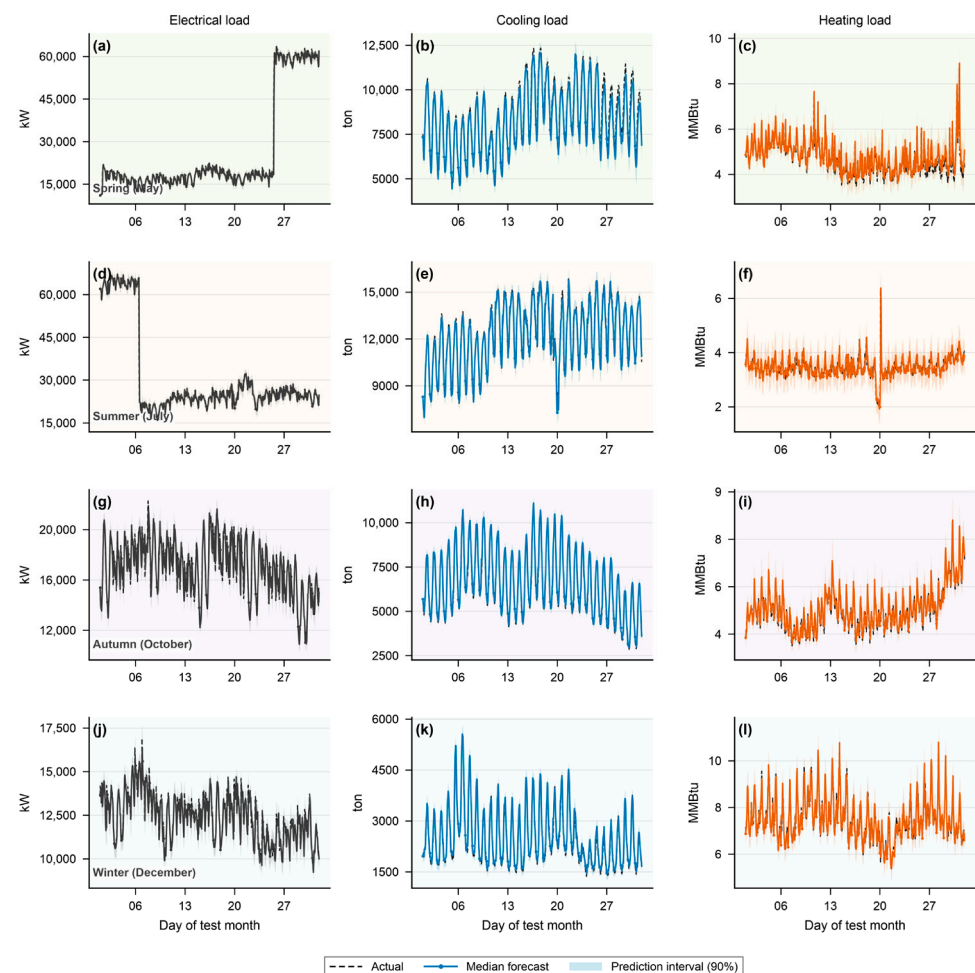


Figure 9. Ninety-percent prediction intervals for electrical, cooling, and heating loads across four seasonal test stages. (a–c) Spring (May): electrical, cooling, and heating load forecasts; (d–f) Summer (July): electrical, cooling, and heating load forecasts; (g–i) Autumn (October): electrical, cooling, and heating load forecasts; (j–l) Winter (December): electrical, cooling, and heating load forecasts. Dashed lines: actual values; solid blue lines: median forecast; light blue shaded regions: 90% prediction intervals.

As shown in Figure 9 and Table 3, the proposed model exhibits clear adaptability to different seasons and load types, although differences in coverage reliability and interval sharpness remain across operating periods. The prediction intervals are dynamically calibrated using the most recent 168 h of available interval nonconformity scores. At each forecast time, the empirical 0.90 quantile of these scores, evaluated using the higher selection rule, is subtracted from the raw lower bound and added to the raw upper bound, while

the median forecast remains unchanged. Because this additive correction is nonnegative, the raw interval is either expanded or left unchanged rather than narrowed. Only scores from observations available before the forecast time are used, thereby preserving the causal forecasting structure.

Table 3. Interval prediction performance for electrical, cooling, and heating loads across four seasonal test stages.

Testing Phase	Load Type	PICP/(%)	PINAW	CWC	MPICD	Average Interval Width
Spring (May)	Electrical Load (kW)	87.63	0.0458	0.1953	642.83	2432.33
Spring (May)	Cooling Capacity (tons)	85.62	0.1399	1.3908	291.90	1085.26
Spring (May)	Heat Load (MMBtu)	87.77	0.2351	0.9524	0.339	1.220
Summer (July)	Electrical Load (kW)	90.73	0.0676	0.0676	769.08	3440.91
Summer (July)	Cooling Capacity (tons)	89.25	0.1241	0.3050	259.56	1112.51
Summer (July)	Heat Load (MMBtu)	97.72	0.2591	0.2591	0.173	1.112
Fall (October)	Electrical Load (kW)	89.78	0.1634	0.3454	426.41	1867.59
Fall (October)	Cooling Capacity (tons)	91.13	0.0767	0.0767	156.65	630.74
Fall (October)	Heat Load (MMBtu)	91.13	0.1960	0.1960	0.156	0.690
Winter (December)	Electrical Load (kW)	90.46	0.1777	0.1777	291.56	1355.34
Winter (December)	Cooling Capacity (tons)	93.95	0.1058	0.1058	90.79	436.78
Winter (December)	Heat Load (MMBtu)	90.86	0.2249	0.2249	0.262	1.118

Note: For PICP, a nominal coverage rate of nearly 90% is optimal; for PINAW and CWC, the lower the value, the better. Since the three load categories have different units of measurement, MPICD is used only to compare the same load across different seasons.

During the spring period, the electricity load was strongly affected by an operating-condition transition in late May. Around 25 May, the mean electricity load increased from 17,463.32 kW to 56,855.96 kW and subsequently reached a peak of 64,029.98 kW at 22:00. Correspondingly, the average prediction interval width increased from 2241.60 kW to 3086.27 kW, indicating that the rolling calibration mechanism responded to the increased forecasting uncertainty associated with the new operating state. Following the transition, the PICP increased to 90.48%, suggesting that the interval calibration gradually adapted to the changed load regime. However, the monthly PICP for electricity load remained at 87.63%, indicating temporary under-coverage during the initial transition stage. This behavior is consistent with the causal nature of the calibration mechanism: the interval width can only be adjusted after new forecasting-error information becomes available, and therefore a short calibration lag may occur when the underlying load distribution changes abruptly. Similar behavior was observed for the cooling and heating loads during spring, with PICPs of 85.62% and 87.77%, respectively. These results indicate that the proposed model can adapt to changing uncertainty levels, but the calibration response may be temporarily insufficient when the load regime changes more rapidly than the 168 h rolling error window can reflect. Thus, the robustness of the proposed model under abrupt operating conditions should be interpreted in terms of both its adaptive response and its short-term calibration lag. In practical deployment, additional recalibration or retraining triggers based on abrupt changes in forecasting errors or load regimes could further improve interval reliability under extreme conditions.

In summer, both electricity and cooling loads operated at relatively high levels, with peak values of 66,910.70 kW and 15,844.65 ton, respectively. The electricity load achieved a PICP of 90.73%, while its PINAW and CWC were both 0.0676, indicating a favorable balance between coverage and interval sharpness. The cooling load achieved a PICP of 89.25%, only 0.75 percentage points below the nominal 90% level, demonstrating satisfactory uncertainty characterization under high cooling demand. The heating load achieved a

PICP of 97.72%, although its PINAW reached 0.2591, indicating that the high coverage was accompanied by relatively wide prediction intervals. During autumn, the three load types gradually declined from their summer levels and exhibited relatively stable fluctuations. The electricity load achieved a PICP of 89.78%, while the cooling and heating loads both reached 91.13%, with coverage generally close to or above the nominal level. The cooling load achieved both a PINAW and CWC of 0.0767, indicating a favorable balance between coverage and sharpness. The heating load also achieved its lowest seasonal MPICD of 0.156 MMBtu, suggesting relatively good interval calibration during this transitional period. During winter, the electricity load achieved a PICP of 90.46%, with an MPICD of 291.56 kW, the lowest among its four seasonal periods. The cooling load achieved a PICP of 93.95%, with PINAW and CWC both equal to 0.1058 and an MPICD of 90.79 ton, indicating satisfactory overall interval performance. The heating load achieved a PICP of 90.86%, although its PINAW remained relatively high at 0.2249, suggesting that a wider uncertainty margin was required to maintain the desired coverage.

Overall, the average PICP values across the three load types for spring, summer, autumn, and winter were 87.01%, 92.56%, 90.68%, and 91.76%, respectively. The relatively lower spring coverage was mainly associated with abrupt operating-condition changes, whereas the coverage in summer, autumn, and winter was generally close to or above the nominal 90% level. The corresponding seasonal average CWC values were 0.8462, 0.2106, 0.2061, and 0.1695, respectively, with winter showing the best overall interval performance. These results demonstrate that interval quality depends jointly on coverage reliability and interval sharpness and should therefore be evaluated using PICP, PINAW, CWC, and MPICD rather than PICP alone. Overall, the proposed rolling calibration strategy enables the prediction intervals to adapt to changing load uncertainty, while its short-term calibration lag under abrupt regime transitions indicates the need for further improvement through adaptive recalibration or regime-change detection.

3.6. Comparison of Point Prediction Results from Different Models

To verify the effectiveness and superiority of the proposed Improved Osprey Optimization Algorithm-based TCN-BiLSTM-QR (IOOA-TCN-BiLSTM-QR) model for short-term probabilistic interval prediction of electricity, cooling, and heating loads in terms of point prediction performance, several representative load forecasting models were selected as benchmark models, and their prediction results were compared with those of the proposed model. The benchmark models included conventional statistical learning, machine learning, and deep learning approaches, specifically CNN-BiLSTM-Attention [13], BiLSTM [16] MTL-LSTM [18], Transformer [20], STGCN (Spatio-Temporal Graph Convolutional Network) [25], MMEMformer [26], AM-RU-BiLSTM [28], CEEMDAN-GWO-MKELM [50], and SP-RF-ARIMA [54]. These representative short-term load forecasting models were comprehensively compared with the proposed IOOA-TCN-BiLSTM-QR model. For the point prediction experiments, the prediction corresponding to the 0.50 quantile of the IOOA-TCN-BiLSTM-QR model was adopted as the final point prediction, i.e., the conditional median prediction was used as the load forecasting result, thereby ensuring consistency with the output format of the other point forecasting models. Meanwhile, all benchmark models were trained and tested using the same electricity, cooling, and heating load datasets, data preprocessing procedures, training–testing data partitioning strategy, input features, and forecasting horizon under identical experimental conditions. Separate point prediction experiments were conducted for short-term electricity, cooling, and heating load forecasting to minimize the influence of different experimental conditions on model performance evaluation. Furthermore, four metrics, namely the coefficient of determination (R^2), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Root Mean Squared

Error (RMSE), were employed to quantitatively evaluate the point prediction performance of all models. The point prediction results for the different load types are summarized in Tables 4–6.

Table 4. Comparison of Point Prediction Performance for Electricity Load among Different Forecasting Models.

Model	R ²		MAPE (%)		MAE		RMSE/(kW)	
	Training Set	Test Set	Training Set	Test Set	Training Set	Test Set	Training Set	Test Set
The model used in this study	0.9868	0.9385	1.7473	2.6117	298	528	679	1053
CNN-BiLSTM-Attention	0.9845	0.9204	2.1825	2.9614	372	584	755	1104
BiLSTM	0.9865	0.9026	2.03	3.2523	352	628	732	1131
MTL-LSTM	0.9861	0.9154	2.0247	2.8941	350	567	730	1096
Transformer	0.9864	0.9124	2.1412	3.1135	371	608	747	1121
STGCN	0.9864	0.8585	2.2411	4.0558	377	765	739	1246
MMEMformer	0.9814	0.8811	2.5027	3.5374	416	672	781	1178
AM-RU-BiLSTM	0.9831	0.8984	2.2614	3.8065	382	794	755	1260
CEEMDAN-GWO-MKELM	0.9845	0.8756	2.3145	4.2445	384	714	756	1325
SP-RF-ARIMA	0.9964	0.9321	1.2614	2.65	219	534	483	1067

Table 5. Comparison of Point Prediction Performance for Cooling Load among Different Forecasting Models.

Model	R ²		MAPE (%)		MAE		RMSE/(Ton)	
	Training Set	Test Set	Training Set	Test Set	Training Set	Test Set	Training Set	Test Set
The model used in this study	0.9974	0.9781	1.5221	2.6712	95	171	130	232
CNN-BiLSTM-Attention	0.9914	0.9382	2.6612	4.2922	160	289	215	377
BiLSTM	0.9931	0.9664	2.2825	3.6925	139	222	189	287
MTL-LSTM	0.9947	0.9692	2.1944	3.4174	134	208	184	279
Transformer	0.9928	0.9641	2.4037	3.7735	148	226	200	297
STGCN	0.9932	0.9605	2.2472	4.0114	137	236	186	304
MMEMformer	0.9841	0.9541	3.0827	4.1327	185	268	248	340
AM-RU-BiLSTM	0.9862	0.9254	2.7014	4.4921	163	307	221	408
CEEMDAN-GWO-MKELM	0.9872	0.9124	2.6587	6.3145	175	312	235	457
SP-RF-ARIMA	0.9944	0.9711	2.0622	3.0841	125	201	174	271

Table 6. Comparison of Point Prediction Performance for Heating Load among Different Forecasting Models.

Model	R ²		MAPE (%)		MAE		RMSE/(MMBtu)	
	Training Set	Test Set	Training Set	Test Set	Training Set	Test Set	Training Set	Test Set
The model used in this study	0.9944	0.9731	1.8910	2.3600	0.1184	0.2151	0.1674	0.3010
CNN-BiLSTM-Attention	0.9634	0.8807	3.9242	4.6380	0.2745	0.3307	0.3875	0.4457
BiLSTM	0.9782	0.8472	3.2473	5.0373	0.2167	0.3514	0.3097	0.4896
MTL-LSTM	0.9802	0.9182	3.1571	3.7717	0.2077	0.2631	0.2974	0.3672
Transformer	0.9744	0.9051	3.4547	4.4221	0.2364	0.2933	0.3364	0.4127
STGCN	0.9781	0.8351	3.2675	5.8277	0.2161	0.3767	0.3099	0.5271
MMEMformer	0.9584	0.9027	4.1952	4.8474	0.2937	0.3255	0.4148	0.4489
AM-RU-BiLSTM	0.9674	0.9484	3.6779	3.8613	0.2512	0.2878	0.3563	0.3958
CEEMDAN-GWO-MKELM	0.9641	0.8962	3.8411	4.6248	0.2717	0.3475	0.5201	0.4912
SP-RF-ARIMA	0.9807	0.8907	3.0617	4.3374	0.2038	0.3162	0.2922	0.4297

For reproducibility, all benchmark models used 24 h input sequences, the same standardized seasonal folds, and a fixed random seed of 100, without test-set tuning. The neural benchmarks were trained using AdamW with a learning rate of 4×10^{-4} , a weight decay of 1×10^{-5} , a batch size of 96, Smooth L1 training loss, gradient clipping at 1.0, and validation-MSE early stopping over a maximum of 20 epochs with a patience of five epochs. The fixed architecture settings were as follows: CNN-BiLSTM-Attention used 24 convolutional channels, a kernel size of 3, and 32 hidden units per BiLSTM direction; BiLSTM used one bidirectional layer with 32 hidden units per direction; MTL-LSTM used 36 LSTM hidden units, a 32-unit shared layer, and three task-specific heads; Transformer used one encoder layer with a model dimension of 40, four heads, an 80-unit feed-forward layer, and a dropout rate of 0.12; STGCN used a fixed three-load graph and two temporal convolution layers with 28 channels and a kernel size of 3; MMEMformer used two encoder layers, a model dimension of 36, four heads, four memory tokens, a 96-unit feed-forward layer, and a dropout rate of 0.10; and AM-RU-BiLSTM used 36 hidden units per direction with residual gating and attention pooling. The CEEMDAN-GWO-MKELM reproduction used 480 RBF random-feature components, with the kernel width selected from {0.01, 0.03, 0.08} and the ridge penalty from {0.03, 0.10, 0.30, 1.00} according to validation MSE. SP-RF-ARIMA used a 180-tree random forest with a maximum depth of 12, a minimum leaf size of 3, 70% feature sampling, and an AR(1)-style residual correction.

As shown in Table 4, the proposed IOOA-TCN-BiLSTM-QR model achieves (R^2), MAPE, MAE, and RMSE values of 0.9868, 1.7473%, 298 kW, and 679 kW, respectively, on the training set for electricity load, and 0.9385, 2.6117%, 528 kW, and 1053 kW, respectively, on the test set. Compared with the training set, the test-set R^2 decreases by only 0.0483, while MAPE, MAE, and RMSE increase by 0.8644 percentage points, 230 kW, and 374 kW, respectively. This relatively small degradation indicates that the proposed model maintains stable predictive performance when confronted with unseen samples and load fluctuations, demonstrating favorable generalization capability. On the test set, the proposed model achieves an R^2 of 0.9385, outperforming CNN-BiLSTM-Attention, BiLSTM, MTL-LSTM, Transformer, STGCN, MMEMformer, AM-RU-BiLSTM, and CEEMDAN-GWO-MKELM, whose corresponding R^2 values are 0.9204, 0.9026, 0.9154, 0.9124, 0.8585, 0.8811, 0.8984, and 0.8756, respectively. The corresponding improvements are 0.0181, 0.0359, 0.0231, 0.0261, 0.0800, 0.0574, 0.0401, and 0.0629, respectively. Compared with SP-RF-ARIMA ($R^2 = 0.9321$), the proposed model still achieves an improvement of 0.0064. These results demonstrate that IOOA-TCN-BiLSTM-QR provides a more accurate representation of the overall variation trend of electricity load. In terms of quantitative forecasting metrics, the proposed IOOA-TCN-BiLSTM-QR model achieves a test-set MAPE of 2.6117% for electrical load forecasting, which is marginally lower than the value of 2.6500% obtained by the optimal baseline SP-RF-ARIMA model. Compared with the other eight benchmark models, the proposed method reduces the MAPE values by 11.81%, 19.70%, 9.76%, 16.12%, 35.61%, 26.17%, 31.39%, and 38.47%, respectively. Correspondingly, the test MAE of the proposed model is 528 kW, with reduction rates of 9.59%, 15.92%, 6.88%, 13.16%, 30.98%, 21.43%, 33.50%, and 26.05% against the eight comparative models. The test RMSE reaches 1053 kW, yielding respective reductions of 4.62%, 6.90%, 3.92%, 6.07%, 15.49%, 10.61%, 16.43%, and 20.53%. Notably, even when compared with the most competitive SP-RF-ARIMA model, the proposed method achieves slight reductions of 1.45% in MAPE, 1.12% in MAE, and 1.31% in RMSE. Overall, these results demonstrate that the proposed IOOA-TCN-BiLSTM-QR model provides competitive predictive performance for electrical load point forecasting, with consistent improvements across the three commonly used error metrics. However, these performance differences should be interpreted in the context of the different model configuration and hyperparameter-tuning strategies adopted for the comparative models.

As shown in Table 5, the proposed model exhibits an even more pronounced advantage in cooling load forecasting. The training-set (R^2) is 0.9974, decreasing to 0.9781 on the test set, corresponding to a decrease of 0.0193. This relatively small decrease indicates that the model can effectively learn the underlying variation patterns of cooling loads and generalize the learned representations to unseen samples. The test-set (R^2) of 0.9781 is higher than those of CNN-BiLSTM-Attention, BiLSTM, MTL-LSTM, Transformer, STGCN, MMEMformer, AM-RU-BiLSTM, and CEEMDAN-GWO-MKELM by 0.0399, 0.0117, 0.0089, 0.0140, 0.0176, 0.0240, 0.0527, and 0.0657, respectively. Compared with SP-RF-ARIMA ($R^2 = 0.9711$), an improvement of 0.0070 is obtained. Thus, IOOA-TCN-BiLSTM-QR achieves the highest (R^2) among all benchmark models for the cooling load test set, demonstrating its strong capability to capture cooling load dynamics. The test-set MAPE of IOOA-TCN-BiLSTM-QR is 2.6712%, corresponding to reductions of 37.77%, 27.66%, 21.84%, 29.21%, 33.41%, 35.36%, 40.54%, and 57.70% relative to CNN-BiLSTM-Attention, BiLSTM, MTL-LSTM, Transformer, STGCN, MMEMformer, AM-RU-BiLSTM, and CEEMDAN-GWO-MKELM, respectively, and a reduction of 13.39% relative to SP-RF-ARIMA. Similarly, the MAE of 171 ton is reduced by 40.83%, 22.97%, 17.79%, 24.34%, 27.54%, 36.19%, 44.30%, 45.19%, and 14.93%, respectively, while the RMSE of 232 ton is reduced by 38.46%, 19.16%, 16.85%, 21.89%, 23.68%, 31.76%, 43.14%, 49.23%, and 14.39%, respectively. The particularly substantial improvement over CEEMDAN-GWO-MKELM, with reductions of 3.6433 percentage points in MAPE, 141 ton in MAE, and 225 ton in RMSE, suggests that signal decomposition combined with conventional optimized machine-learning approaches may have limitations in characterizing the complex nonlinear, periodic, and externally driven variations in cooling loads. In contrast, IOOA-TCN-BiLSTM-QR can more effectively extract local fluctuation patterns and long-term temporal dependencies from cooling load sequences.

As shown in Table 6, the results for heating load forecasting further confirm the effectiveness of the proposed model. IOOA-TCN-BiLSTM-QR achieves a training-set R^2 of 0.9944 and a test-set R^2 of 0.9731, with test-set MAPE, MAE, and RMSE values of 2.3600%, 0.2151 MMBtu, and 0.3010 MMBtu, respectively. The decrease in R^2 from the training to test set is only 0.0213, indicating favorable generalization capability. On the test set, the proposed model improves (R^2) by 0.0924, 0.1259, 0.0549, 0.0680, 0.1380, 0.0704, 0.0247, 0.0769, and 0.0824 over CNN-BiLSTM-Attention, BiLSTM, MTL-LSTM, Transformer, STGCN, MMEMformer, AM-RU-BiLSTM, CEEMDAN-GWO-MKELM, and SP-RF-ARIMA, respectively. Notably, the R^2 improvement reaches 0.1380 over STGCN and 0.1259 over BiLSTM, highlighting the limitations of single recurrent architectures and graph-based approaches in capturing the complex temporal dynamics of heating loads. The proposed model also achieves substantial improvements in error metrics. Its test-set MAPE of 2.3600% is reduced by 49.12%, 53.15%, 37.43%, 46.63%, 59.50%, 51.31%, 38.88%, 48.97%, and 45.59% compared with the nine benchmark models, respectively. The MAE of 0.2151 MMBtu is reduced by 34.96%, 38.79%, 18.24%, 26.66%, 42.90%, 33.92%, 25.26%, 38.10%, and 31.97%, respectively, while the RMSE of 0.3010 MMBtu is reduced by 32.47%, 38.52%, 18.03%, 27.07%, 42.90%, 32.95%, 23.95%, 38.72%, and 29.95%, respectively. In particular, the RMSE decreases from 0.4896 MMBtu for BiLSTM to 0.3010 MMBtu, corresponding to an absolute reduction of 0.1886 MMBtu. Compared with STGCN, the RMSE decreases from 0.5271 to 0.3010 MMBtu, representing a reduction of 0.2261 MMBtu. These results demonstrate that the proposed model is particularly effective in reducing large prediction deviations in heating load forecasting.

Overall, the results in Tables 4–6 demonstrate that the proposed IOOA-TCN-BiLSTM-QR model achieves consistently strong point-forecasting performance across electricity, cooling, and heating loads. The test-set R^2 values are 0.9385, 0.9781, and 0.9731 for electricity, cooling, and heating loads, respectively, while the corresponding MAPEs are 2.6117%,

2.6712%, and 2.3600%. Although the proposed model performs well across all three load types, the magnitude of improvement varies substantially among them. This difference should be interpreted in terms of both the intrinsic characteristics of the target load series and the performance levels of the competing models, rather than as evidence that the proposed architecture is universally more suitable for cooling and heating loads.

For electricity load, SP-RF-ARIMA already provides a highly competitive benchmark, achieving a test-set R^2 of 0.9321, an MAPE of 2.6500%, an MAE of 534 kW, and an RMSE of 1067 kW. Consequently, the remaining performance margin is relatively limited. Nevertheless, the proposed model still improves the MAPE, MAE, and RMSE by 1.45%, 1.12%, and 1.31%, respectively, relative to SP-RF-ARIMA, while increasing R^2 to 0.9385. In contrast, the larger performance gains obtained for cooling and heating loads indicate that these two load types contain temporal patterns that are less fully represented by the competing models. For cooling load, the proposed model reduces MAPE, MAE, and RMSE by 13.39%, 14.93%, and 14.39%, respectively, compared with SP-RF-ARIMA. For heating load, the corresponding reductions reach 45.59%, 31.97%, and 29.95%. Therefore, the larger improvements for cooling and heating are attributable not only to the proposed architecture itself but also to the greater performance headroom relative to the benchmark models. The different improvement margins can also be related to the heterogeneous temporal characteristics of the three energy-load series. Electricity, cooling, and heating loads are jointly affected by historical demand, periodic patterns, meteorological conditions, operating schedules, and other external factors, but the relative importance and temporal response of these factors differ among load types. The proposed TCN-BiLSTM architecture is designed to represent such heterogeneous temporal behavior through complementary mechanisms. Specifically, TCN extracts local and multi-scale temporal features through causal and dilated convolutions, allowing short-term fluctuations and changes at different temporal scales to be captured, while BiLSTM further models longer-range temporal dependencies within the historical input sequence. This complementary structure is particularly beneficial when the load dynamics cannot be adequately characterized by a single temporal scale or by autoregressive relationships alone.

The comparative results further support this interpretation. Traditional machine-learning and signal-decomposition approaches, such as CEEMDAN-GWO-MKELM and SP-RF-ARIMA, can achieve competitive performance for specific load types, but their effectiveness depends strongly on how well the underlying decomposition, autoregressive, and nonlinear feature relationships match the target load dynamics. Recurrent models such as BiLSTM and MTL-LSTM can capture long-term temporal dependencies, but they may provide less explicit multi-scale local feature extraction. Transformer-, STGCN-, and MMEMformer-based models offer alternative mechanisms for modeling global or cross-variable relationships, but they do not consistently achieve the best performance on the present datasets. In contrast, the proposed framework combines multi-scale local feature extraction by TCN with long-range temporal dependency modeling by BiLSTM, while IOOA adaptively optimizes the key model hyperparameters. This combination provides a flexible representation of load dynamics across different temporal scales. The contribution of IOOA should also be understood as being task-dependent. Because the statistical characteristics and forecasting difficulty of electricity, cooling, and heating loads differ, their corresponding optimal TCN-BiLSTM configurations may also differ. IOOA enables the model to search for suitable hyperparameter combinations rather than relying exclusively on manually selected settings. Therefore, its contribution is not expected to produce an identical performance gain for all load types. The overall comparative and ablation results indicate that the optimization strategy and temporal feature-learning components jointly contribute to the observed forecasting performance.

It is also important to note that the larger improvements for cooling and heating do not imply that the proposed model is immune to challenging operating conditions. As discussed in the seasonal analysis, abrupt changes in load regimes can temporarily reduce forecasting and probabilistic calibration performance. In particular, the electricity load experienced a substantial increase around 25 May, and although the prediction interval expanded in response to the increased uncertainty and the PICP subsequently recovered toward the nominal coverage level, temporary under-coverage was observed during the initial transition stage. This finding indicates that the proposed architecture can effectively model complex temporal patterns under normal operating conditions but may require additional rolling recalibration or model updating when the underlying load-generation mechanism changes abruptly. Finally, for the point-forecasting comparison, the 0.50 quantile is adopted as the central point estimate of the probabilistic forecasting model and is evaluated using the same R^2 , MAPE, MAE, and RMSE metrics as the deterministic benchmark models. The 0.50 quantile represents the conditional median of the predictive distribution and therefore provides a statistically meaningful central estimate while remaining fully consistent with the quantile-regression framework. Importantly, the use of the median forecast does not imply that it is universally superior to a separately optimized deterministic forecast based on MSE or MAE. A separately optimized deterministic model could potentially achieve different performance under a specific point-loss objective. Rather, the present comparison is intended to examine whether the proposed probabilistic framework can maintain competitive central forecasting accuracy while simultaneously generating prediction intervals. The results demonstrate that the proposed framework achieves strong point-prediction performance across all three load types while retaining its probabilistic uncertainty-quantification capability.

Overall, the results in Tables 4–6 verify the effectiveness of the proposed IOOA-TCN-BiLSTM-QR framework for multi-energy load forecasting. The model maintains competitive point-prediction accuracy across electricity, cooling, and heating loads, while the substantially larger improvements observed for cooling and heating can be attributed to the interaction between their temporal characteristics, the representational advantages of multi-scale temporal modeling, and the performance headroom of the competing models. These findings support the effectiveness of integrating IOOA-based adaptive hyperparameter optimization with complementary TCN-BiLSTM temporal representation and quantile regression, while also highlighting the need for rolling calibration and adaptive updating under abrupt changes in operating conditions.

To assess whether the point-forecasting differences in Tables 4–6 exceeded sampling variation, paired tests were additionally performed on the hourly out-of-sample absolute errors from the four seasonal test months, comprising 2976 forecasts per model and load type. Two-sided Diebold–Mariano tests used a Newey–West heteroskedasticity- and autocorrelation-consistent variance estimator with a lag of 24 h. Holm adjustment was applied across the nine benchmark comparisons within each load type. In Table 7, a positive DM statistic indicates a lower MAE for the proposed model, a negative statistic indicates a lower MAE for the benchmark, and an adjusted p value below 0.05 indicates a statistically significant difference. The first row identifies the proposed model as the reference, represented by 0.000 and 1.000.

As shown in Table 7, after Holm adjustment, the proposed model achieved significantly lower MAE than all nine benchmarks for both electrical and cooling loads and five of the nine benchmarks for heating load ($p < 0.05$). The DM statistics were positive in all heating-load comparisons, indicating that the proposed model also obtained lower MAE than every heating-load benchmark, although four differences were not statistically significant. These findings confirm the proposed model's overall advantage in MAE across all three load types.

Table 7. Diebold–Mariano statistics and Holm-adjusted p values for paired hourly absolute-error comparisons.

Model	Electrical Load		Cooling Load		Heating Load	
	DM Statistic	Adjusted p	DM Statistic	Adjusted p	DM Statistic	Adjusted p
The model used in this study	0.000	1.000	0.000	1.000	0.000	1.000
CNN-BiLSTM-Attention	7.220	<0.001	6.861	<0.001	4.861	<0.001
BiLSTM	9.095	<0.001	3.259	0.002	5.993	<0.001
MTL-LSTM	5.669	<0.001	3.137	0.003	1.931	0.189
Transformer	8.294	<0.001	4.774	<0.001	1.849	0.193
STGCN	13.504	<0.001	6.303	<0.001	4.866	<0.001
MMEMformer	10.546	<0.001	8.584	<0.001	1.990	0.186
AM-RU-BiLSTM	10.890	<0.001	5.854	<0.001	7.484	<0.001
CEEMDAN-GWO-MKELM	10.106	<0.001	26.940	<0.001	16.464	<0.001
SP-RF-ARIMA	2.530	0.011	4.590	<0.001	0.955	0.340

3.7. Model Applicability Analysis

To further evaluate the applicability of the proposed IOOA–TCN–BiLSTM–QR model across different data sources, load scales, and meteorological conditions, the publicly available HEEW 2022 integrated energy dataset was selected for external validation. The dataset integrates regional-level shared meteorological observations with all energy-related time series using consistent hourly timestamps, covering 8760 consecutive hourly samples from 00:00 on 1 January to 23:00 on 31 December 2022. The prediction targets are the total regional electrical, cooling, and heating loads. Data quality inspection identified no missing values, duplicate timestamps, temporal discontinuities, or negative load values. The eight -1 °F observations in the dew-point temperature series were confirmed to be valid low-temperature measurements rather than missing-value indicators and were therefore retained. The meteorological variables in the HEEW dataset include temperature, dew-point temperature, humidity, wind speed, wind gust, atmospheric pressure, and precipitation. Calendar information includes month, hour, day of week, weekend indicator, U.S. federal holiday indicator, and working-day indicator. However, the dataset does not provide solar zenith angle, cloud cover, wind direction, or surface albedo. To avoid introducing unverifiable information, the missing variables were not artificially imputed; instead, the three historical load series, available meteorological variables, and temporal encoding features were jointly used as model inputs. Temperature, dew-point temperature, wind speed, atmospheric pressure, and precipitation were uniformly converted to °C, m/s, hPa, and mm, respectively. All normalization parameters were fitted exclusively on the training data of each forecasting stage.

The external validation retained the model architecture and hyperparameter configuration determined in Section 3.4.1, including a 24 h input window, two TCN layers, 48 convolutional channels, a bidirectional BiLSTM with 64 hidden units per direction, a dropout rate of 0.15, an initial learning rate of 0.0003, and quantile outputs at 0.05, 0.50, and 0.95. The 168 h causal rolling calibration strategy was also maintained. Data partitioning followed the four-season rolling expanding-window scheme described in Section 3.1, with 744 hourly samples included in each test month. It should be emphasized that, for each season, the network was retrained and the prediction intervals were calibrated exclusively using HEEW observations available prior to the corresponding prediction time. Therefore, this experiment evaluates the transferability and external applicability of the model archi-

texture and fixed configuration to an independent dataset, rather than zero-shot transfer of model weights trained on the original dataset.

As shown in Table 8, across the 2976 hourly samples combined from the four test months, the proposed model achieved R^2 values of 0.9064, 0.9936, and 0.9760 for electrical, cooling, and heating loads, respectively, with corresponding MAPE values of 3.56%, 3.03%, and 3.71%. Notably, the MAE for cooling and heating loads was reduced by 42.11% and 7.74%, respectively, compared with the previous-hour persistence baseline, indicating that the local periodic feature extraction of TCN and the cross-temporal dependency modeling of BiLSTM retain meaningful predictive gains on the external dataset. For electrical load, the overall MAE was 8.52% higher than that of the persistence baseline; however, the proposed model still outperformed the baseline during the spring, autumn, and winter stages, with the disadvantage mainly concentrated in summer. The summer electrical load patterns in the HEEW dataset differ from the high-load plateau observed in the original dataset, while the external dataset lacks solar geometry and equipment operating-state variables. Consequently, the model may be more prone to prediction deviations during systematic load regime transitions. These results indicate that the proposed model exhibits strong cross-dataset adaptability for cooling and heating load forecasting, whereas cross-regional application to electrical load forecasting may benefit from incorporating local occupancy schedules, equipment start-stop states, or demand-control information.

Table 8. Point forecasting performance on the external HEEW 2022 dataset.

Load	Unit	R^2	MAPE (%)	MAE	RMSE	Percentage Change in MAE Relative to the Baseline (%)
Electric Load	kW	0.9064	3.56	881.25	1335.83	8.52
Cooling Load	ton	0.9936	3.03	207.27	303.17	−42.11
Heat Load	MMBtu	0.9760	3.71	0.189	0.276	−7.74

As shown in Table 9, the average PICP across the 12 “season $\times$ load” scenarios was 90.41%, with 8 scenarios achieving or exceeding the nominal 90% coverage level. When averaged arithmetically across the four seasons by load type, the PICP values for electrical, cooling, and heating loads were 90.26%, 90.15%, and 90.83%, respectively, indicating that the 168 h rolling calibration strategy was still able to maintain the overall coverage close to the target level on the external dataset. The average PINAW for cooling load was 0.1152, substantially lower than those for electrical load (0.3185) and heating load (0.2590), indicating that the cooling-load prediction intervals achieved greater compactness while maintaining satisfactory coverage.

From the perspective of seasonal boundary conditions, the spring cooling-load PICP was 86.69%, representing the most pronounced under-coverage scenario in the external validation. The summer heating-load and winter electrical-load PICPs were 89.65% and 89.78%, respectively, also slightly below the nominal target, with corresponding CWC values increasing to 0.7383 and 0.8428. These deviations mainly occurred during periods characterized by substantial changes in load formation mechanisms or operating modes, indicating that a unified fixed configuration cannot completely eliminate distribution shifts across seasons and regions. In practical deployment, rolling calibration should therefore be retained, while independent warning and retraining triggers should be established for scenarios such as the onset of spring cooling operation, low heating demand in summer, and regime changes in winter electrical loads.

Figure 10 further reveals the seasonal differences in external applicability. The R^2 values for cooling load remained stable at 0.9587–0.9839 across the four seasons, with MAE reductions of 30.59–56.23% relative to the persistence baseline, demonstrating the

most stable cross-dataset improvement. For electrical load, the MAE decreased by 15.59%, 6.89%, and 4.91% in spring, autumn, and winter, respectively, whereas it increased by 57.49% in summer, corresponding to a reduction in R^2 from 0.8111 for the persistence baseline to 0.6601. This result clearly illustrates the applicability boundary caused by abrupt changes in operating regimes and the absence of external variables. For heating load, the MAE reductions were -3.31% and -9.96% in spring and summer, respectively, while improvements of 6.96% and 21.12% were achieved in autumn and winter. Although the point forecasting gain was limited in summer, the PICP shown in Figure 10b remained at 89.65%, indicating that rolling calibration could maintain uncertainty coverage close to the nominal level even when point prediction errors increased.

Table 9. Seasonal interval forecasting performance on the external HEEW 2022 dataset.

Stage	Load	Unit	PICP (%)	PINAW	CWC	MPICD
Spring	Electric Load	kW	89.92	0.3575	0.7296	835.15
Spring	Cooling Load	ton	86.69	0.0880	0.5476	171.06
Spring	Heat Load	MMBtu	91.26	0.2659	0.2659	0.197
Summer	Electric Load	kW	90.05	0.2979	0.2979	1336.55
Summer	Cooling Load	ton	90.46	0.1444	0.1444	310.03
Summer	Heat Load	MMBtu	89.65	0.3370	0.7383	0.135
Fall	Electric Load	kW	91.26	0.2197	0.2197	782.67
Fall	Cooling Load	ton	92.61	0.1095	0.1095	225.00
Fall	Heat Load	MMBtu	91.40	0.2293	0.2293	0.189
Winter	Electric Load	kW	89.78	0.3988	0.8428	1014.18
Winter	Cooling Load	ton	90.86	0.1189	0.1189	130.85
Winter	Heat Load	MMBtu	90.99	0.2038	0.2038	0.251

Overall, the external validation supports the applicability of IOOA-TCN-BiLSTM-QR as a regional short-term probabilistic forecasting framework for integrated energy systems. When the target region provides at least several months of continuous hourly electricity, cooling, and heating load data together with synchronized meteorological and calendar information, the model can achieve relatively high point forecasting accuracy and near-nominal interval coverage through local retraining without re-optimizing the structural hyperparameters. Its primary scope of applicability includes campus-, park-, and regional-scale integrated energy systems characterized by consistent data granularity, significant short-term load persistence, and periodic model updating. For scenarios with no local historical data, frequent abrupt changes in operating regimes, or a requirement for direct zero-shot transfer, the current evidence remains insufficient. Further enhancement through transfer learning, equipment operating-state features, and online distribution-drift detection is therefore warranted.

Although the primary dataset contains one year of hourly observations from a university integrated energy system, the proposed framework was additionally evaluated on the independent HEEW 2022 dataset to examine its cross-dataset applicability. The two datasets differ in regional characteristics, operating conditions, load magnitudes, seasonal patterns, and the statistical relationships between energy demand and exogenous variables. The external validation demonstrates that the proposed model can maintain relatively high point-forecasting accuracy and near-nominal probabilistic coverage under a different data environment. Nevertheless, the observed performance degradation under several seasonal boundary conditions also indicates that distribution shifts cannot be completely eliminated by a fixed model configuration. Therefore, the results should be interpreted as evidence of cross-dataset transferability rather than universal generalization.

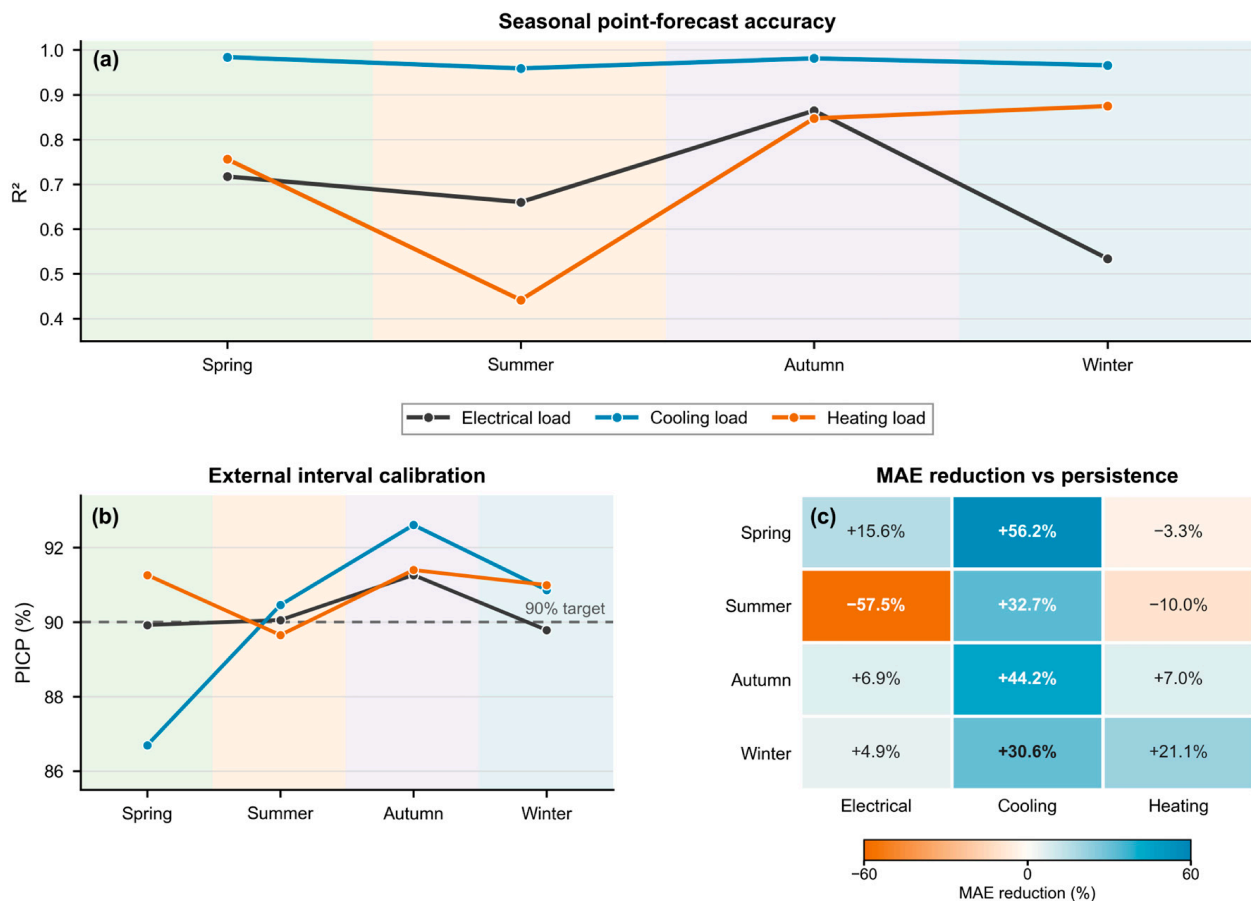


Figure 10. Seasonal model applicability on the external HEEW 2022 dataset. (a) Seasonal R^2 for the three loads; (b) PICP of the 90% prediction intervals; (c) MAE reduction relative to the persistence baseline.

In practical applications, the framework can potentially be extended to commercial, residential, and industrial integrated energy systems because its input structure and forecasting architecture are not inherently restricted to university systems. However, different application sectors may exhibit distinct load-generation mechanisms. Commercial systems are generally influenced by occupancy and operating schedules, residential systems may contain stronger behavioral stochasticity, while industrial systems can be strongly affected by production schedules and process operating states. These differences may lead to substantial distribution shifts. Therefore, practical deployment should incorporate local data preprocessing, model retraining, and rolling calibration using system-specific historical observations. Broader multi-site validation involving commercial, residential, and industrial integrated energy systems is therefore required to further establish the robustness and generalizability of the proposed framework. In addition, the difference in hyperparameter tuning budgets between the proposed model and baseline models is explicitly clarified. The superior forecasting performance of the IOOA-optimized framework is verified to stem from the structural advantages of the integrated model and algorithm optimization mechanism, rather than excessive manual or automated tuning investment.

3.8. Ablation Experiment Analysis

To investigate the actual contributions of the core modules to the probabilistic interval prediction performance of the proposed model, six model variants were constructed by individually removing or combining IOOA, TCN, and BiLSTM while retaining the quantile regression (QR) output layer, as presented in Table 10. Specifically, BiLSTM-QR and

TCN-QR were used to establish the baseline performance of individual temporal modeling architectures; TCN-BiLSTM-QR was employed to examine the synergistic effect of convolution-based feature extraction and bidirectional temporal modeling; IOOA-BiLSTM-QR and IOOA-TCN-QR were designed to evaluate the optimization effect of IOOA on different base networks; and IOOA-TCN-BiLSTM-QR represents the complete proposed model. All models adopted the same four-season rolling expanding-window forecasting scheme described in Section 3.1, with a 24 h input window and a 1 h forecasting horizon. The models simultaneously output the 0.05, 0.50, and 0.95 quantiles and employed the same 168 h causal rolling calibration strategy. Each seasonal test month contained 744 hourly samples, and the random seed was fixed at 100. For models without IOOA, unified pre-defined hyperparameters were adopted, whereas models incorporating IOOA used their corresponding optimized configurations. Therefore, the comparison was designed to isolate the effects arising from model architecture and hyperparameter optimization, thereby providing a controlled basis for quantifying the contribution of each core component to probabilistic interval prediction performance.

Table 10. Ablation Experiment Design.

Model	TCN	BiLSTM	IOOA	QR	Purpose of the Experiment
BiLSTM-QR	×	✓	×	✓	Evaluate the temporal modeling capability of a single recurrent network
TCN-QR	✓	×	×	✓	Evaluate the temporal feature extraction capability of a single convolutional network
TCN-BiLSTM-QR	✓	✓	×	✓	Examine the synergistic effect of TCN and BiLSTM
IOOA-BiLSTM-QR	×	✓	✓	✓	Evaluate the optimization effect of IOOA on BiLSTM
IOOA-TCN-QR	✓	×	✓	✓	Evaluate the optimization effect of IOOA on TCN
IOOA-TCN-BiLSTM-QR	✓	✓	✓	✓	Complete model

As shown in Figure 11, the arithmetic mean results for the four test periods—spring, summer, autumn, and winter—are reported separately for each load type. PICP is evaluated against the nominal coverage level of 90%, with values closer to 90% indicating better calibration. When coverage reliability is comparable, lower PINAW, CWC, and MPICD values indicate better performance. Because electricity, cooling, and heating loads have different units and scales, MPICD is compared only within the same load type and is not used for cross-load ranking. Across the 12 seasonal load scenarios, the complete model achieves an average PICP of 90.50%, with a mean absolute deviation of only 2.16 percentage points from the nominal coverage level, both representing the best performance among the six models. Meanwhile, its average PINAW is 0.1513, also the lowest among all models. Compared with the unoptimized TCN-BiLSTM-QR, IOOA reduces the average PINAW from 0.1610 to 0.1513, corresponding to a reduction of 6.02%, while decreasing the average PICP from 91.76% to 90.50%, bringing the coverage closer to the target level. These results indicate that the primary advantage of the complete model does not arise from simply widening the prediction intervals to achieve higher coverage, but rather from improving interval sharpness while maintaining appropriate nominal coverage.

For electricity load, IOOA-TCN-BiLSTM-QR achieves a PICP of 89.65%, a PINAW of 0.1136, and an MPICD of 532.47 kW, with both PINAW and MPICD being the best among the six models, while its PICP deviates from the nominal 90% level by only 0.35 percentage points. Compared with TCN-BiLSTM-QR, the complete model reduces PINAW and MPICD by 13.91% and 7.64%, respectively, indicating that IOOA optimization effectively eliminates redundant safety margins, while the combination of TCN and BiLSTM facilitates the simultaneous characterization of intraday local fluctuations and cross-period temporal

dependencies. This advantage is particularly relevant to high electricity-load plateaus, rapid load ramps, and demand-side capacity management during summer. Narrower prediction intervals with smaller center deviations can reduce unnecessary procurement and reserve-capacity requirements while still retaining appropriate risk margins for peak-load excursions. The cooling-load results reveal a more distinct structural preference. The complete model achieves a PICP of 89.99%, which is the closest to the nominal 90% coverage level among the six models. IOOA-TCN-QR achieves the lowest PINAW (0.1062) and MPICD (178.16 ton), indicating that cooling loads are more strongly influenced by short-term variations, equipment start-stop operations, and weather-related factors, for which the optimized TCN can extract most of the relevant local information. The complete model obtains a CWC of 0.4696, higher than the 0.1636 achieved by BiLSTM-QR. This difference is primarily attributable to the exponential penalty triggered by under-coverage during the spring period rather than to generally wider prediction intervals throughout the year. Therefore, the complete model is more suitable for integrated scheduling applications where stable coverage constraints are prioritized. If the operational objective places greater emphasis on narrow intervals and low center deviation under normal operating conditions, IOOA-TCN-QR can serve as a lightweight alternative. For heating load, IOOA-TCN-QR achieves the best PICP of 90.12% and the lowest PINAW of 0.2251, indicating favorable coverage calibration and interval sharpness. However, its CWC reaches 1.9052, suggesting severe under-coverage during certain seasons, particularly during the winter-spring transition. This finding demonstrates that seasonal average PICP may conceal structural risks associated with specific operating periods. TCN-BiLSTM-QR achieves the lowest MPICD of 0.206 MMBtu, whereas the complete model maintains the PICP at 91.87% and reduces PINAW to 0.2288, representing a further 2.80% reduction in interval width compared with the unoptimized combined model. These results indicate that BiLSTM contributes to modeling the persistence and lagged characteristics of heating loads, while IOOA helps reduce excessively conservative interval widths. Therefore, for boiler and heat-pump reserve-capacity assessment, cross-seasonal average coverage alone is insufficient, and exceedance risks during transitional seasons should be examined separately.

From the perspective of module contributions, IOOA consistently reduces the overall average PINAW across the corresponding BiLSTM, TCN, and TCN-BiLSTM structures, with reductions of 3.97%, 7.31%, and 6.02%, respectively. This demonstrates that hyper-parameter optimization provides a consistent improvement in interval sharpness. TCN primarily enhances the extraction of local abrupt variations and multi-scale periodic patterns, whereas BiLSTM complements this capability by capturing cross-period temporal dependencies. However, the ranking based on CWC does not completely coincide with that based on PINAW. The complete model achieves an average CWC of 0.3581, higher than the 0.2001 obtained by BiLSTM-QR, mainly because CWC applies an exponential penalty to insufficient coverage, causing the spring operating-condition transition to exert a disproportionate influence on the overall score. Overall, IOOA-TCN-BiLSTM-QR achieves the most stable overall balance between coverage calibration and interval sharpness, particularly for electricity-load forecasting. For cooling and heating loads, however, the appropriate safety margin should still be selected according to seasonal operating conditions, with additional consideration given to load-change detection, calendar effects, and equipment start-stop information during the spring transition period. These findings further demonstrate that ablation conclusions should be drawn from the joint assessment of PICP, PINAW, CWC, and MPICD rather than from any single metric, and that no individual module can be claimed to be universally superior across all load types and operating scenarios.

An additional implementation-level ablation experiment was conducted to isolate the effects of the two improvement mechanisms introduced into IOOA, namely Logistic

chaotic initialization and the Lévy flight strategy. Rather than using a nested three-stage design, the experiment adopted a four-condition ablation design comprising: (1) the original OOA without either improvement; (2) OOA with only Logistic chaotic initialization; (3) OOA with only the Lévy flight strategy; and (4) the complete IOOA incorporating both Logistic chaotic initialization and Lévy flight. This design enables the individual contribution of each mechanism and their combined effect to be examined under consistent experimental conditions.

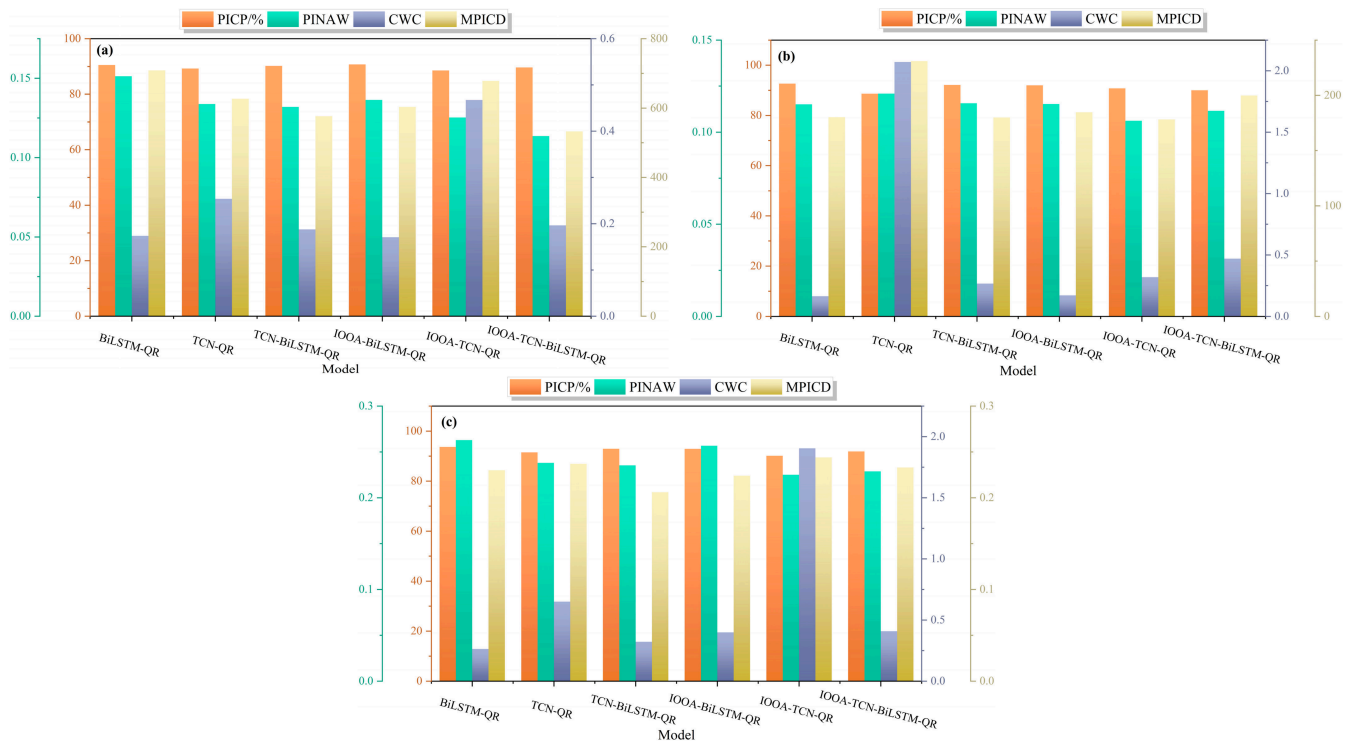


Figure 11. Results of ablation experiments for different load types. (a) Electrical load; (b) cooling load; (c) heating load.

The forecasting fitness surface comprised 27 combinations of one to three TCN residual blocks, 40 to 56 convolutional channels, and 48 to 80 BiLSTM hidden units. All other settings were fixed. Each candidate configuration was evaluated using the mean standardized validation pinball loss over the four seasonal folds with a model-training seed of 100. To assess the search reliability of the four optimizer variants under the same population size and iteration limit, each variant was independently repeated 5000 times using identical initial seeds, a population of four, and four iterations. The resulting mean pinball loss and optimum-hit rate were used to compare the search performance and reliability of the four conditions.

Table 11 presents the four-condition ablation results for the two core improvement mechanisms embedded in IOOA: Logistic chaotic initialization and the Lévy flight strategy. The original OOA, without either improvement mechanism, achieves a mean pinball loss of 0.01982014 and an optimum-hit rate of 59.50%, and is therefore used as the reference condition. When Logistic chaotic initialization is introduced alone, the mean pinball loss decreases slightly to 0.01982008, while the optimum-hit rate increases to 64.84%, corresponding to a gain of 5.34 percentage points over the original OOA. When Lévy flight is introduced alone, the mean pinball loss decreases to 0.01981972 and the optimum-hit rate increases to 62.18%, corresponding to a gain of 2.68 percentage points. When both mechanisms are incorporated, the complete IOOA achieves the lowest mean pinball loss of 0.01981935 and the highest optimum-hit rate of 67.66%, representing an 8.16 percentage-

point improvement over vanilla OOA. These four experimental conditions allow the individual and combined effects of the two mechanisms to be distinguished under the same search budget and evaluation procedure. The results indicate that both Logistic chaotic initialization and Lévy flight contribute positively to the search performance, while their combination provides the strongest overall search reliability among the four tested variants. Although the differences in mean pinball loss are numerically small, the consistent improvement in optimum-hit rate provides additional evidence that the two mechanisms mainly enhance the reliability and stability of the hyperparameter search. Logistic chaotic initialization increases the diversity of the initial population and improves the coverage of the search space, whereas Lévy flight introduces long-range search movements that can strengthen exploration and reduce the tendency toward premature convergence. The complementary roles of these two mechanisms contribute to the improved search capability of IOOA for the considered hyperparameter optimization task.

Table 11. Internal Ablation of the Two Implemented IOOA Improvement Mechanisms.

Variant	Logistic Initialization	Lévy Flight	Mean Pinball Loss	Optimum-Hit Rate (%)	Gain vs. OOA (p.p.)
OOA	×	×	0.01982014	59.50	+0.00
OOA + Logistic	✓	×	0.01982008	64.84	+5.34
OOA + Logistic + Lévy	✓	✓	0.01981935	67.66	+8.16
OOA + Lévy	×	✓	0.01981972	62.18	+2.68

3.9. Feature Contribution Analysis

To improve the interpretability of the proposed IOOA–TCN–BiLSTM–QR forecasting framework, a surrogate-model-based SHAP analysis was conducted on the four-season rolling test samples. It should be emphasized that SHAP was not applied directly to the original IOOA–TCN–BiLSTM–QR architecture, which takes continuous 24 h multivariate sequences as inputs. Instead, the SHAP analysis was performed on separate gradient-boosting surrogate models constructed to approximate the median ($q = 0.50$) forecasts generated by the original forecasting model. Because the original model operates on sequential multivariate inputs, several aggregated features were first extracted from the 24 h input window preceding each forecasting time point result analysis. These features included the latest value, 24 h mean, 24 h standard deviation, and 24 h change, which were used to characterize the recent load level, average state, fluctuation intensity, and short-term change, respectively. The corresponding ($q = 0.50$) forecasts generated by the original IOOA–TCN–BiLSTM–QR model were then used as the target outputs for training separate gradient-boosting surrogate models for electricity, cooling, and heating loads. TreeSHAP was subsequently applied to these surrogate models to quantify the contribution of the aggregated input features to the surrogate predictions. Features representing the same physical variable were further grouped to obtain 17 interpretable feature groups. The surrogate models were evaluated using temporally ordered hold-out samples within each season, with the final 20% of test time points in each season reserved exclusively for fidelity assessment. The surrogate models achieved (R^2) values of 0.9699, 0.9762, and 0.9752 for electricity, cooling, and heating loads, respectively, when reproducing the median forecasts of the original model, with corresponding MAEs of 1821.80 kW, 426.33 ton, and 0.173 MMBtu. These relatively high fidelity values indicate that the surrogate models can reproduce a large proportion of the input–output variation in the original model within the evaluated samples. However, these results should be interpreted as surrogate-model fidelity rather than forecasting accuracy. Moreover, high surrogate fidelity does not imply that the surrogate model is mathematically equivalent to the original TCN–BiLSTM–QR

architecture. Therefore, the SHAP results presented below should be understood as an interpretable approximation of the dominant relationships encoded in the original model outputs, rather than a direct decomposition of the internal nonlinear representations of the original deep-learning model.

Figure 12 presents the ten feature groups with the highest SHAP contributions for each load type. The horizontal coordinate represents the positive or negative contribution of a feature to an individual surrogate prediction, while the color gradient indicates increasing feature values. A symmetric logarithmic scale was adopted for the horizontal axis to retain extreme samples associated with seasonal operating-regime transitions while clearly displaying the dense distribution of contributions around zero. No extreme observations were removed during the analysis. Figure 12 shows that recent historical load information has the largest contribution to the surrogate predictions for all three load types. The normalized mean absolute SHAP contributions associated with the previous-hour load reach 70.01%, 68.04%, and 63.77% for electricity, cooling, and heating load forecasts, respectively. Higher historical load values are generally associated with positive SHAP contributions, whereas lower values tend to be associated with negative contributions. These results suggest that the input–output relationship learned by the original forecasting model can be well approximated, in the aggregated feature space, by a strong short-term persistence relationship. However, because these SHAP values are obtained from the surrogate models, they should not be interpreted as a direct measurement of the internal feature attribution of the original TCN–BiLSTM–QR network. The secondary feature contributions differ among the three load types. For electricity load, precipitable water, solar zenith angle, and month show relatively high contributions, with normalized values of 9.58%, 8.33%, and 5.41%, respectively. For cooling load, solar zenith angle, month, and hour contribute 14.47%, 7.53%, and 3.44%, respectively. For heating load, solar zenith angle, hour, and precipitable water contribute 18.83%, 6.07%, and 5.20%, respectively. These results indicate that, within the surrogate representation, environmental and temporal variables provide complementary information beyond recent load states. The cross-energy historical variables also show non-negligible contributions: historical cooling load contributes 1.16% to electricity-load forecasting, historical electricity load contributes 1.22% to cooling-load forecasting, and historical heating load contributes 0.53% to heating-load forecasting. These relatively small contributions indicate that cross-energy information can offer supplementary cues for approximating model outputs, while the forecasts are still dominated by the temporal dynamics of the target load series itself.

Figure 13 further illustrates representative SHAP relationships obtained from the surrogate models. Figure 13a shows the normalized global contributions, while Figure 13b–d present the SHAP dependence relationships for previous-hour electricity load, temperature and cooling load, and temperature and heating load, respectively. The results show that the surrogate models reproduce several physically interpretable patterns in the outputs of the original forecasting model. For example, when the previous-hour electricity load remains within its normal operating range, its SHAP contribution generally fluctuates around zero. When the load enters the high-demand summer operating regime, its contribution becomes increasingly positive, indicating that the original model’s median forecast tends to increase with the recent electricity-load state. Similarly, the surrogate SHAP dependence indicates that higher temperatures are generally associated with larger positive contributions to cooling-load forecasts, whereas lower temperatures tend to be associated with larger positive contributions to heating-load forecasts. These relationships provide an interpretable approximation of how recent load states and environmental conditions are associated with the median forecasts generated by the original IOOA–TCN–BiLSTM–QR model. Nevertheless, they should not be interpreted as direct causal effects or as a complete explanation of the internal decision process of the original deep-learning architecture. The original model

receives multivariate sequential inputs and extracts nonlinear temporal representations through TCN and BiLSTM layers, whereas the surrogate models operate on a reduced set of aggregated statistical features. Consequently, some temporal interactions, nonlinear dependencies, and internal representations learned by the original network may not be fully preserved in the surrogate feature space.

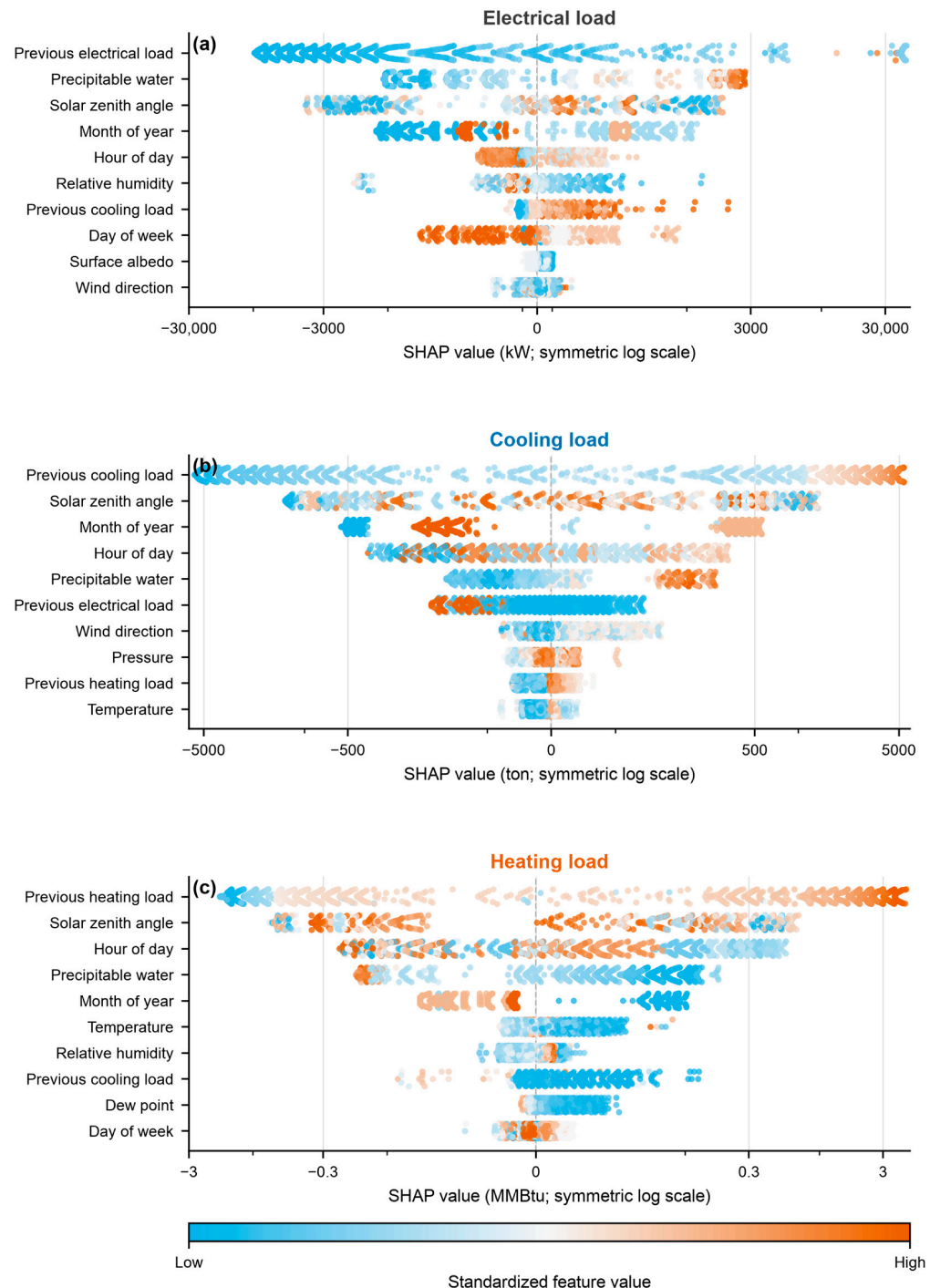


Figure 12. Grouped SHAP beeswarm plots for the three load forecasts. (a) Electrical load; (b) cooling load; (c) heating load.

Therefore, the SHAP analysis in this study is primarily intended to provide model-output-level interpretability and an approximate characterization of the dominant information associated with the original model predictions, rather than an exact feature attribution of the original neural network. The relatively high fidelity of the surrogate models supports

the usefulness of this approximation for the evaluated samples, but the transferability of these interpretations beyond the surrogate feature space and the evaluated operating conditions remains limited. Future work could further investigate direct attribution methods designed specifically for sequential deep-learning architectures, such as sequence-level perturbation or gradient-based attribution, to provide a more direct interpretation of the original TCN–BiLSTM–QR model.

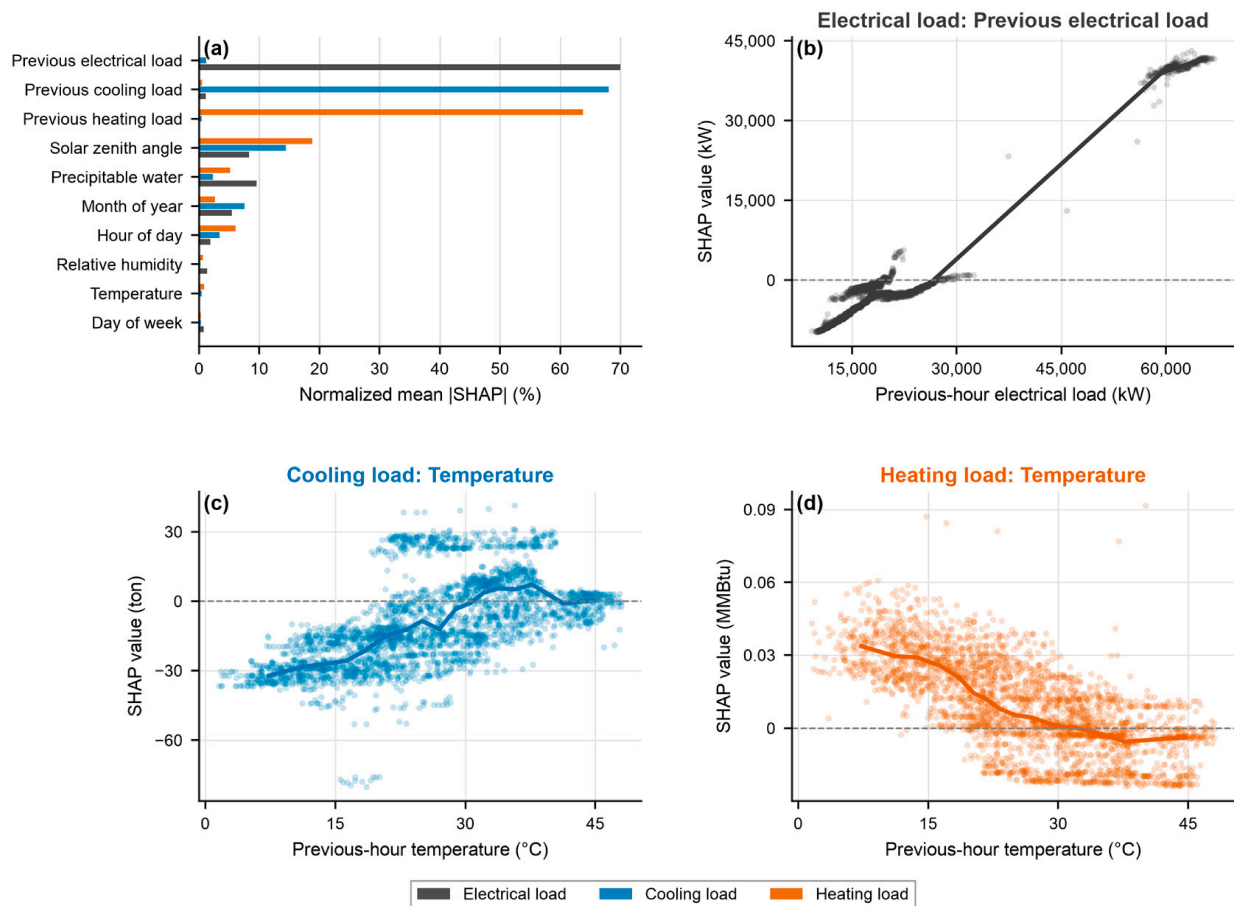


Figure 13. Cross-load global contributions and key SHAP dependence relationships. (a) Normalized mean absolute SHAP values; (b) effect of previous-hour electrical load; (c) temperature effect on cooling load; (d) temperature effect on heating load.

4. Discussion and Future Outlook

4.1. Discussion

The proposed IOOA-TCN-BiLSTM-QR model is designed to address the strong fluctuations, complex temporal dependencies, and difficulties in quantifying prediction uncertainty associated with electricity, cooling, and heating loads in integrated energy systems (IESs). Experimental results demonstrate that the proposed model achieves favorable overall performance across all three load types while maintaining high point-prediction accuracy and 90% interval coverage that is close to the nominal level on average. The performance improvement can be attributed primarily to the synergistic effects of the individual modules. Specifically, TCN employs causal and dilated convolutions to extract multi-scale local temporal features, while BiLSTM further captures long-term temporal dependencies. IOOA adaptively optimizes the key hyperparameters, thereby improving the prediction accuracy and stability of the overall model. The ablation experiments further confirm the effectiveness and complementarity of these modules. Furthermore, the incorporation of quantile regression (QR) extends the model from conventional deterministic

forecasting to probabilistic interval prediction. The prediction uncertainty is quantitatively characterized using PICP, PINAW, CWC, and MPICD from the perspectives of coverage reliability, interval width, and overall interval performance. Compared with conventional methods that provide only a single deterministic prediction, the proposed framework can simultaneously characterize the expected future levels of electricity, cooling, and heating loads and their potential fluctuation ranges, thereby providing more comprehensive predictive information for risk-aware operation and optimal scheduling of integrated energy systems. Considering the differences among the three load types in terms of periodicity, volatility, and external influencing factors, the proposed model provides a unified temporal feature-learning framework for effectively characterizing their distinct variation patterns. From a practical perspective, IOOA-TCN-BiLSTM-QR can support short-term operational scheduling and energy allocation in integrated energy systems. The point prediction results can assist in formulating energy supply plans, while the upper and lower bounds of the prediction intervals can provide valuable information for high-load risk identification, energy storage configuration, and the scheduling of cooling and heating source equipment, thereby improving the operational reliability and flexibility of integrated energy systems.

To further illustrate how the predicted intervals can support practical operational decisions, a simple numerical example is provided. Suppose that for a certain future hour, the proposed model predicts an electricity load of 50,000 kW at the 0.50 quantile, with a 90% prediction interval of 47,000–53,000 kW (i.e., the 0.05 and 0.95 quantiles). If the operator considers only the point forecast, the planned electricity supply capacity may be set close to 50,000 kW. However, when the prediction interval is explicitly considered, the upper bound indicates that the actual demand may reach approximately 53,000 kW under the modeled uncertainty. Therefore, an additional reserve of approximately 3000 kW can be considered for short-term scheduling, energy procurement, or battery discharge capacity. Conversely, the lower bound of 47,000 kW indicates that maintaining the full upper-bound capacity at all times may lead to unnecessary resource allocation when demand is lower than expected. Thus, the prediction interval provides a quantitative basis for balancing operational reliability against reserve costs rather than relying solely on a single deterministic forecast.

A limitation of the present benchmark design is that the compared models were not subjected to identical hyperparameter-search budgets. The proposed IOOA-TCN-BiLSTM-QR framework uses adaptive optimization over an eight-dimensional search space, whereas the baseline neural models were evaluated using fixed configurations. Therefore, the reported performance improvements represent the combined benefit of model architecture and adaptive configuration optimization, rather than an isolated measurement of architectural superiority. A fully budget-matched comparison, in which all competing models receive the same number of optimization evaluations and comparable training resources, would provide a more controlled assessment of architecture-specific contributions and is left for future work.

4.2. Future Outlook

Although the proposed IOOA-TCN-BiLSTM-QR model achieves favorable performance in probabilistic interval forecasting of electricity, cooling, and heating loads, several limitations remain. First, the model relies primarily on historical load data for forecasting. Although it can extract complex temporal features from multivariate time series, its consideration of external factors such as weather conditions, energy prices, user behavior, and equipment operating states remains relatively limited. In practical integrated energy systems, variations in external environmental conditions may have a substantial impact on cooling and heating loads. Therefore, relying solely on historical load information may

constrain the model's predictive capability under extreme weather and abnormal operating conditions. Second, a unified TCN-BiLSTM architecture is employed to forecast electricity, cooling, and heating loads. Although this design can effectively capture common variation patterns across different energy loads, these loads exhibit distinct periodic characteristics and dynamic response mechanisms. Future studies could investigate load-specific feature extraction mechanisms and cross-energy attention mechanisms to preserve load-specific characteristics while exploiting shared information, thereby further improving the model's ability to characterize complex multi-energy coupling relationships. Third, although IOOA can effectively optimize the model hyperparameters, metaheuristic optimization algorithms generally require multiple rounds of model training and fitness evaluations. Consequently, their computational cost may increase with the growth of dataset size and model complexity. For integrated energy systems requiring real-time updating and online forecasting, further improvements in model training efficiency are therefore necessary [49]. Future research could explore lightweight network architectures, incremental learning, and on-line adaptive optimization strategies to reduce the computational burden associated with repeated model training and improve deployment efficiency. In addition, the current QR head outputs only the 0.05, 0.50, and 0.95 quantiles, which define a central 90% prediction interval but do not reconstruct the full predictive distribution. Accordingly, PICP, PINAW, CWC, and MPICD evaluate interval-specific properties such as empirical coverage, width, the coverage-width trade-off, and center deviation; they are not full-distribution calibration tests. Future studies should estimate a denser quantile grid or an explicit conditional density and assess distributional calibration using quantile reliability diagrams, probability integral transform diagnostics, and proper scoring rules such as the continuous ranked probability score. Future research will focus on the following five directions: (1) incorporating multi-source external information, including weather conditions, energy prices, user behavior, and equipment operating states, to develop more comprehensive multimodal load forecasting models; (2) integrating graph neural networks, attention mechanisms, or Transformer architectures to further explore the dynamic spatiotemporal coupling relationships among electricity, cooling, and heating loads [28]; (3) combining lightweight network architectures with online learning strategies to improve real-time forecasting capability in large-scale integrated energy systems; (4) further improving probabilistic forecasting and uncertainty calibration mechanisms to enhance the reliability of prediction intervals under extreme load conditions and abnormal operating states [33]; and (5) integrating probabilistic load forecasting results with energy optimization scheduling and risk-aware decision-making to extend the research from high-accuracy load forecasting toward coordinated forecasting–decision frameworks. Overall, future research should focus not only on further improving predictive accuracy but also on promoting deeper integration among multi-energy coupling mechanisms, uncertainty quantification, computational efficiency, and practical scheduling decisions, thereby facilitating the transition of probabilistic load forecasting models from experimental validation toward practical applications in integrated energy systems [7,17,24].

5. Conclusions

This study developed an Improved Osprey Optimization Algorithm-based TCN-BiLSTM-QR (IOOA-TCN-BiLSTM-QR) framework for short-term probabilistic forecasting of electricity, cooling, and heating loads in integrated energy systems. The main conclusions are summarized as follows.

- (1) The proposed IOOA, incorporating Logistic chaotic mapping and Lévy flight, effectively improves the optimization performance of the original OOA. Benchmark experiments on six 30-dimensional functions demonstrate enhanced convergence speed,

optimization accuracy, and search stability compared with PSO, GWO, WOA, and OOA. The improved optimization capability provides effective support for adaptive hyperparameter selection in the proposed forecasting framework.

- (2) The proposed model achieves accurate point forecasts and 90% prediction intervals whose average empirical coverage is close to the nominal level across the four seasonal test periods. The average PICP and PINAW are 90.50% and 0.1513, respectively. The test-set R^2 values for electricity, cooling, and heating loads are 0.9385, 0.9781, and 0.9731, with corresponding MAPE values of 2.6117%, 2.6712%, and 2.3600%. Comparative and ablation experiments confirm the complementary contributions of IOOA, TCN, and BiLSTM to forecasting accuracy and interval quality. However, abrupt changes in operating conditions can temporarily cause interval under-coverage, indicating that further improvement in adaptive interval calibration is needed for sharp load transitions.
- (3) Independent-dataset validation further demonstrates that the proposed framework maintains satisfactory forecasting performance under unseen load conditions, providing supporting evidence of its applicability across different datasets and load types. The SHAP analysis indicates that recent historical load information is the primary driver of the forecasts, while meteorological, solar-geometry, and calendar-related variables provide complementary information for capturing variations in cooling and heating demand. Overall, the proposed framework provides both accurate load forecasts and quantitative uncertainty information, offering a useful basis for short-term load monitoring, reserve-capacity assessment, and risk-aware operational decision-making in integrated energy systems.

Author Contributions: Conceptualization, Y.Z. and X.X.; methodology, Y.Z., X.X. and R.Z.; software, Y.Z. and X.X.; validation, Y.Z., X.X., L.X. and M.W.; formal analysis, Y.Z., X.X. and M.W.; investigation, Y.Z. and X.X.; data curation, X.X. and L.X.; data preprocessing, X.X. and L.X.; model development, Y.Z. and X.X.; algorithm design and optimization, Y.Z., X.X. and R.Z.; experimental design, Y.Z., X.X. and M.W.; experiments, Y.Z. and X.X.; performance evaluation, Y.Z., X.X. and M.W.; result analysis and interpretation, Y.Z., X.X., M.W. and R.Z.; visualization, Y.Z. and X.X.; writing—original draft preparation, Y.Z. and X.X.; writing—review and editing, M.W., L.X., H.W. and R.Z.; resources, R.Z.; supervision, H.W. and R.Z.; project administration, R.Z.; funding acquisition, R.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (52279005) and National Key Research and Development Program of China (2023YFC3205600).

Data Availability Statement: The data supporting the findings of this study are not publicly available due to data confidentiality and access restrictions. The data may be available from the corresponding authors upon reasonable request and with permission from the data provider.

Acknowledgments: The authors have no additional acknowledgments to declare.

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

Abbreviations

The following abbreviations are used in this manuscript:

IES	Integrated Energy System
IOOA	Improved Osprey Optimization Algorithm
OOA	Osprey Optimization Algorithm
TCN	Temporal Convolutional Network
BiLSTM	Bidirectional Long Short-Term Memory

QR	Quantile Regression
PICP	Prediction Interval Coverage Probability
PINAW	Prediction Interval Normalized Average Width
CWC	Coverage Width-based Criterion
MPICD	Mean Prediction Interval Coverage Deviation
R ²	Coefficient of Determination
MAPE	Mean Absolute Percentage Error
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
SHAP	Shapley Additive Explanations
PSO	Particle Swarm Optimization
GWO	Grey Wolf Optimizer
WOA	Whale Optimization Algorithm
BiLSTM-Attention	Bidirectional Long Short-Term Memory with Attention
MTL-LSTM	Multi-Task Learning Long Short-Term Memory
STGCN	Spatio-Temporal Graph Convolutional Network
MMEMformer	Multi-Modal Energy Management Transformer
AM-RU-BiLSTM	Attention Mechanism-Robust Unit-Bidirectional Long Short-Term Memory
CEEMDAN	Complete Ensemble Empirical Mode Decomposition with Adaptive Noise
MKELM	Multiple Kernel Extreme Learning Machine
SP-RF-ARIMA	Seasonal Period Random Forest-AutoRegressive Integrated Moving Average
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
HEEW	Hierarchical Energy, Emissions, and Weather

References

1. Al-Shetwi, A.Q. Sustainable development of renewable energy integrated power sector: Trends, environmental impacts, and recent challenges. *Sci. Total Environ.* **2022**, *822*, 153645. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Wang, K.; Wang, C.; Yao, W.; Zhang, Z.; Liu, C.; Dong, X.; Yang, M.; Wang, Y. Embedding P2P transaction into demand response exchange: A cooperative demand response management framework for IES. *Appl. Energy* **2024**, *367*, 123319. [\[CrossRef\]](#)
3. Song, D.; Meng, W.; Dong, M.; Yang, J.; Wang, J.; Chen, X.; Huang, L. A critical survey of integrated energy system: Summaries, methodologies and analysis. *Energy Convers. Manag.* **2022**, *266*, 115863. [\[CrossRef\]](#)
4. Xiao, L.; Shao, W.; Liang, T.; Wang, C. A combined model based on multiple seasonal patterns and modified firefly algorithm for electrical load forecasting. *Appl. Energy* **2016**, *167*, 135–153. [\[CrossRef\]](#)
5. Huang, N.; Wang, Y.; Li, B.; Cai, G.; Zhang, L. Heterogeneous ensemble short-term integrated load forecasting considering differentiated cycle-trend properties. *Eng. Appl. Artif. Intell.* **2025**, *156*, 111198. [\[CrossRef\]](#)
6. Jiang, W.; Xie, X. Short-term load forecasting based on the improved bi-directional long short-term memory model. *Int. J. Electr. Power Energy Syst.* **2026**, *177*, 111789. [\[CrossRef\]](#)
7. Yildiz, B.; Bilbao, J.I.; Sproul, A.B. A review and analysis of regression and machine learning models on commercial building electricity load forecasting. *Renew. Sustain. Energy Rev.* **2017**, *73*, 1104–1122. [\[CrossRef\]](#)
8. Kuster, C.; Rezgui, Y.; Mourshed, M. Electrical load forecasting models: A critical systematic review. *Sustain. Cities Soc.* **2017**, *35*, 257–270. [\[CrossRef\]](#)
9. Zhang, R.; Chen, Y.; Li, Z.; Jiang, T.; Li, X. Two-stage robust operation of electricity-gas-heat integrated multi-energy microgrids considering heterogeneous uncertainties. *Appl. Energy* **2024**, *371*, 123690. [\[CrossRef\]](#)
10. Magalhães, B.; Bento, P.; Pombo, J.; Calado, M.d.R.; Mariano, S. Short-term load forecasting based on optimized random forest and optimal feature selection. *Energies* **2024**, *17*, 1926. [\[CrossRef\]](#)
11. Niu, D.; Wang, Y.; Wu, D.D. Power load forecasting using support vector machine and ant colony optimization. *Expert Syst. Appl.* **2010**, *37*, 2531–2539. [\[CrossRef\]](#)
12. Abumohsen, M.; Owda, A.Y.; Owda, M. Electrical load forecasting using LSTM, GRU, and RNN algorithms. *Energies* **2023**, *16*, 2283. [\[CrossRef\]](#)
13. Luo, S.; Wang, B.; Gao, Q.; Wang, Y.; Pang, X. Stacking integration algorithm based on CNN-BiLSTM-Attention with XGBoost for short-term electricity load forecasting. *Energy Rep.* **2024**, *12*, 2676–2689. [\[CrossRef\]](#)
14. Peng, D.; Liu, Y.; Wang, D.; Zhao, H.; Qu, B. Multi-energy load forecasting for integrated energy system based on sequence decomposition fusion and factors correlation analysis. *Energy* **2024**, *308*, 132796. [\[CrossRef\]](#)

15. Liu, G.; Yu, J.; Luo, X.; Del Pero, C.; Zhao, S.; Huang, W.; Zhao, X. An integrated method for electric load forecasting in large public buildings based on load patterns clustering and multi-scale temporal feature selection. *Energy* **2026**, *347*, 140438. [\[CrossRef\]](#)
16. Guo, Y.; Li, Y.; Qiao, X.; Zhang, Z.; Zhou, W.; Mei, Y.; Lin, J.; Zhou, Y.; Nakanishi, Y. BiLSTM multitask learning-based combined load forecasting considering the loads coupling relationship for multienergy system. *IEEE Trans. Smart Grid* **2022**, *13*, 3481–3492. [\[CrossRef\]](#)
17. Zhao, P.; Hu, W.; Cao, D.; Zhang, Z.; Huang, Y.; Dai, L.; Chen, Z. Probabilistic multienergy load forecasting based on hybrid attention-enabled transformer network and gaussian process-aided residual learning. *IEEE Trans. Ind. Inform.* **2024**, *20*, 8379–8393. [\[CrossRef\]](#)
18. Tan, M.; Liao, C.; Chen, J.; Cao, Y.; Wang, R.; Su, Y. A multi-task learning method for multi-energy load forecasting based on synthesis correlation analysis and load participation factor. *Appl. Energy* **2023**, *343*, 121177. [\[CrossRef\]](#)
19. Zhuang, W.; Fan, J.; Xia, M.; Zhu, K. A multi-scale spatial-temporal graph neural network-based method of multienergy load forecasting in integrated energy system. *IEEE Trans. Smart Grid* **2023**, *15*, 2652–2666. [\[CrossRef\]](#)
20. Saravana, M.K.; Roopa, M.S.; Arunalatha, J.S.; Venugopal, K.R. Transformers for multivariate time series forecasting: Comprehensive analysis, challenges, research opportunities and future prospects. *IEEE Access* **2026**, *14*, 11424–11457. [\[CrossRef\]](#)
21. Koenker, R.; Bassett, G., Jr. Regression quantiles. *Econom. J. Econom. Soc.* **1978**, *46*, 33–50. [\[CrossRef\]](#)
22. Luo, J.; Chen, F.; Kong, L.; Liu, H. An INRBO-SSA-LSTM Hybrid Framework for Short-Term Power Load Forecasting in Smart Microgrids. *Electronics* **2026**, *15*, 3044. [\[CrossRef\]](#)
23. Quan, H.; Srinivasan, D.; Khosravi, A. Uncertainty handling using neural network-based prediction intervals for electrical load forecasting. *Energy* **2014**, *73*, 916–925. [\[CrossRef\]](#)
24. Yang, Y.; Li, S.; Li, W.; Qu, M. Power load probability density forecasting using Gaussian process quantile regression. *Appl. Energy* **2018**, *213*, 499–509. [\[CrossRef\]](#)
25. Cao, J.; Liu, C.; Chen, C.-L.; Qu, N.; Xi, Y.; Dong, Y.; Feng, R. A short-term load forecasting method for integrated community energy system based on STGCN. *Electr. Power Syst. Res.* **2024**, *232*, 110265. [\[CrossRef\]](#)
26. Wang, D.; Peng, D.; Huang, D.; Zhao, H.; Qu, B. MMEMformer: A multi-scale memory-enhanced transformer framework for short-term load forecasting in integrated energy systems. *Energy* **2025**, *322*, 135762. [\[CrossRef\]](#)
27. Dai, Y.; Yu, W. Short-term power load forecasting based on Seq2Seq model integrating Bayesian optimization, temporal convolutional network and attention. *Appl. Soft Comput.* **2024**, *166*, 112248. [\[CrossRef\]](#)
28. Wang, S.; Wang, X.; Wang, S.; Wang, D. Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting. *Int. J. Electr. Power Energy Syst.* **2019**, *109*, 470–479. [\[CrossRef\]](#)
29. Xu, T.; He, J.; Li, Y.; Li, X.; Tang, J. A VMD-Bayesian-Optimized XGBoost-BiLSTM Hybrid Model for Short-Term Load Forecasting. *Electronics* **2026**, *15*, 2507. [\[CrossRef\]](#)
30. Shi, P.; Tang, M.; Wang, Q.; Ma, X. Optimization of TCN-BiLSTM for dissolved oxygen prediction based on improved sparrow search algorithm. *Sci. Rep.* **2025**, *15*, 30790. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Lei, W.; Han, J.; Wu, X. Improved osprey optimization algorithm with multi-strategy fusion. *Biomimetics* **2024**, *9*, 670. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Zhang, Y.; Liu, P. Research on reactive power optimization based on hybrid osprey optimization algorithm. *Energies* **2023**, *16*, 7101. [\[CrossRef\]](#)
33. Yuan, Y.; Yang, Q.; Ren, J.; Mu, X.; Wang, Z.; Shen, Q.; Zhao, W. Attack-defense strategy assisted osprey optimization algorithm for PEMFC parameters identification. *Renew. Energy* **2024**, *225*, 120211. [\[CrossRef\]](#)
34. Xie, H.; Wan, C.; Quan, D.; Gui, G.; Yao, Y. Movable Antenna Enhanced Beamforming via an Improved Osprey Optimization Algorithm. *IEEE Wirel. Commun. Lett.* **2026**, *15*, 3104–3108. [\[CrossRef\]](#)
35. Ermiş, S.; Taşdemir, O.; Al-Hajj, R. Multi-objective techno-economic and environmental optimization of hydrogen-based hybrid renewable energy system using osprey optimization algorithm. *Sci. Rep.* **2026**, *16*, 15618. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Yuan, Y.; Zhao, X.; Yin, X.; Liu, Y. Optimization of tool edge preparation using magnetoelastic abrasives and development of a wear prediction model based on logistic chaotic mapping and sparrow algorithm. *Mech. Syst. Signal Process.* **2025**, *235*, 112909. [\[CrossRef\]](#)
37. Xu, X.; Mao, X.; Zou, W. An improved goose optimization algorithm (I-GOOSE) based on self-adapting attenuation factors and lévy flight. *Neurocomputing* **2026**, *671*, 132717. [\[CrossRef\]](#)
38. Shen, X.; Luo, X. An Improved Comprehensive Learning Jaya Algorithm with Lévy Flight for Engineering Design Optimization Problems. *Electronics* **2025**, *14*, 3776. [\[CrossRef\]](#)
39. Barthelemy, P.; Bertolotti, J.; Wiersma, D.S. A Lévy flight for light. *Nature* **2008**, *453*, 495–498. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Wang, X.; Hu, M.; Luo, X.; Guan, X. Spatio-temporal photovoltaic power forecasting via Bayesian-optimized dynamic graph convolutional networks with temporal convolutional networks. *J. Clean. Prod.* **2026**, *542*, 147641. [\[CrossRef\]](#)
41. Shi, J.; Teh, J.; Lai, C.-M. Wind power prediction based on improved self-attention mechanism combined with Bi-directional Temporal Convolutional Network. *Energy* **2025**, *322*, 135666. [\[CrossRef\]](#)

42. Mahmoud, A.; Mohammed, A. Enhancing time series forecasting: A hybrid TCN-transformer approach. *Neural Comput. Appl.* **2026**, *38*, 342. [[CrossRef](#)]
43. Fei, W.; Dai, W.; Li, C.; Zou, J.; Xiong, H. On centralization and unitization of batch normalization for deep ReLU neural networks. *IEEE Trans. Signal Process.* **2024**, *72*, 2827–2841. [[CrossRef](#)]
44. Bai, J.; Zhu, W.; Liu, S.; Ye, C.; Zheng, P.; Wang, X. A temporal convolutional network–bidirectional long short-term memory (TCN-BiLSTM) prediction model for temporal faults in industrial equipment. *Appl. Sci.* **2025**, *15*, 1702. [[CrossRef](#)]
45. Sun, K.; Yang, D.; Wang, D.; Zhang, F. Development of a physics-guided bidirectional long short-term memory for wind power forecasting. *Eng. Appl. Artif. Intell.* **2026**, *167*, 113798. [[CrossRef](#)]
46. Nadimi, R.; Goto, M. A novel decision support system for enhancing long-term forecast accuracy in virtual power plants using bidirectional long short-term memory networks. *Appl. Energy* **2025**, *382*, 125273. [[CrossRef](#)]
47. Wang, S.; Cao, W.; Hu, X.; Zhong, H.; Sun, W. A selective overview of quantile regression for large-scale data. *Mathematics* **2025**, *13*, 837. [[CrossRef](#)]
48. Shen, L.; Bao, Y.; Hasan, N.; Huang, Y.; Zhou, X.; Deng, C. Intelligent crude oil price probability forecasting: Deep learning models and industry applications. *Comput. Ind.* **2024**, *163*, 104150. [[CrossRef](#)]
49. Geng, G.; He, Y.; Zhang, J.; Qin, T.; Yang, B. Short-term power load forecasting based on PSO-optimized VMD-TCN-attention mechanism. *Energies* **2023**, *16*, 4616. [[CrossRef](#)]
50. Li, T.; Qian, Z.; He, T. Short-Term Load Forecasting with Improved CEEMDAN and GWO-Based Multiple Kernel ELM. *Complexity* **2020**, *2020*, 1209547. [[CrossRef](#)]
51. Zhao, H.; Zhou, Z.; Zhang, P. Forecasting of the short-term electricity load based on WOA-BiLSTM. *Int. J. Pattern Recognit. Artif. Intell.* **2023**, *37*, 2359018. [[CrossRef](#)]
52. 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. [[CrossRef](#)]
53. Chen, Q.; Kang, J.; Peng, C.; Zhang, R.; Tian, J.; Yang, C.; Chen, A. Prediction model of spontaneous combustion temperature intervals of coal samples from multiple mining regions. *Fuel* **2026**, *425*, 139383. [[CrossRef](#)]
54. Baesmat, K.H.; Shokoohi, F.; Farrokhi, Z. SP-RF-ARIMA: A sparse random forest and ARIMA hybrid model for electric load forecasting. *Glob. Energy Interconnect.* **2025**, *8*, 486–496. [[CrossRef](#)]

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