



ARTICLE

Short-Term Electric Load Forecasting by Cross-Feature Analysis and Multimodal Selection

Li-Ling Peng¹, Tong Li¹, Guo-Feng Fan¹, Xin-Yu Yang¹ and Wei-Chiang Hong^{2,*}

¹School of Mathematics & Statistics, Ping Ding Shan University, Ping Ding Shan, China

²College of Shipbuilding Engineering, Harbin Engineering University, Harbin, China

*Corresponding Author: Wei-Chiang Hong. Email: samuelsonhong@gmail.com

Received: 09 June 2026; Accepted: 08 September 2026; Published: 21 October 2026

ABSTRACT: Accurate load forecasting has become a critical foundation for ensuring the stable operation of power systems, optimizing generation scheduling, and supporting the efficient functioning of electricity markets. In this paper, the cross-scale and meso-scale characteristics of the complexity of power loads are analyzed, and the meteorological factors, the impact of the emergence law, and the uncertainty are also considered simultaneously. The short-term coupled forecasting model of power loads based on AI technology is proposed. Firstly, the cross-scale and meso-scale characteristics of short-term power loads are explored, and the nonlinear effects between different scales are analyzed in terms of peaks and valleys. Secondly, the modal decomposition method with the Aquila optimizer algorithm is employed to perform stable decomposition and key feature extraction on the strongly nonlinear load sequence. Subsequently, the Gated Recurrent Unit is used to capture the inherent temporal dynamics and nonlinear evolution patterns in the load data. At the same time, the Bidirectional Long Short-Term Memory neural network model optimized by the Coati optimization algorithm is introduced to achieve the deep integration of meteorological features and load data, effectively handling coupled nonlinear and random uncertainties. Finally, the results for the two types of data are reconstructed. Experimental results show that the proposed model can effectively handle various nonlinear problems. The model demonstrates high accuracy and robustness across different time scales and is helpful for the power industry to better cope with power market fluctuations and changes.

KEYWORDS: Cross-scale; mesoscale; short-term power load; meteorological characteristics; nonlinear effects

1 Introduction

Short-term power load forecasting has always been a fundamental task in power system operation, and its prediction accuracy is critical to system security and economic efficiency. Electricity demand is substantial in numerous scenarios such as residential and industrial applications [1]. Nevertheless, the rising complexity of modern power grids introduces notable uncertainties into the forecasting process, among which meteorological conditions are identified as the dominant external influencing factor. Against the backdrop of frequent extreme weather events and profound social impacts induced by climate change, incorporating meteorological data into load forecasting has become an extremely important research topic in the energy sector. As the most crucial external driving factor governing electricity load evolution, meteorological conditions exhibit strong coupling with the variation patterns of residential, industrial and commercial power consumption, and directly determine the fluctuation trends as well as peak-valley characteristics of short-term loads. Therefore, exploring the time-series evolution characteristics of power loads, revealing the

nonlinear coupling relationship between meteorological factors and power loads, and establishing a short-term load forecasting model with high prediction accuracy and anti-interference capability are of important theoretical and engineering value for maintaining the safe and stable operation of power grids and improving the resource utilization efficiency of power systems.

With the continuous deepening of market-oriented reforms in the electricity market, the operation mode of power systems has become increasingly complex and exhibits obvious dynamic characteristics. Traditional load forecasting methods (such as regression analysis [2], grey system theory [3–5] and time series models [6,7]) generally have limitations in accuracy and reliability for short-term power load forecasting, which drives researchers to take neural network structure optimization as the core approach to improve forecasting performance. Under this circumstance, machine learning algorithms [8] have gradually occupied a dominant position due to their outstanding practicability. In particular, deep learning technologies show remarkable potential in short-term load forecasting tasks by virtue of end-to-end feature extraction and multi-layer nonlinear mapping, and have gradually become mainstream tools in this field. Models frequently used for load forecasting include convolutional neural networks [9,10], long short-term memory neural networks [11,12], and bidirectional long short-term memory neural networks [13,14]. Mining deep features of power load data based on machine learning and deep learning can effectively improve load forecasting results and better meet the operational requirements of modern power systems, and this direction has become a core research hotspot in power load forecasting.

The reliability of energy systems can be substantially enhanced by leveraging analytical capabilities across weather and climate big data, artificial intelligence, supercomputing, and remote sensing, provided that these tools are effectively integrated for operational management. In response, numerous researchers have conducted extensive studies on power load forecasting, focusing on signal decomposition techniques, optimization algorithms, and hybrid ensemble models.

In terms of algorithm optimization and model integration, Ren et al. [15] introduced the AO algorithm for optimization in wind power forecasting research. The results demonstrate that the Aquila Optimizer (AO) achieves better optimization performance compared with the particle swarm optimization (PSO) algorithm. Liu et al. [16] implemented improvements to the hyperparameters of BiLSTM using OOA and enhanced it with spectral attention mechanism, thereby enhancing the global search capability of the model.

In the field of multi-model ensemble research, Chen et al. [17] designed an integrated architecture combining an adaptive noise regulation mechanism, gated recurrent unit (GRU), and improved grey wolf optimizer. Experimental results show that the proposed framework achieves higher prediction accuracy and stronger robustness than traditional single-model schemes when handling complex load variation scenarios. Goswami and Kandali [18] systematically compared three typical sequence learners, namely long short-term memory (LSTM), bidirectional LSTM and GRU. They found that prediction models based on composite architectures significantly outperform single network structures. For power load forecasting tasks, Li and Zhang [19] put forward a technical route integrating GRU and equivariant Kolmogorov-Arnold operator (EKAN). The GRU module extracts spatiotemporal dynamic correlation information, while EKAN identifies intrinsic symmetry features within data, thereby effectively improving the generalization ability of the model. Targeting the requirements of specific business scenarios, Fan et al. [20] constructed a high-precision time series modeling platform consisting of empirical wavelet transform preprocessing, convolutional neural network-based feature extraction and long short-term memory network. Experiments verify that its practical performance exceeds that of existing mainstream commercial software tools. Manjili et al. [21] proposed a design scheme that embeds Transformer modules into the classic XGBoost framework and introduces residual correction terms, which achieves competitive performance across multiple public benchmark tests. Meanwhile, Niu et al. [22] established a learning pipeline that integrates temporal convolutional

network (TCN) with attention mechanism and improved fuzzy C-means clustering, which also delivers outstanding overall performance. The above studies greatly promote theoretical breakthroughs and technical upgrades for multi-source heterogeneous time series data management and decision support systems in smart grid environments.

In the research of signal decomposition and model integration, Yıldız [23] adopted variational mode decomposition (VMD) to extract key features from historical load data. With this method, a forecasting system was established whose performance outperforms various classic and state-of-the-art deep learning algorithms. Ribeiro et al. [24] introduced the VMD method into multivariate data analysis, which greatly improves the processing efficiency and prediction accuracy for complex datasets. Targeting the short-term household electrical load forecasting problem, Wu et al. [25] designed a hybrid architecture integrating wavelet threshold denoising, VMD and bidirectional long short-term memory (BiLSTM). The proposed method achieves remarkable performance improvement for household-level short-term load forecasting tasks. Yu et al. [26] put forward a novel short-term load forecasting scheme based on adaptive mode decomposition. By accurately identifying and separating stationary and non-stationary components within load fluctuations, this method substantially enhances the reliability and accuracy of short-term load forecasting.

In the research of feature representation and model integration, Huang et al. [27] proposed an innovative framework that integrates variational mode decomposition (VMD) with multi-layer bidirectional long short-term memory (BiLSTM). Relying on signal decomposition and dynamic sequence modeling capabilities, this framework deeply analyzes complex nonlinear load data. The method can not only effectively extract multi-dimensional spatiotemporal correlation information, but also significantly improve prediction accuracy. Chen et al. [28] developed a power load forecasting system based on long short-term memory (LSTM). This system fully integrates multi-source data including historical electricity consumption records, meteorological variables and environmental factors. It has been applied to real-time monitoring and early warning tasks in power grid operation and maintenance, providing important technical support for improving the stability of power supply. Aiming at short-term load forecasting, Deng et al. [29] put forward a comprehensive solution combining cluster analysis, adaptive weight adjustment mechanism, bidirectional gated recurrent unit (BiGRU) and Bayesian optimization algorithm. More accurate and reliable forecasting results are achieved through multi-level collaborative optimization. The above studies verify that the proposed methods exhibit excellent performance and broad application prospects when coping with practical scenario requirements. Segura and Meléndez [30] proposed that the Local Flexibility Market (LFM) realizes flexibility transactions at the distribution network level through market mechanisms. Its effective operation imposes clear requirements on load forecasting algorithms, which need to take into account factors such as data granularity, prediction duration, and the degree of participant aggregation. This study focuses on the load forecasting model (LFM) under the background of data infrastructure, explores its optimization routes, and deeply analyzes the potential impacts of algorithm fusion on system flexibility. Liu et al. [31] proposed that complex nonlinear systems can be effectively modeled and stabilized via variable substitution techniques, piecewise Lyapunov analysis, and adaptive intermittent control, which provides a theoretical basis for nonlinear time-series research. Neda [32] developed an integrated machine learning framework for transformer condition monitoring that fuses multi-source operational data. By combining a Support Vector Classifier, Isolation Forest, and Random Forest Regression models, the framework simultaneously accomplishes transformer fault classification, anomaly identification, and remaining useful life prediction. This multi-task machine learning system achieves commendable performance metrics in fault classification, degradation pattern capture, and lifespan estimation.

Although existing studies have greatly improved the accuracy of power load forecasting, the demand for higher precision has not yet been fully satisfied. To tackle this challenge, academia has widely adopted

comprehensive approaches such as algorithm optimization, improved feature selection, and multi-model fusion, and systematically applied them to power load forecasting research.

Kim et al. [33] used load and temperature datasets to test the SD-CNN-LSTM model. Comparative results show that this model outperforms other benchmark models in prediction accuracy. For short-term load forecasting, Kiruthiga and Manikandan [34] adopted the Levy particle swarm optimization algorithm to optimize the BiLSTM network parameters. The study indicates that the model optimized by this algorithm achieves significantly better prediction performance than PSO and other existing methods. Fan et al. [35] developed the EWT-S-ALOSVR hybrid intelligent forecasting model that integrates feature extraction with the ant lion optimizer (ALO). The coupled support vector regression (SVR) and ALO framework can capture the nonlinear characteristics and fluctuation patterns of power loads, and the optimized model achieves superior prediction accuracy and robustness compared with other models. Wang et al. [36] integrated VMD, t-SNE, LSSVM, Tent mapping and the chaotic sparrow search algorithm (CSSA) to develop a hybrid short-term load forecasting model, which achieves good fitting performance for actual load curves. Fan et al. [37] integrated the Improved Sparrow Search Algorithm (ISSA), BiTCN and LSTM to establish a forecasting model. Based on ablation experiments and multiple groups of optimization comparison analyses of the model structure, the superior prediction performance of the model was verified. Neda [38] established an LSTM-XGBoost hybrid framework based on seasonal-trend decomposition and multi-feature engineering. Meteorological variables, time-lag features and calendar information were introduced to fully explore the temporal dependencies and nonlinear correlations of load data, thereby improving the accuracy of multi-scale load forecasting. Li and Wang [39] constructed the VMD-SPOA-BiLSTM model. The pelican optimization algorithm (POA) was improved by introducing Logistic chaotic mapping, nonlinear weight factors, Cauchy mutation and sparrow vigilance strategies. Experiments demonstrate that this model achieves higher forecasting accuracy than other competing models. Huang et al. [40] proposed a hybrid deep learning load forecasting method named DCFformer (DFT-CNN-FEDformer). DFT is adopted to eliminate temporal noise in load series and suppress data fluctuations caused by unexpected events; CNN is used to extract local features from the denoised sequences; the extracted features are fed into the FEDformer network for load forecasting. Experimental results verify the favorable prediction performance of the proposed model.

The aforementioned studies have advanced the research on short-term load forecasting from various perspectives, yet they share several common limitations. First, most models employ a single optimizer for the entire parameter tuning process, lacking a task-divided design tailored to the distinct optimization objectives of the decomposition stage and forecasting stage. Second, the forecasting networks generally adopt a single-branch architecture, which fails to separately model the relatively stable inherent patterns embedded in load sequences and the uncertain fluctuations induced by meteorological and other external factors. Third, theoretical analysis regarding the root causes of load complexity remains insufficient; model construction is largely driven by engineering experience rather than the intrinsic characteristics extracted from data.

Although existing hybrid forecasting models have achieved favorable prediction performance, several urgent challenges and research deficiencies remain in current studies on short-term power load forecasting. Firstly, most existing researches merely focus on superficial fitting and feature mining of load time-series data, while ignoring the inherent cross-scale and mesoscale evolutionary characteristics of urban power load. Few studies conduct multi-scale feature analysis targeting typical distribution patterns including M-shaped load peaks and V-shaped load valleys. Moreover, insufficient exploration is carried out on the nonlinear coupling effects among load components at different scales, making it difficult to accurately characterize the complex evolutionary mechanism of load. Secondly, meteorological influencing factors are utilized in a relatively simplistic manner in existing literature, mostly through straightforward feature concatenation. The emergent evolutionary rules of load variations induced by meteorological nonlinear factors from the

perspective of embodied intelligence cannot be fully excavated, which restricts effective adaptation to uncertain load fluctuations under complex meteorological scenarios. Finally, conventional hybrid models suffer from inadequate pertinence in parameter optimization, weak capability for separated extraction of shallow and deep features, and poor prediction robustness for strongly fluctuating components. Such models struggle to balance explicit temporal feature extraction of load and fusion of implicit nonlinear meteorology-load features, leaving room for further improvement in generalization capability under multi-time-scale forecasting scenarios.

To address the aforementioned challenges, this paper proposes an intelligent hybrid prediction model that integrates cross-scale and meso-scale analyses with nonlinear fusion of meteorological information and power load data. The model is designed to capture the short-term load variation patterns and consumption behaviors under varying meteorological conditions in a target city. Specifically, the raw data are first processed via an optimized variational mode decomposition (AO-VMD) technique, which leverages the AO algorithm's fast convergence and strong global search capability to resolve the issues of mode number presetting and boundary effects inherent in standard VMD. Following physical decomposition, the resulting modal components are classified according to their feature characteristics: components carrying surface-level features are fed into a GRU model, while those embedding latent feature information, combined with meteorological factors, are input into an MSCOA-BiLSTM model. This establishes the AO-VMD-GRU-MSCOA-BiLSTM prediction framework (hereafter referred to as the MSHM model), which reconstructs the decomposed predictions to yield final forecast values. The proposed approach is of significant practical value for enhancing the safe operation and optimization of power grids.

The core contributions of this paper are summarized in the following three aspects. First, meteorological conditions and electricity load exhibit complicated nonlinear interactions across different scales. From the perspective of load peaks and valleys, this paper conducts statistical analysis on large-scale and mesoscale features in short-term load forecasting and explores the emergent laws governing how nonlinear meteorological factors affect load variation within the embodied intelligence framework to realize accurate load forecasting. Second, the Aquila optimizer algorithm-optimized variational mode decomposition (AO-VMD) method decomposes raw data into multiple Intrinsic Mode Function (IMF) components. This approach effectively addresses the inherent drawbacks of conventional VMD, including the artificial preset of mode number and boundary effects. Moreover, it stabilizes strongly nonlinear sequences during decomposition and reliably captures the complex nonlinear characteristics of power load. Third, the GRU forecasting model can effectively mitigate the gradient vanishing problem for long-sequence data while capturing long-range dependencies and nonlinear temporal dynamics embedded in load series. By integrating cross-scale and mesoscale features, the proposed MSCOA-BiLSTM alleviates low parameter sensitivity, achieves a favorable balance between global and local search, and supports efficient parallel computation. The model can effectively handle coupled nonlinearities and reduce the parameter scale of artificial intelligence models to cope with load uncertainties, thereby significantly boosting computational efficiency.

The remaining chapters of this paper are organized as follows. [Section 2](#) systematically elaborates the conceptual framework of MSHM and deeply analyzes its theoretical foundations. Combined with case studies, [Section 3](#) focuses on exploring cross-scale and mesoscale interactions as well as nonlinear characteristics, and empirically demonstrates the key basis for the technical feasibility of MSHM. [Section 4](#) applies MSHM to practical scenarios of power load forecasting, compares and analyzes its forecasting performance, verifies the effectiveness of the method through consistency tests, and highlights the interpretability advantage of the model. [Section 5](#) serves as the conclusion, summarizing the core findings of this research and proposing directions for future work.

2 Methodology

The proposed methodology begins with the AO-VMD technique, which combines the AO algorithm with variational mode decomposition to preprocess the urban short-term electricity data, yielding a set of intrinsic modal components (IMFs) and a residual component. The IMFs exhibiting relatively stable fluctuation patterns are then fed into a GRU model for prediction, whereas the BiLSTM model hybridized with the COA is employed to extract features from meteorological data and capture their coupling effects. The predictions from both branches are subsequently reconstructed to form the MSHM forecasting framework, which effectively characterizes the variation patterns of urban short-term electricity load and consumption behavior under diverse meteorological conditions.

2.1 The Aquila Optimizer Algorithm (AO)

In 2021, Abualigah proposed the AO algorithm, a metaheuristic optimization technique inspired by natural laws. As a typical representative of swarm intelligence algorithms, this algorithm generates individual instances by constructing an initial population and imposing boundary constraints, and the specific operations are implemented according to the rules defined by Eqs. (1) and (2).

$$X = \begin{bmatrix} x_{1,1} & \dots & x_{1,j} & \dots & x_{1,Dim} \\ \vdots & \dots & \vdots & \dots & \vdots \\ x_{i,1} & \dots & x_{i,j} & \dots & x_{i,Dim} \\ \vdots & \dots & \vdots & \dots & \vdots \\ x_{N,1} & \dots & x_{N,j} & \dots & x_{N,Dim} \end{bmatrix} \quad (1)$$

$$x_{i,j} = rand \times (UB_j - LB_j) + LB_j, i = 1, 2, \dots, N, j = 1, 2, \dots, Dim \quad (2)$$

where $x_{i,j}$ is the position of the i -th eagle of the j -th population; $rand$ is a random number between $[0, 1]$; N is the number of eagles in the population (taken as 20); Dim is the number of eagles in the population; UB and LB is the upper and lower bounds of the problem to be solved.

Step 1: The high-altitude flying behavior in a vertical dive is calculated by Eq. (3).

$$X_1(t+1) = X_{best}(t) \times (1 - \frac{t}{T}) + (X_M(t) - X_{best}(t) \times rand) \quad (3)$$

where $X_M(t)$ is obtained by Eq. (4),

$$X_M(t) = \frac{1}{N} \sum_{i=1}^N X_i(t), \forall j = 1, 2, \dots, Dim \quad (4)$$

where $X_1(t+1)$ is the position of the next iteration of method 1; $X_{best}(t)$ is the optimal point before the iteration; t and T are the current and maximum number of iterations; $X_M(t)$ is the average of the position before the iteration.

Step 2: A contour flight with a short glide is obtained by Eq. (5).

$$X_2(t+1) = X_{best}(t) \times Levy + X_R(t) + (y - x) \times rand \quad (5)$$

where $X_2(t+1)$ is the next iteration position of method 2; $X_R(t)$ is any one of the positions; $Levy$ is the flight distribution function as Eq. (6),

$$Levy = s \times \frac{u \times \sigma}{|v|^{1/\beta}} \quad (6)$$

where s is a fixed value of 0.01; u and v are random numbers between 0 and 1; σ is expressed as in Eq. (7),

$$\sigma = \left(\frac{\Gamma(1 + \beta) \times \sin(\frac{\pi\beta}{2})}{\Gamma(\frac{1+\beta}{2}) \times \beta \times 2^{\frac{\beta-1}{2}}} \right) \quad (7)$$

where $\Gamma(x)$ is the gamma function; β is a fixed value, take 1.5.

In addition, in Eq. (5), $y = r \times \cos(\theta)$ and $x = r \times \sin(\theta)$ are spiral, r is the helix radius, θ is the helix angle, and they are illustrated as in Eqs. (8) and (9), respectively,

$$r = r_1 \times U \times D_1 \quad (8)$$

$$\theta = -\omega \times D_1 \times \theta_1 \quad (9)$$

where r_1 is the number of search cycles, taking the value between 1 to 20; U is a fixed value, taking the value between 0.005 to 65; D_1 is an integer between 1 to Dim; ω is a fixed value, taking the value of 0.005; $\theta_1 = 3\pi/2$ is the initial angle.

Step 3: A slow descent at low altitude is calculated using Eq. (10).

$$X_3(t+1) = (X_{best}(t) - X_M(t)) \times \alpha - rand + ((UB - LB) \times rand + LB) \times \delta \quad (10)$$

where $X_3(t+1)$ is the location of the next iteration of method 3; α and δ are both fixed values of 0.1.

Step 4: Catch prey for walking by Eq. (11).

$$X_4(t+1) = QF \times X_{best}(t) - (G_1 \times X(t) \times rand) - G_2 \times Levy + rand \times G_1 \quad (11)$$

where $X_4(t+1)$ is the position of the next iteration of method 4, $X(t)$ is the current position; $QF(t) = t \frac{2 \times rand - 1}{(1-T)^2}$ is the equilibrium search quality function; $G_1 = 2 \times rand - 1$ is the winter planting campaign of the Skyhawk; $G_2 = 2 \times (1 - t/T)$ is the slope of the eagle's flight from the starting point to the end point, taking values from 2 to 0.

2.2 Variational Modal Decomposition (VMD)

The VMD [41] is a novel, adaptive, and completely non-recursive signal decomposition method. The core idea of the algorithm is to construct and solve the optimal solution of the variational problem, and the specific steps are shown in Eqs. (12) and (13).

Step 1: Establish the constrained variational problem. Assume that the original signal $f(t)$ is decomposed into K modal components u_k , ensuring that the decomposition sequence consists of modal components with central frequencies within a finite bandwidth, and the sum of the estimated bandwidths of each modal component is minimized. The constraint condition is that the sum of all modal components is equal to the original signal. Then, the constrained variational expression is established as shown in Eq. (12),

$$\begin{cases} \min_{\{u_k\}, \{\omega_k\}} \left\{ \sum_{k=1}^K \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \\ s.t. \sum_{k=1}^K u_k = f(t) \end{cases} \quad (12)$$

where $\{u_k\}$ and $\{\omega_k\}$ are the k -th modal component and the corresponding frequency center of the signal decomposition, respectively; $\delta(t)$ is the Dirac distribution function; and $*$ is the convolution operator.

Step 2: Solve the constrained variational problem, as shown in Eq. (13),

$$\begin{cases} \hat{u}_k^{(n+1)}(\omega) = \frac{\hat{f}(\omega) - \sum_{i \neq k} \hat{u}_i^{(n)}(\omega) + \frac{\hat{\lambda}^{(n)}(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k^{(n)})^2} \\ \omega_k^{(n+1)} = \frac{\int_0^\infty \omega |\hat{u}_k^{(n+1)}(\omega)|^2 d\omega}{\int_0^\infty |\hat{u}_k^{(n+1)}(\omega)|^2 d\omega} \\ \hat{\lambda}^{n+1}(\omega) = \hat{\lambda}^n(\omega) + \tau(\hat{f}(\omega) - \sum_{k=1}^K \hat{u}_k^{n+1}(\omega)) \end{cases} \quad (13)$$

2.3 The Gated Recurrent Unit (GRU)

The GRU [42] belongs to a special class of Recurrent neural networks (RNNs). Designed to address the gradient vanishing problem of traditional RNNs for long sequential data, it also features a simpler structure than the LSTM network.

The neurons of the GRU model are mainly composed of two parts, the update gate z and the reset gate r . The GRU prediction model process is shown in Eqs. (14)–(17).

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (14)$$

$$\tilde{h}_t = g[W_h x_t + U_h(r_t \odot h_{t-1}) + b_h] \quad (15)$$

There are neural-network gates shown in Eqs. (16) and (17), respectively,

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (16)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (17)$$

where g is the excitation function; x_t is the current moment input; $\tilde{h}_t$ is the current unit update information; h_{t-1} is the last moment unit output.

2.4 The Coati Optimization Algorithm (COA)

The COA is a meta-heuristic intelligent optimization algorithm for simulating coatis in nature, published by Dehghani M and other scholars in 2023. The algorithm models two natural behaviors of coatis: hunting iguanas and coatis fleeing from predators. The COA process is shown in Eqs. (18)–(25).

Step 1: Initialization, as shown in Eq. (18),

$$x_{i,j} = b_j^L + r(b_j^U - b_j^L) \quad (18)$$

where $x_{i,j}$ is an individual; b_j^L is the lower boundary of the search for superiority; b_j^U is the upper boundary of the search for superiority; r is a random number between $[0, 1]$.

Step 2: Iguana hunting and attack strategies (exploration phase)

Updating the first stage of modelling a coatis population in search space: simulating the strategy of a coatis when attacking an iguana (Iguana), with the location mathematically modelled, as shown in Eq. (19),

$$\begin{cases} X_{i,j}^{P1} = x_{i,j} + r(G_j - Ix_{i,j}) \\ i = 1, 2, \dots, \left[\frac{N}{2}\right]; j = 1, 2, \dots, m \end{cases} \quad (19)$$

where $X_{i,j}^{P1}$ is the new position of the i -th coatis in the j -th dimension; r is a random number between $[0, 1]$; G_j is the position of the iguana in the j -th dimension, which refers to the position of the best member; I

is a randomly selected number from the set $\{1, 2\}$; N is the number of coatis; $\lceil \frac{N}{2} \rceil$ is the largest integer not exceeding $\lceil \frac{N}{2} \rceil$; m is the number of decision variables.

The following two equations (Eqs. (20) and (21)) are used to simulate a random position in the search space where the iguana is placed after it has fallen to the ground.

$$G_j^g = b_j^L + r(b_j^U - b_j^L) \quad (20)$$

where G_j^g is the position of the iguana on the ground in the j -th dimension.

$$\begin{cases} x_{i,j}^{P1} = \begin{cases} x_{i,j} + r(G_j^g - x_{i,j}), & F_{G,j}^g \leq F_{i,j} \\ x_{i,j} + r(x_{i,j} - G_j^g), & \text{other} \end{cases} \\ i = \lceil \frac{N}{2} \rceil + 1, \lceil \frac{N}{2} \rceil + 2, \dots, N \end{cases} \quad (21)$$

where $F_{G,j}^g$ is the value of the objective function for the iguana after falling to the ground in the j -th dimension, and $F_{i,j}$ is the value of the objective function for the i -th coatis in the j -th dimension.

If the updated individual is better, update the current individual; otherwise, leave it as Eq. (22).

$$X_i = \begin{cases} X_i^{P1}, & F_i^{P1} \leq F_i \\ X_i, & \text{other} \end{cases} \quad (22)$$

where F_i^{P1} is the value of the objective function for the i -th coatis at the new position; F_i is the value of the objective function for the i -th coatis at the previous position.

Step 3: Escape from predators (exploitation phase)

The second stage of the process of updating coatis locations in the search space was mathematically modeled based on the natural behavior of coatis when encountering a predator and fleeing from a predator. To model this behavior, a random location is generated near the location of each coatis based on Eqs. (23) and (24),

$$b_{j,L}^{loc} = \frac{b_j^L}{t}, b_{j,U}^{loc} = \frac{b_j^U}{t}, t = 1, 2, \dots, T \quad (23)$$

where $b_{j,L}^{loc}$ is the local lower bound of the j -th decision variable; $b_{j,U}^{loc}$ is the local upper bound of the j -th decision variable; t is the number of iterations; T is the maximum number of iterations.

$$\begin{cases} X_i^{P2} = x_{i,j} + (1 - 2r)(b_{j,L}^{loc} + r(b_{j,U}^{loc} - b_{j,L}^{loc})) \\ i = 1, 2, \dots, N \end{cases} \quad (24)$$

where X_i^{P2} is the new position of the i -th coatis in the j -th dimension.

Update the current individual if the updated individual is better; otherwise, leave it as Eq. (25),

$$X_i = \begin{cases} X_i^{P2}, & F_i^{P2} \leq F_i \\ X_i, & \text{other} \end{cases} \quad (25)$$

where F_i^{P2} is the value of the objective function for the i -th coati at the new position; F_i is the value of the objective function for the i -th coati at the previous position.

2.5 Bidirectional Long Short-Term Neural Network Model (BiLSTM)

The BiLSTM neural networks [43] (as shown in Fig. 1) are bi-directional neural network structures combining forward LSTM channels and reverse LSTM channels, such that each input sequence passes through the recurrent neural network in both directions. In this paper's model, the electric load time-series signals are passed through the input layer. The information is fed into the BiLSTM network layer, and the input sample signal passes through the forward LSTM layer output, $\vec{h}_t$, and the backward LSTM layer output $\overleftarrow{h}_t$, which together determine the value passed into the hidden layer to obtain the BiLSTM output, y_t , whose formula is shown in Eqs. (26)–(28),

$$\vec{h}_t = LSTM(x_t, \vec{h}_{t-1}) \quad (26)$$

$$\overleftarrow{h}_t = LSTM(x_t, \overleftarrow{h}_{t-1}) \quad (27)$$

$$y_t = \overrightarrow{W} \vec{h}_t + \overleftarrow{W} \overleftarrow{h}_t + b_y \quad (28)$$

where $\overrightarrow{W}$ is the weight matrix of the forward LSTM layer to the output layer; $\overleftarrow{W}$ is the weight matrix of the backward LSTM layer to the output layer; b_y denotes the bias matrix of the output layer.

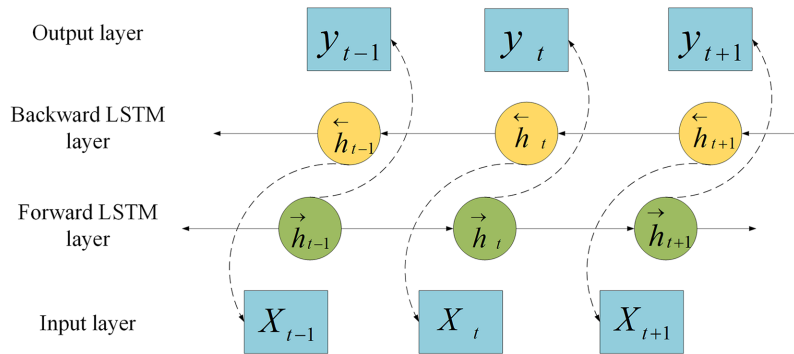


Figure 1: The structure of the BiLSTM.

2.6 Research Methodological Framework

This paper synthesizes the above approaches and proposes the following architecture:

Step 1: The AO-VMD algorithm is adopted to decompose the original load time series, and six intrinsic mode functions (IMFs) are finally obtained.

Step 2: Analyses are carried out by combining cross-scale characteristics, mesoscale laws and nonlinear effects under scale variations to deeply mine the latent features hidden in each modal component. Based on statistical theory, the modal components decomposed from the original data and meteorological influencing factors are separated to complete the classification and screening of modes.

Step 3: Feature components with significant fluctuation amplitudes are screened out and combined with the meteorological factor dataset as the input variables of the MSCOA-BiLSTM model. Among them, the COA module is responsible for iteratively solving the optimal parameter combination of the objective function, while the Bidirectional Long Short-Term Memory (BiLSTM) network undertakes the forecasting task of short-term power load.

Step 4: The feature information components with gentle fluctuation and strong stability are input into the gated recurrent unit (GRU) model to realize load forecasting for corresponding components.

Step 5: The prediction results from the above branches are integrated to construct the MSHM hybrid forecasting model and output the final power load forecasting data.

The overall flowchart framework of the model proposed in this paper is shown in Fig. 2.

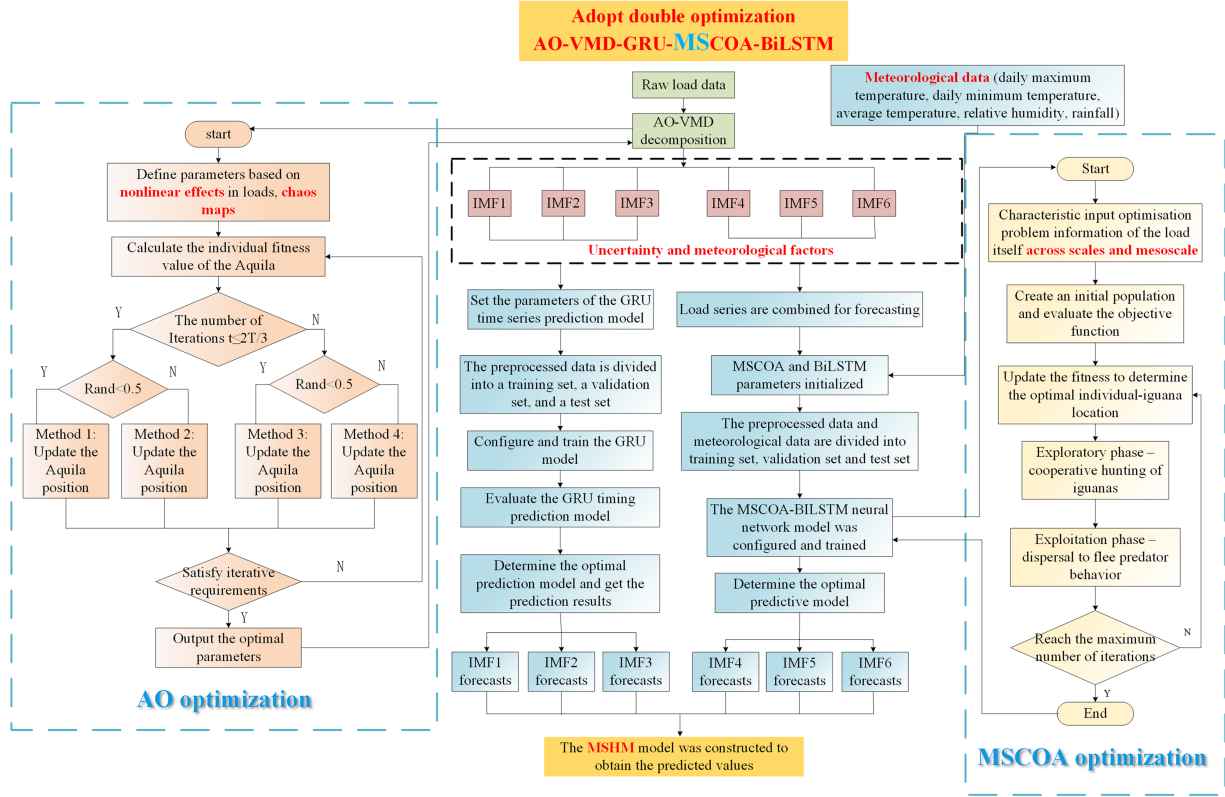


Figure 2: Flowchart of the network structure.

3 Case Studies

3.1 Data Source and Data Preprocessing

The experimental data in this study are collected from the electricity consumption records of a city in China in 2022. The time-varying curve of the original load series is shown in Fig. 3. It can be observed from the figure that the dataset takes 96 time points as one variation cycle with a sampling interval of 15 min, which exactly corresponds to the daily electricity consumption pattern. The power load time series exhibits certain random disturbances overall, while its fluctuation pattern shows obvious similarity and typical daily periodic characteristics. In terms of load distribution, the electricity load of this city remains at a relatively high level from Monday to Saturday, and drops significantly on Sundays. A total of 2016 power load samples are collected in this experiment. Such a data scale can effectively avoid experimental bias and accidental results caused by insufficient samples, and meanwhile prevent redundancy induced by excessive data. On this basis, load data for 21 consecutive days are selected for experimental analysis. This sample duration can sufficiently reflect the temporal variation characteristics and evolution rules of power load.

In this paper, we adopt a two-stage data preprocessing scheme including missing value imputation and feature normalization to standardize the raw power load and meteorological time-series data, so as to eliminate data defects and dimensional differences and lay a solid data foundation for subsequent decomposition and prediction modeling.

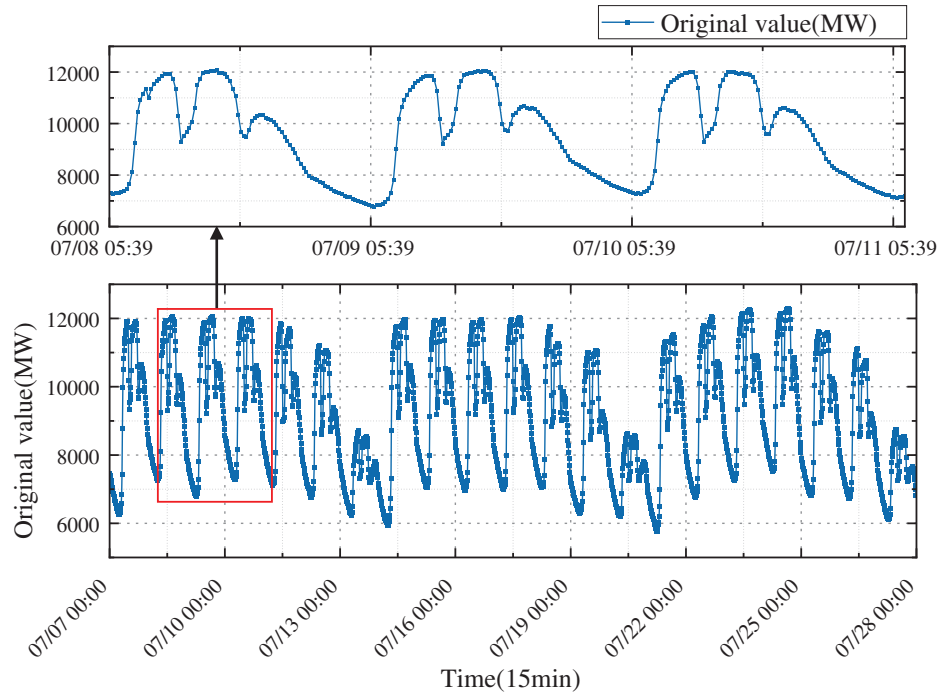


Figure 3: Raw data variation curve.

The linear interpolation method is adopted to complete missing value imputation for power load time-series data. The specific implementation procedure is as follows. First, we traverse the original power load time-series dataset to accurately locate time instants with missing data caused by equipment acquisition faults and signal interference. The valid load data adjacent to missing points are extracted as benchmark samples for interpolation, and a local short-time interpolation interval is defined to guarantee fitting effectiveness. Next, taking time as the independent variable and power load values as the dependent variable, a local linear fitting model is constructed based on adjacent valid data. The load values at missing moments are calculated via the linear interpolation formula. The trend of short-term load rise and fall is restored according to the load variation slope of adjacent periods to realize precise filling of missing data. Finally, the original missing positions are replaced with interpolated results to reconstruct a complete and continuous time-series dataset. We further verify the overall data fluctuation characteristics and daily periodic rules to ensure no feature distortion occurs after imputation.

After missing data repair, power load, temperature, humidity and other multi-dimensional meteorological features differ greatly in magnitude and physical dimension; direct input into the neural network will cause bias in parameter updating, slow down model convergence and interfere with feature learning effects. Therefore, we employ Min-Max normalization to map all feature data uniformly to the interval $[0, 1]$.

The calculation formula is defined as:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (29)$$

where x denotes the original sample value, $x_{\max}$ and $x_{\min}$ respectively represent the maximum and minimum values of the corresponding feature sequence, x^* stands for the normalized data.

This transformation eliminates the dimensional influence between load and meteorological variables while retaining the relative fluctuation law and daily periodic characteristics of the original data. Upon

normalization, the fluctuation amplitude difference between stable base load and meteorology-induced random fluctuation is reasonably preserved, which is conducive to AO-VMD adaptive modal decomposition and the subsequent dual-branch network feature extraction.

3.2 Cross-Scaling and Mesoscale

3.2.1 Across Scales

Cross-scale variation in power load refers to the distribution and fluctuation of electricity demand across different temporal scales. At the daily level, this manifests as peak-load periods during daytime and valley-load periods at night, alongside notable differences between weekday and weekend consumption patterns. At the seasonal level, it is reflected in the increased electricity demand for air conditioning in summer and heating in winter. A thorough understanding of these cross-scale load characteristics is essential for power system planning, operational scheduling, and energy policy formulation. Effective forecasting and dispatch informed by such multi-scale insights can enhance system stability, optimize resource allocation, and facilitate the integration of renewable energy sources while supporting carbon reduction objectives. Accordingly, this paper analyzes electricity consumption patterns at three distinct temporal scales: seasonal, weekly, and hourly.

As shown in Fig. 4, the quarterly level electrical load characteristics are analyzed every month, where March–May is spring, June–August is summer, September–November is autumn, and December–February is winter. In this paper, the electricity load data of a city for the year 2022 is selected to draw its box line diagram. Among them, each month contains data for the average load of each day of the month, and the unit is unified into MW. From the quarterly box plot, it can be seen that the overall electricity consumption load of this city shows a pattern of being higher in summer and autumn, and lower in spring and winter. Among them, the load was the highest in July and the lowest in February. From the perspective of load fluctuations, the fluctuation range was generally larger in summer and autumn, but the fifth month of spring also showed obvious volatility; the fluctuations in winter were relatively more stable. Overall, in summer and autumn due to factors such as cooling load and industrial production activities, the load dispersion was higher; although the overall average was lower in spring, the load fluctuations in May, due to the change of seasons or variable climate, were also not negligible.

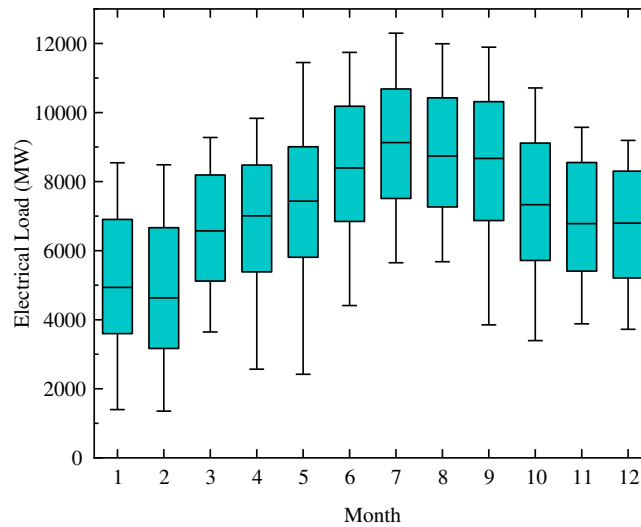


Figure 4: Monthly box plots grouped by quarter.

As shown in Fig. 5, the weekly electrical load characteristics are analyzed in terms of days of the week, thus having seven sets of data to obtain weekly box-plot data of electrical loads for a city for the year 2022. In addition, it can be observed that the electricity load in the city generally shows high electricity consumption from Monday to Saturday and low electricity consumption on rest days (Sunday), where the highest average daily load is observed on Wednesday, and the lowest average daily load is observed on Sunday.

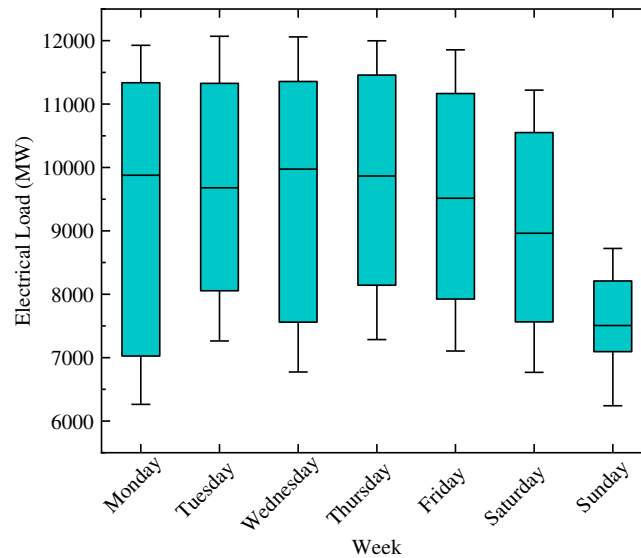


Figure 5: Weekly box plots.

As can be seen from Fig. 6, the hourly power load characteristics are analyzed for 24 h in a day with 24 sets of data, and the hourly power load box-plot data for a city for the year 2022. In this case, the overall trend of the electrical load shows a pattern of high daytime load and low nighttime load performance. The load starts to rise gradually from 6 a.m. and reaches its maximum at around 11:00 a.m., which is closely related to people's work and rest, and living habits.

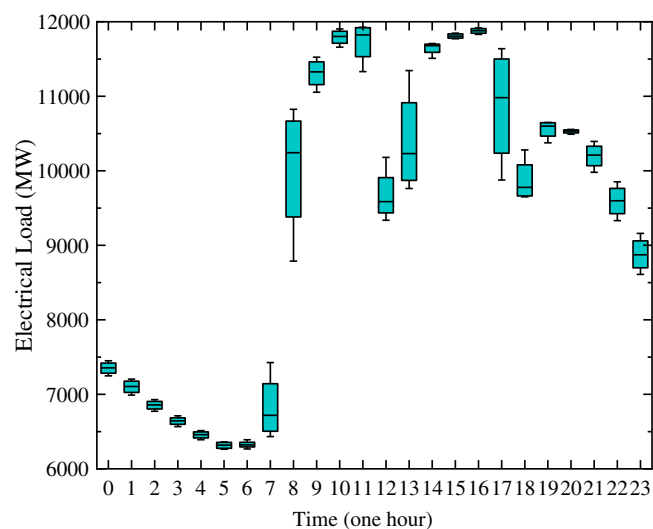


Figure 6: Hourly box plot.

3.2.2 Nonlinear Effects in Mesoscale

The mesoscale is defined as the range of intermediate scales bridging the unit scale and the overall system level. Analyzing the scale-dependent nonlinear interactions between meteorological conditions and energy consumption offers a novel perspective for understanding the underlying uncertainty in their relationship. Through a mesoscale theoretical framework, two representative load patterns—the “M”-shaped peak curve and the “V”-shaped valley curve—are derived to effectively characterize the extreme regimes of electricity demand. Both patterns are formulated as quadratic polynomial curves, providing an accurate mathematical representation of load evolution dynamics. This formulation substantiates the transition of the mesoscale concept from theoretical abstraction to practical application in load analysis.

Taking a full week as the research period, Figs. 7 and 8 depict the electricity load variations from 00:00 on 07 July 2022 to 23:45 on 13 July 2022, with a 15-min sampling interval. Between 08:15 and 18:15 each day, the load curve presents a distinct M-shaped profile, which recurs seven times throughout the week. Each M-shaped cycle contains five characteristic inflection points. From Monday to Saturday, the load peaks typically emerge at 11:15 and 16:30–17:00, indicating that workdays correspond to peak electricity consumption periods. By contrast, the load curve follows a V-shaped trend from 19:45 to 09:15 the following day, with the trough occurring between 05:30 and 06:30 during nighttime rest hours. Overall load variation trends are consistent across Monday to Saturday. Although the total electricity consumption on Sundays is markedly lower than that on weekdays, the intraday fluctuation patterns remain comparable. These phenomena demonstrate that regional electricity usage is strongly correlated with industrial activities and residents’ daily lives. Subsequently, we scale the 10-h M-type load data and the 13.5-h V-type load data, from which three quadratic polynomial fitting functions are derived.

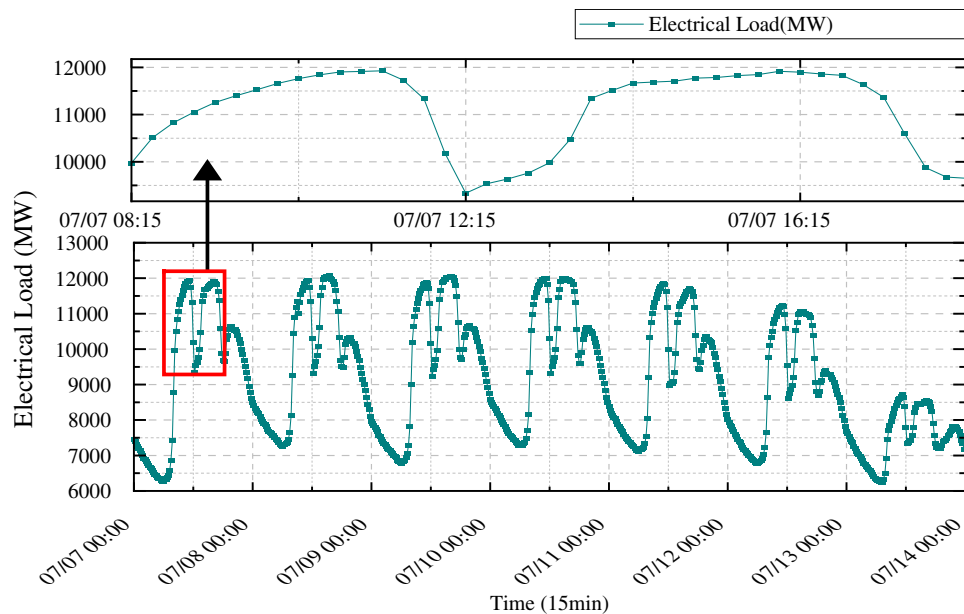


Figure 7: Change curve of electricity load for M-function.

As can be seen from the fitting equations in Table 1, the curve is a quadratic function with a downward slope, showing an ‘M’ pattern in the two time periods from 08:15 to 12:15 and from 12:45 to 18:15 on 7 July. The electricity load value shows a trend of increasing and then decreasing, with a slow increase before reaching the peak and a rapid decrease after the peak, indicating that the waking-up time and lunchtime of each

household in this area are not the same. During the period from 19:45 on 7 July to 09:15 on 8 July, the function shows a ‘V’ pattern, with a decreasing and then increasing trend in the load value, indicating that people use less electricity at night, and their electricity consumption behavior is more homogeneous and less affected by meteorological factors, and that the rest time of each household in this area is different. It is indicated that the nonlinear effect varies with the change of scale. There are complexities, differences, and uncertainties in the load, and nonlinear problems are difficult to describe. They affect the prediction accuracy of power load. Therefore, this paper considers the nonlinear relationships between different scales, different meteorological factors, and meteorological loads. The cross-scale and mesoscale characteristics of the load are input into the optimization problem information. Temperature, relative humidity, rainfall, and other meteorological factors are used as the input of the MSCOA-BiLSTM model. By combining statistical features with nonlinear features, the prediction accuracy of MSHM has been significantly improved.

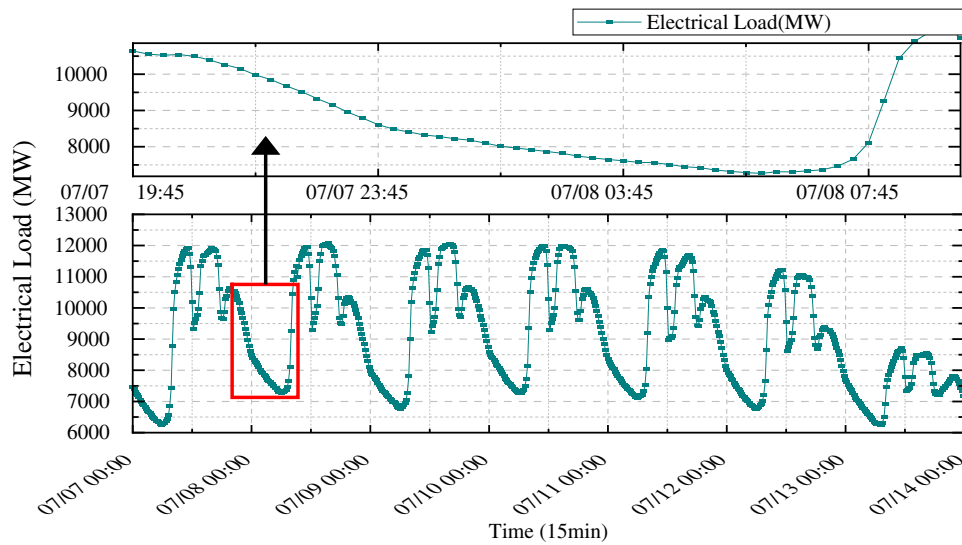


Figure 8: V-function electricity load change curve.

Table 1: The nonlinear effect.

Period	Fitting Formula
7 July 08:15 to 7 July 12:15	$y_t = 11952.92 + 10.23t - 32.03t^2$
7 July 12:45 to 7 July 18:15	$y_t = 12042.62 + 2.58t - 21.41t^2$
7 July at 19:45 to 8 July at 09:15	$y_t = 7562.04 - 27.76t + 4.56t^2$

3.3 Signal Expression

After decomposition via the AO-VMD algorithm, the original load signal is split into six intrinsic mode functions (IMFs) sorted in ascending order of frequency. Fig. 9 displays the time-series waveforms of all decomposed components, where IMF 1 to IMF 6 are arranged sequentially from top to bottom. These modal components collectively capture varying degrees of trend and periodic information inherent in the original power load data. Each component features distinct waveforms, with easily distinguishable amplitudes and frequencies.

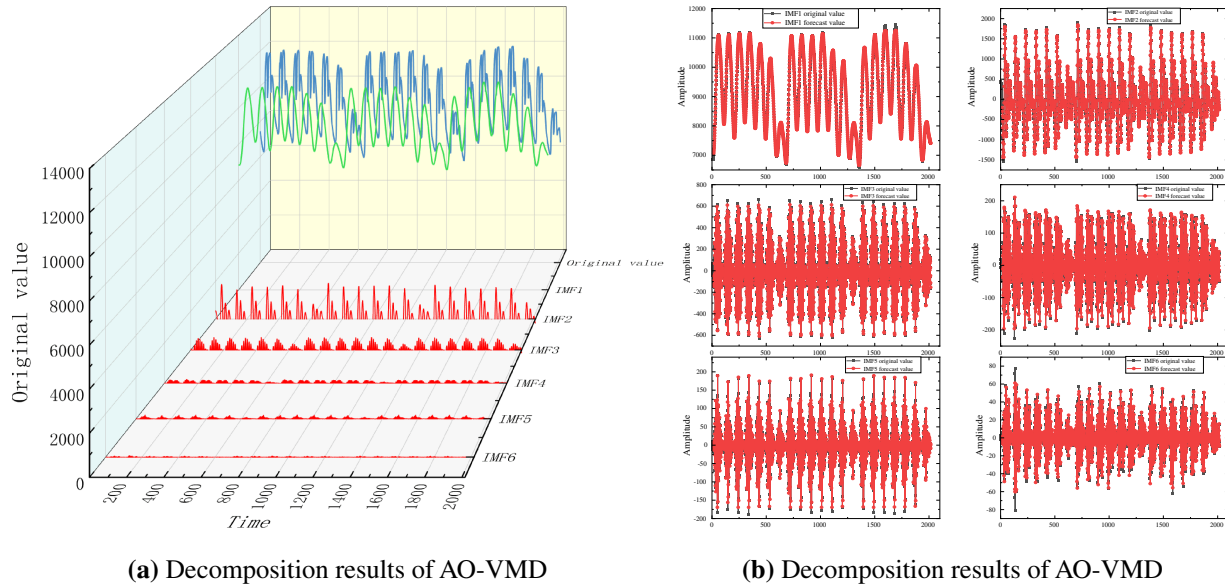


Figure 9: Decomposition results of AO-VMD.

3.4 Study of Fluctuation Characteristics

3.4.1 Fluctuation

As shown in Table 2, kurtosis characterizes the impulsive features of time-series signals and is sensitive to abrupt components. For a standard normal distribution, the theoretical kurtosis value is approximately 3. A substantial deviation from 3 indicates strong abrupt components within the signal. On one hand, a higher kurtosis value implies greater uncertainty in the electricity load series, which increases the difficulty of electricity load forecasting. In terms of skewness, the skewness of IMF1 is less than 0, showing a left-skewed distribution, while the skewness of IMF2 to IMF6 is greater than 0, showing right-skewed distributions.

Table 2: Signal characteristics.

Signal Characteristics	IMF1	IMF2	IMF3	IMF4	IMF5	IMF6
Kurtosis	1.75	3.20	3.04	2.85	4.93	4.70
Skewness	-0.02	0.39	0.17	0.06	0.07	0.16
Waveform factor	1.01	1.32	1.34	1.30	1.49	1.46
Crest factor	0.53	5.08	5.13	5.73	7.09	8.87
Impulse factor	0.53	6.70	6.89	7.44	10.60	12.93
Margin factor	0.53	8.28	8.79	9.16	14.75	17.34

In general, crest factor, impulse factor and margin factor share similar physical meanings. Crest factor and impulse factor can identify abrupt components in signals, and margin factor is used to characterize the extreme amplitude features of signal waveforms. Among the six modal components from IMF1 to IMF6, IMF5 and IMF6 exhibit larger values of waveform factor, crest factor, impulse factor and margin factor, indicating that the fluctuation risk of regional load time-series signals under external disturbances is more prominent.

3.4.2 Volatility

In this study, two evaluation metrics—the variance contribution rate and correlation coefficient—are adopted to analyze the decomposed components obtained via AO-VMD. These indicators quantify the contribution of each IMF to the fluctuation characteristics of the original time series and depict how periodic components of different cycles affect the composition of the raw data. Notably, the cumulative sum of the variance contribution rates does not equal unity.

Table 3 reveals that the variance contributions of the six IMF components are relatively balanced. The first three modal components carry more surface feature information and are thus more readily extractable, whereas IMFs 4–6 contain less explicit surface information, necessitating deeper investigation into their latent features to improve prediction performance. In terms of correlation with the original series, IMFs 1–3 exhibit higher correlation coefficients than the remaining components, indicating that these earlier modes more faithfully represent the original data. Given the incomplete feature extraction from IMFs 4–6, the model accuracy can be enhanced by incorporating additional explanatory factors. In contrast, the first three IMFs already contain pronounced feature information and do not require such supplementation.

Table 3: Volatility.

	IMF1	IMF2	IMF3	IMF4	IMF5	IMF6
Variance contribution	0.234	0.173	0.166	0.158	0.147	0.122
Correlation coefficient	0.664	0.593	0.231	0.140	0.060	0.085

Overall, from the perspective of fluctuation properties, IMFs 5 and 6 present higher uncertainty. By comparison, IMFs 1–3 better retain the inherent characteristics of the original time series and contain distinct feature information.

3.5 Load Uncertainty Analysis Considering Influencing Factors

Power load is inherently correlated with a variety of influencing factors, and the comprehensive consideration of these factors is essential for enhancing forecasting accuracy. To this end, the entropy weight method is employed in this study to quantitatively analyze the impacts of these factors.

As shown in Table 4, among all the influencing factors, rainfall exhibits the highest weight, reaching 57.786%, which is considerably higher than the other factors. It is inferred that the amount of rainfall in this region significantly affects residents' travel frequency, outdoor working hours, and the usage intensity of indoor cooling and dehumidification equipment, thereby directly causing substantial fluctuations in power load. This indicates that rainfall is a key external variable affecting the operational stability of the power system in this area. The second and third highest weights are attributed to average temperature and relative humidity, at 17.669% and 11.001%, respectively, suggesting that daily mean temperature and air humidity also exert relatively important influences on electricity demand. Changes in perceived temperature indirectly affect the load curve through the switching of air conditioning, heating, and other equipment. In contrast, minimum temperature (8.52%) and maximum temperature (5.023%) have comparatively smaller weights, implying that single-day extreme temperatures have limited explanatory power for overall load variations. Therefore, in this paper, rainfall, average temperature, relative humidity, minimum temperature, and maximum temperature are sequentially incorporated into the forecasting model as input features in descending order of their weights.

Table 4: Influencing factors and their attribute descriptions.

Evaluation Item	Weight (%)
Maximum temperature (°C)	5.023
Minimum temperature (°C)	8.52
Average temperature (°C)	17.669
Relative humidity	11.001
Rainfall (mm)	57.786

3.6 Application of Hybrid Models

The power load dataset comprises 2016 observations recorded at 15-min intervals, spanning from 00:00 on 07 July 2022, to 23:45 on 27 July 2022. The dataset is partitioned into a training set and a test set, containing 1920 and 96 samples, respectively. The training set covers the period from 00:00 on 07 July 2022, to 23:45 on 26 July 2022, while the test set corresponds to the subsequent 24-h period from 00:00 to 23:45 on 27 July 2022. This study mainly adopts the July 2022 dataset as the primary research object. To further verify the generalization capability of the MSHM model across different geographical regions and time scales, several supplementary experiments are added. We use datasets covering January 1–21, April 1–21, and October 1–21 of 2022, together with the publicly available 2023 Belgian power load dataset for validation. All experiments follow exactly the same experimental procedures as the July 2022 experiment.

The MSCOA-BiLSTM component comprises three parts: the MSCOA optimization module, the BiLSTM network, and the data input module. The optimal parameters identified by the MSCOA are decoded and passed to the BiLSTM, which is then trained using the training set from the data module. The mean squared error (MSE) between the actual outputs and the desired targets is computed and returned to the MSCOA module as the fitness value for iterative optimization. Meanwhile, the GRU model is configured and trained with its hyperparameters independently specified for time-series prediction. Once the optimal prediction model is obtained through training, the final forecast results are generated.

All experiments in this study were conducted on the Windows 10 operating system. The experimental environment configuration is detailed in Table 5, and the model parameter settings are summarized in Table 6. The development was carried out using MATLAB 2021a.

Table 5: Experimental environment configuration.

Designation	Configuration Info
Operating system	Windows10
Programming language	Matlab2021a
CPU	11th Gen Intel(R) Core(TM) i3-1115G4
GPU	Iris Xe
RAM	256 G

As shown in Table 6, the COA algorithm is adopted in this paper to automatically optimize the learning rate (search range: [0.001, 0.05]) and the number of hidden layer units ([4,20]) of BiLSTM, with the number of search agents set to 10 and the maximum number of iterations set to 30. The determination criteria for the above search intervals are elaborated as follows: the lower bound of the learning rate (0.001) follows the general prior knowledge in deep learning—values below this threshold will drastically slow

down convergence and make the model prone to local optima; the upper bound (0.05) is defined through preliminary experiments—training loss fluctuates drastically and the accuracy on the validation set degrades once this limit is exceeded. The lower bound of hidden layer units (4) satisfies the minimum representation requirement for time series prediction; the upper bound (20) is constrained by the size of the dataset used in this study and verified via preliminary experiments: when the number of units exceeds 20, the model suffers from obvious overfitting, the training time rises sharply, while prediction accuracy barely improves. For the AO-VMD model, the traditional VMD algorithm requires the decomposition level K to be manually predefined. In this work, the AO algorithm is employed to search for the optimal value of K within the preset interval $K \in [3, 10]$. The lower bound of 3 is derived from the physical prior of power load signals, which at least contain daily periodic components, weekly periodic components and trend components; the upper bound of 10 is determined based on the theoretical constraints of VMD decomposition and preliminary experiments: modal aliasing and frequency overlap occur persistently when $K > 10$. Taking the minimum sum of the bandwidths of all decomposed modes as the fitness function, the AO algorithm converges to the optimal decomposition level $K = 6$ for the dataset within 30 iterations. With regard to the training hyperparameters of GRU and BiLSTM, the maximum epoch is set to 500 and the initial learning rate is set to 0.01. A piecewise learning rate decay strategy is introduced, where the learning rate is reduced to 10% of its initial value at the 400th epoch to balance fast convergence in the early training stage and fine-tuned fitting in the later stage. An early stopping mechanism is also incorporated to prevent overfitting.

Table 6: Parameter settings for each model.

Model	Parameter Settings
MSCOA-BiLSTM	SearchAgents_no = 10; Max_iteration = 30; lb = [4, 0.001]; ub = [20, 0.05]; dim = 2;
GRU	MaxEpochs = 500, InitialLearnRate = 0.01, LearnRateSchedule, piecewise, LearnRateDropFactor = 0.1, LearnRateDropPeriod = 400
BiLSTM	MaxEpochs = 500, InitialLearnRate = 0.01, LearnRateSchedule, piecewise, LearnRateDropFactor = 0.1, LearnRateDropPeriod = 400
LSTM	MaxEpochs = 100, GradientThreshold = 1, InitialLearnRate = 0.005, LearnRateSchedule, piecewise, LearnRateDropPeriod = 25, LearnRateDropFactor = 0.16

As shown in Fig. 10, guided by the mesoscale analytical framework, the prediction task is assigned as follows: the GRU model is employed for the “V”-shaped trough segments, the MSCOA-BiLSTM model for the “M”-shaped peak segments, and the hybrid GRU-MSCOA-BiLSTM model for the remaining data portions. The resulting predicted curves exhibit strong concordance with the actual load profiles, demonstrating superior prediction accuracy and a higher degree of fit. This MSHM framework effectively overcomes the limitations of conventional models in characterizing low-order nonlinear chaotic time series, while faithfully capturing the overall load variation trend. Benefiting from the AO-VMD decomposition, the six IMF components carry distinct feature information. Specifically, IMFs 1–3 exhibit weaker correlations with external influencing factors and thus contain relatively simplified data structures, whereas IMFs 4–6 encode more complex latent feature information. To further enhance the representation of the original data, these higher-order IMFs are combined with meteorological variables—including maximum, minimum, and average temperatures, relative humidity, and rainfall—thereby enabling deeper mining of the underlying feature information.

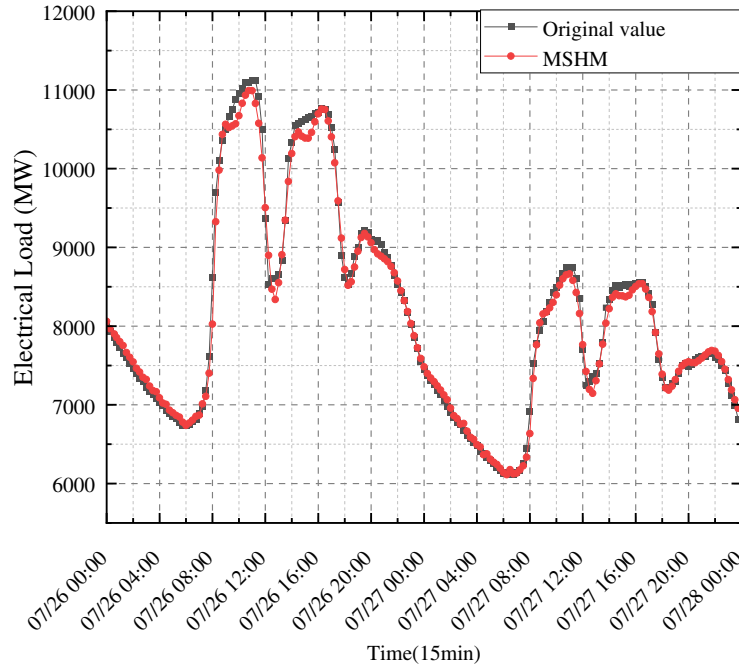


Figure 10: Prediction results of AO-VMD-GRU-MSCOA-BiLSTM.

4 Analysis of Results and Discussion

4.1 Indicators for Evaluation

Evaluation metrics can be used to measure whether the effectiveness of a model has been improved and how models compare to each other. In this paper, five evaluation metrics are used to evaluate the effectiveness of the prediction model. The calculation formula is in Eqs. (30)–(34).

Root mean square error (RMSE), which indicates how discrete the sample is; the smaller the value, the better.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (30)$$

Mean Absolute Error (MAE) represents the average of the absolute error between the predicted value and the true value; the smaller the value, the higher the accuracy of the model.

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (31)$$

Mean Absolute Percentage Error (MAPE) is a measure of how accurately a prediction is, with smaller values indicating more accurate predictive models.

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (32)$$

The mean square error (MSE) is the square of the difference between the expected value and the actual value: it indicates how discrete the sample is, and the smaller the value, the better.

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (33)$$

R^2 indicates the quality of the model fitting data, and the value range is generally $[0, 1]$:

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

where $\hat{y}_i$ ($i = 1, 2, \dots, n$) is the predicted value of sample i ; y_i ($i = 1, 2, \dots, n$) is the true value of sample i ; n is the number of samples.

4.2 Comparison of Models

Firstly, the prediction of the undecomposed data was carried out using models such as the LSTM, the BiLSTM, the GRU, and the MSCOA-BiLSTM. The results of the model prediction for undecomposed data are shown in Fig. 11.

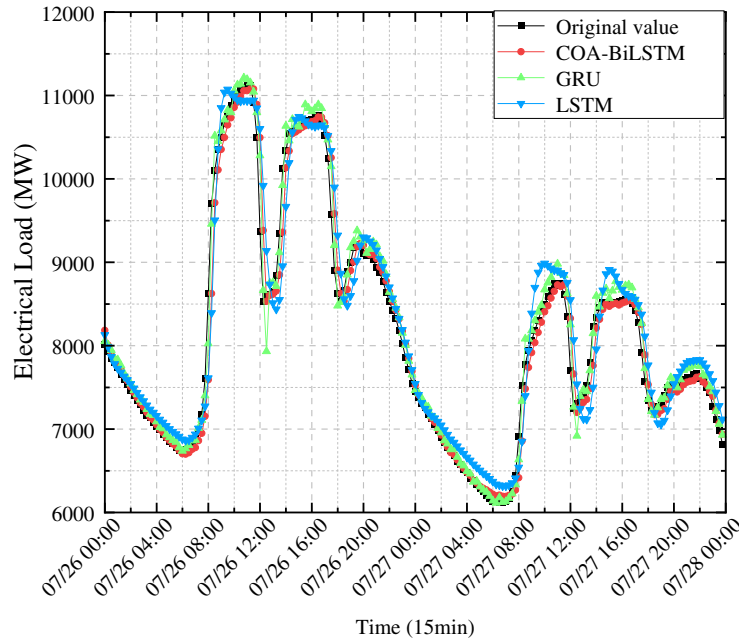


Figure 11: Predictions for raw data (undecomposed).

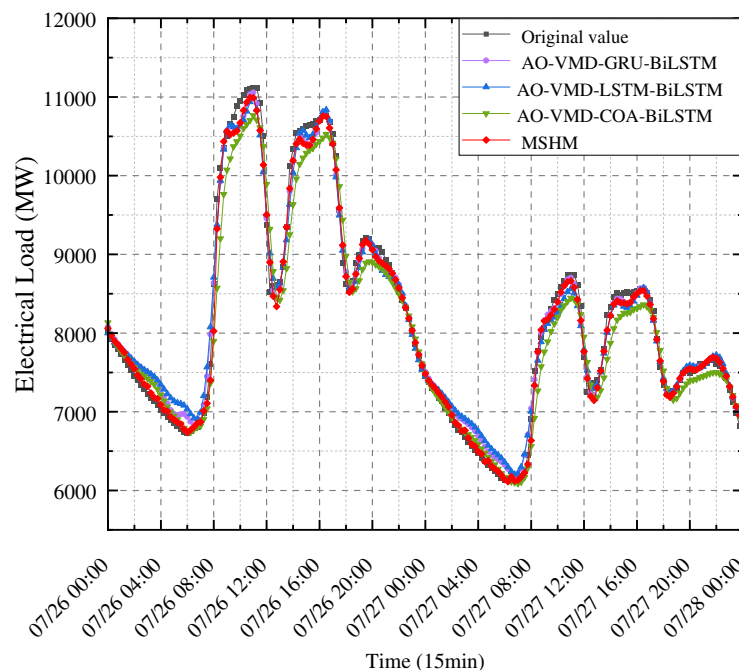
As shown in Fig. 11, the MSCOA-BiLSTM and GRU models achieve superior fitting performance compared to the other two models. On the one hand, this demonstrates that the BiLSTM enhanced by the coati optimization algorithm (COA) yields a significantly higher goodness of fit, substantially improving prediction accuracy and underscoring the effectiveness of the MSCOA-BiLSTM model. Nevertheless, its predictions still exhibit considerable volatility and nonlinearity relative to the actual values. On the other hand, the GRU model benefits from faster training speed and a lower risk of overfitting. By incorporating reset and update gates, the GRU flexibly regulates the utilization of historical information, thereby enhancing both modeling efficiency and predictive accuracy. These findings are further corroborated by the prediction error metrics presented in Table 7.

Table 7: Calculation of evaluation indicators (undecomposed).

Model	MSE	RMSE	MAE	MAPE (%)	R ²	Runtime
MSCOA-BiLSTM	65295.369	255.5296	151.726	1.7	0.974	2699 s
GRU	37428.977	193.4657	135.706	1.5	0.988	88 s
BiLSTM	2731317.798	1652.6699	1388.494	16.4	0.126	82 s
LSTM	131557.262	362.7082	252.370	8.4	0.948	10 s
XGBoost	2694643.596	1641.537	1405.411	64.4	0.123	15 s
SVR	3029266.501	1740.479	1548.948	70.3	0.015	13 s
RF	2693984.582	1641.336	1405.357	64.4	0.124	12 s
Transformer	57086.589	238.928	188.549	2.4	0.968	50 s
Informer	600797.016	775.11	448.517	5.7	0.81	45 s

From Table 7, it can be seen that in the case of data undecomposed, the GRU has the best prediction effect, which reaches 0.988, and compared with other models, the MSE, RMSE, MAE, and MAPE of the GRU are smaller than other models, indicating that the prediction value of the GRU is the closest to the original sequence among these models. The prediction accuracy of the model can be improved by decomposing the data.

Despite the superior performance of the MSCOA-BiLSTM and GRU models, their predictions still exhibit considerable deviations from the actual values, and the overall forecasting accuracy remains less than satisfactory. To address this limitation, we reviewed a substantial body of related literature, which highlights the advantages of variational mode decomposition (VMD) in signal processing and the critical role of machine learning techniques in improving prediction performance. Motivated by these insights, we further evaluated the predictive effectiveness of AO-VMD and standard VMD decomposition. Figs. 12 and 13 present the corresponding prediction results.

**Figure 12:** Prediction of electrical load based on AO-VMD decomposition (July 2022).

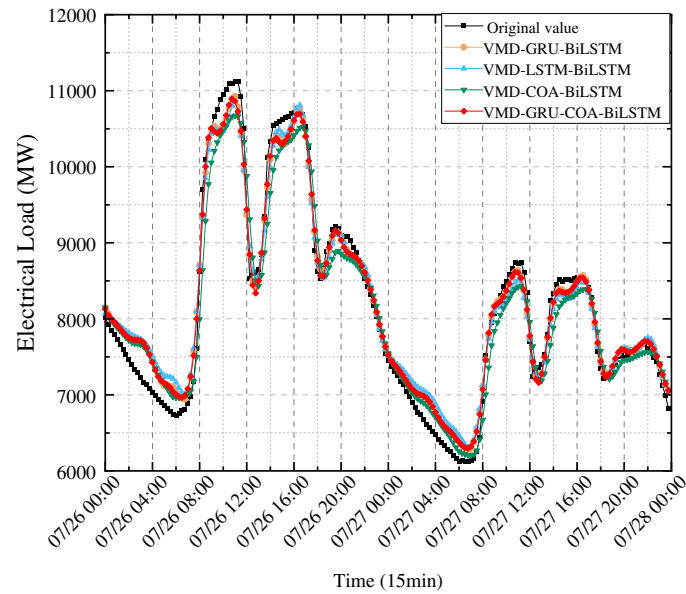


Figure 13: Prediction of electrical load based on VMD decomposition (July 2022).

As shown in Figs. 12 and 13, the forecasting results based on AO-VMD decomposition perform better in fitting the trend of actual data compared with the VMD decomposition method. The fusion model proposed in this paper achieves higher forecasting accuracy and minimizes the deviation between its output values and observed data. When analyzing the forecasting performance after decomposition, the selection of additional penalty factors and decomposition layers is critical for VMD decomposition; inappropriate parameter settings may degrade the overall algorithm performance. In contrast, AO-VMD adopts an adaptive optimization mechanism to dynamically adjust the optimal parameter combination, which significantly improves forecasting accuracy. The evaluation indices given in Tables 8 and 9 further verify this conclusion.

Table 8: Calculation of evaluation indicators (VMD decomposition).

Model	MSE	RMSE	MAE	MAPE (%)	R^2	Runtime
VMD-GRU-COA-BiLSTM	56987.349	238.720	198.887	2.4	0.990	2699 s
VMD-GRU-BiLSTM	61232.685	247.452	204.583	2.5	0.989	98 s
VMD-BiLSTM	3418789.547	1848.997	1592.127	19.8	0.003	98 s
VMD-LSTM-BiLSTM	80532.126	283.782	236.402	2.9	0.984	98 s
VMD-COA-BiLSTM	114718.417	338.701	279.833	3.2	0.967	2699 s

Table 9: Calculation of evaluation indicators (AO optimization).

Model	MSE	RMSE	MAE	MAPE (%)	R^2	Runtime
AO-VMD-GRU-BiLSTM	27408.0	165.55	131.057	1.57	0.994	230 s
AO-VMD-BiLSTM	3720127.2	1928.7	1650.98	20.54	0.025	218 s
AO-VMD-LSTM-BiLSTM	40834.56	202.08	163.265	1.99	0.990	218 s
AO-VMD-COA-BiLSTM	90280.29	300.46	227.799	3.54	0.972	2819 s
MSHM	17956.15	134	97.18	1.07	0.995	2819 s

Note: Bold values indicate the best performance among all compared models.

As shown in Table 8, the original sequence is decomposed into six sub-sequences via VMD, each of which is predicted individually, and the final forecast is obtained by summing the sub-sequence predictions. The computational results indicate that, compared with the non-decomposed model, the VMD-based approach achieves a general improvement in R^2 and a consistent reduction across all four evaluation metrics, demonstrating that the decomposition strategy effectively enhances predictive performance. Furthermore, the VMD-GRU-COA-BiLSTM model outperforms the VMD-GRU-BiLSTM variant, confirming the necessity and effectiveness of incorporating the COA optimization. Building on these findings, we proceed to further optimize the decomposition-based model and evaluate its prediction performance in the subsequent analysis.

To better validate the prediction performance of the MSHM model proposed in this paper, it focuses on analyzing some special points in the data. Fig. 14 shows the prediction results of each model after the AO-VMD decomposition.

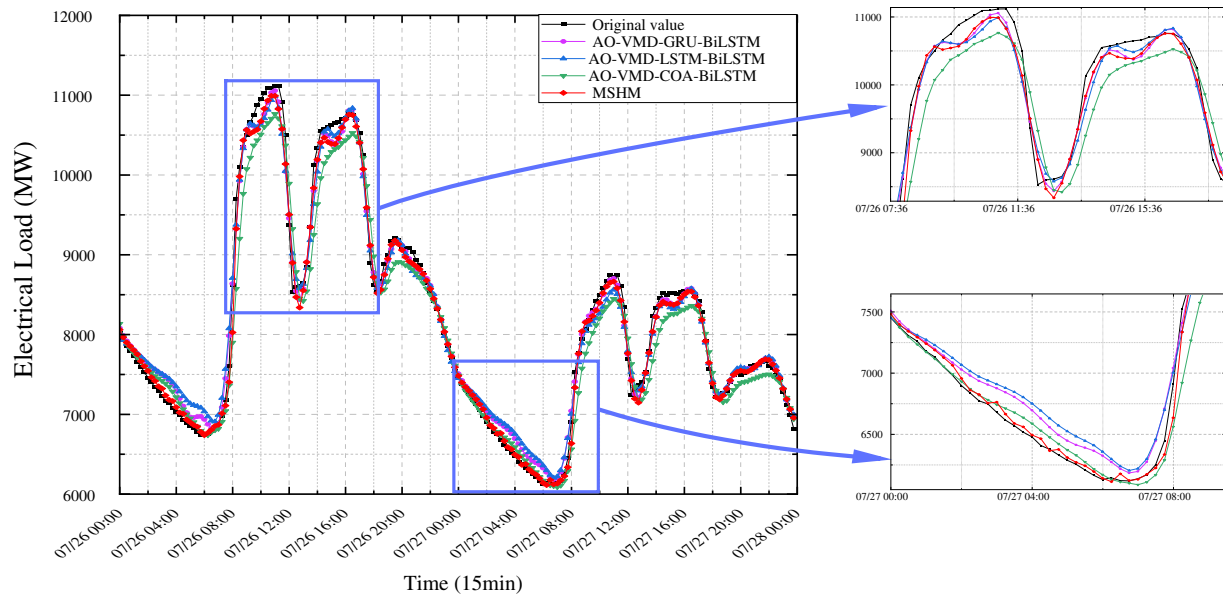


Figure 14: Prediction results of each model with AO-VMD decomposition (July 2022).

Fig. 14 compares the predicted values of the evaluated models against the actual load measurements. The proposed MSHM model, denoted by the red solid line, exhibits the closest agreement with the actual load profile (black solid line), with relatively small deviations throughout the prediction horizon. To provide a more detailed assessment of its trend-capturing capability, a local zoomed-in view is included in Fig. 14. This enlarged panel clearly demonstrates that the MSHM model consistently outperforms the other competing models in tracking the actual load variations, with predictions that most closely match the real measurements. These results confirm that the MSHM model possesses strong information fusion and feature-capturing capabilities.

a. Integration capacity

In this paper, we propose a hybrid coupled prediction framework that integrates artificial intelligence techniques. Specifically, the proposed model combines the AO-VMD decomposition method with a GRU-based time series prediction module and an MSCOA-BiLSTM module featuring enhanced information fusion capability.

b. Information Capture Capability

At the cross-scale level, we observe that electricity load is generally higher in summer and autumn, and lower in spring and winter. At the mesoscale level, the nonlinear effects are characterized through the “M”-shaped and “V”-shaped functions, which effectively capture the city’s electricity consumption patterns and enhance the model’s information extraction capability.

The decomposed model is optimized, and as shown in Table 9, the model optimized by the AO has better prediction results, and it also shows the necessity of optimization; the MSHM model R^2 reaches 0.995, which is the highest among the five models, while the MSE, RMSE, MAE, and MAPE are the smallest indexes. Therefore, it is correct and necessary to optimize the model with the AO in this paper.

To verify the reliability of the MSHM model, this study selects the four seasons—spring, summer, autumn, and winter in 2022—for forecasting.

As shown in Table 10, the MSHM model performs better during winter and summer (with strong load patterns) than during spring and autumn transition seasons (with high climate fluctuations), with MAPE consistently below 2.7% and R^2 above 0.98, indicating high prediction accuracy and robustness of the model. To further validate the generalization capability of the MSHM model across different geographical regions and time scales, this study additionally conducts a cross-regional prediction experiment using the publicly available 2023 Belgian power load dataset. As shown in Table 10, the model achieves an R^2 of 0.988 on the Belgian dataset, indicating that it explains over 98.8% of the variance in the target variable, demonstrating a highly satisfactory goodness-of-fit. Meanwhile, the corresponding Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) remain at relatively low levels, further corroborating the model’s robustness on unseen data. These results not only strongly verify that the MSHM model maintains excellent predictive performance in regions with distinctively different load patterns, but also suggest that the model exhibits favorable transferability and adaptability in characterizing nonlinear factors such as meteorological variables. This provides more substantial experimental support for the practical deployment and application of the model in real-world power systems.

Table 10: Performance evaluation metrics of the MSHM model under different seasons and the 2023 Belgian load dataset.

Month	MSE	RMSE	MAE	MAPE (%)	R^2	Runtime
January	14311.27092	119.62972	83.858913	1.53	0.991	2819 s
April	68509.03307	261.74230	196.27069	2.66	0.982	2819 s
July	17956.15	134	97.18	1.07	0.995	2819 s
October	68303.24811	261.34890	192.97972	2.3	0.988	2819 s
The 2023 Belgian dataset	15974.90092	126.39185	102.1393	0.97	0.988	2819 s

From the results in Table 11, it can be seen that the basic BiLSTM model has a certain time series fitting ability in power load forecasting, but the overall error is large and the coefficient of determination is low, indicating that it is insufficient in characterizing the complex nonlinear features of the load sequence. After introducing VMD, the model performance did not significantly improve, suggesting that relying solely on the decomposition strategy is difficult to effectively improve the prediction accuracy, and if the decomposition parameters are not adequately optimized, it may lead to information loss or feature redundancy. Further introducing AO to optimize the VMD parameters did not result in significant improvement in model performance, indicating that relying solely on parameter optimization is not sufficient to significantly enhance the prediction ability of complex load sequences. When GRU is added, the error indicators of

the model significantly decreased, and the R^2 increased significantly to 0.994, indicating that GRU can effectively extract the obvious time series features of the load sequence and form a good complement with the multi-scale information after decomposition. At the same time, introducing COA to optimize BiLSTM also improved the model accuracy, indicating that the optimization algorithm helps to enhance the modeling ability of BiLSTM for the coupling relationship between meteorological factors and load. Finally, the AO-VMD-GRU-MSCOA-BiLSTM model achieved the best results in all indicators, verifying the effectiveness and superiority of the proposed framework in load forecasting. This result indicates that the synergy of multi-scale decomposition, time series feature extraction, and intelligent optimization algorithms can significantly improve the accuracy and robustness of short-term power load forecasting.

Table 11: Results of the ablation experiment.

Model	MSE	RMSE	MAE	MAPE (%)	R^2	Runtime
BiLSTM	2731317.79	1652.67	1388.494	16.42	0.126	82 s
VMD-BiLSTM	3418789.55	1848.99	1592.127	19.8	0.003	98 s
AO-VMD-BiLSTM	3720127.2	1928.7	1650.98	20.54	0.025	218 s
AO-VMD-GRU-BiLSTM	27408.0	165.55	131.057	1.57	0.994	230 s
AO-VMD-COA-BiLSTM	90280.29	300.46	227.799	2.54	0.972	2819 s
MSHM	17956.15	134	97.18	1.07	0.995	2819 s

4.3 Methodological Testing

4.3.1 Normality Test

To verify the consistency of the proposed model, we conduct a normality test on the power load dataset and select the appropriate testing approach in accordance with the statistical properties of the data. The null hypothesis assumes that the sampled data follow a normal distribution derived from the overall population, while the alternative hypothesis holds that the sample population does not conform to a normal distribution.

As shown in Table 12, the p -values of both the Kolmogorov–Smirnov and Shapiro–Wilk tests are far below the 0.05 significance level, indicating that the load data do not follow a normal distribution and can therefore be treated as non-parametric. To further substantiate this conclusion, a Q-Q plot is provided in Fig. 15, where the data points deviate considerably from the fitted straight line at both ends, visually confirming the non-normality of the distribution. Accordingly, the Kolmogorov–Smirnov test is adopted in this study for comparative evaluation of prediction model accuracy.

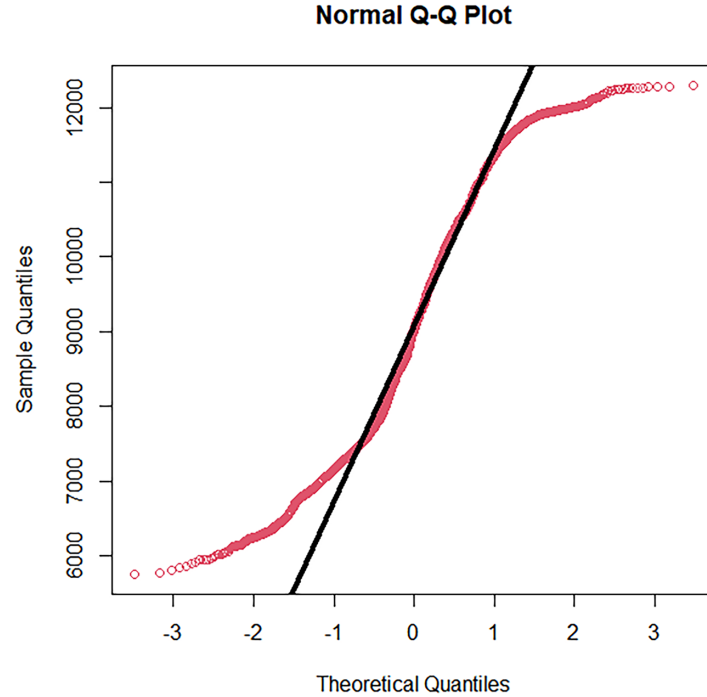
4.3.2 The K-S Test of the Proposed Model

The Kolmogorov–Smirnov Predictive Accuracy (KSPA) test serves as a supplementary statistical method for evaluating the forecasting accuracy between two sets of prediction results. As a nonparametric test, it is developed based on the fundamental principle of the Kolmogorov–Smirnov (K-S) test. The KSPA test is categorized into two variants: the two-sample two-sided test and the two-sample one-sided test. The two-sided test is applied to detect statistically significant discrepancies between the two prediction error distributions. The one-sided test further judges whether the model yielding the minimum overall error also possesses the smallest random error, thereby enabling quantitative comparisons of forecasting performance across different models.

Table 12: Tests of normality.

	Kolmogorov-Sminov ^a			Shapiro-Wilk		
	Statistics	df	Significance	Statistics	df	Significance
VAR00001	0.098	2016	0.000	0.944	2016	0.000

Note: ^aRicci's significance correction.

**Figure 15:** QQ chart.

The results of the prediction models are shown in [Table 13](#). The p -values of the unilateral and bilateral KSPA tests comparing the hybrid model with other models are less than 0.01 in this table, indicating statistically significant differences among the prediction models.

To further quantitatively verify whether the differences in forecasting performance between the proposed MSHM model and comparative models are statistically significant, this paper introduces the Wilcoxon signed-rank test to conduct nonparametric statistical significance analysis on the forecasting results of each model. Unlike traditional parametric tests, the Wilcoxon signed-rank test does not require the strict assumption of sample normal distribution. It fits the non-normal and discrete distribution characteristics of time series forecasting error samples, and is suitable for pairwise difference verification of prediction accuracy between two groups of models.

Table 13: Predictive accuracy test results (KSPA).

Comparison Model	Bilateral (<i>p</i> -Value)	Unilateral (<i>p</i> -Value)
MSHM vs. AO-VMD-GRU-BiLSTM	<0.01**	<0.01**
MSHM vs. AO-VMD-BiLSTM	<0.01**	<0.01**
MSHM vs. AO-VMD-LSTM-BiLSTM	<0.01**	<0.01**
MSHM vs. AO-VMD-COA-BiLSTM	<0.01**	<0.01**
MSHM vs. VMD-GRU-COA-BiLSTM	<0.01**	<0.01**
MSHM vs. VMD-GRU-BiLSTM	<0.01**	<0.01**
MSHM vs. VMD-BiLSTM	<0.01**	<0.01**
MSHM vs. VMD-LSTM-BiLSTM	<0.01**	<0.01**
MSHM vs. VMD-COA-BiLSTM	<0.01**	<0.01**

Note: **Indicates that the *p*-value for MSHM are <0.01, and are statistically significant.

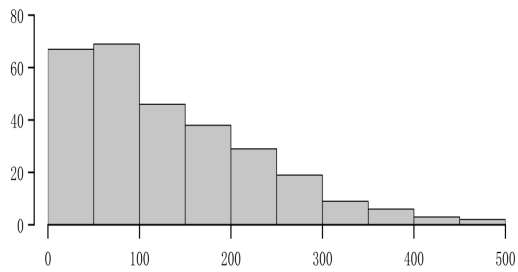
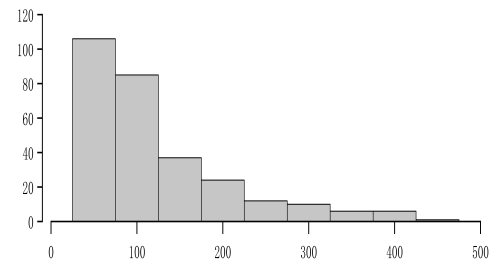
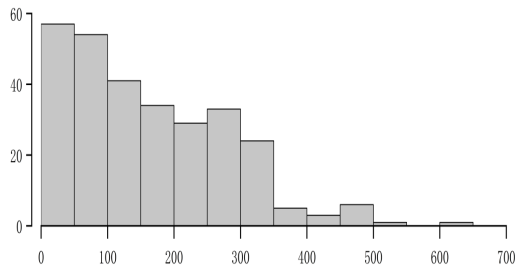
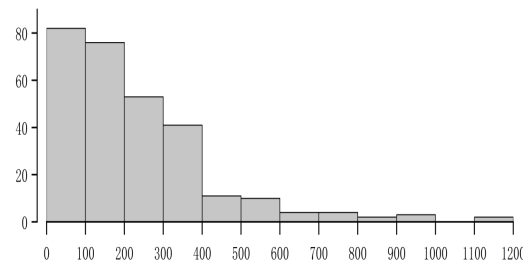
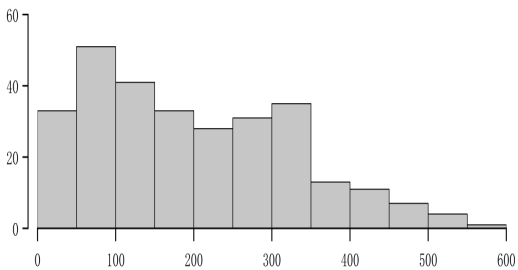
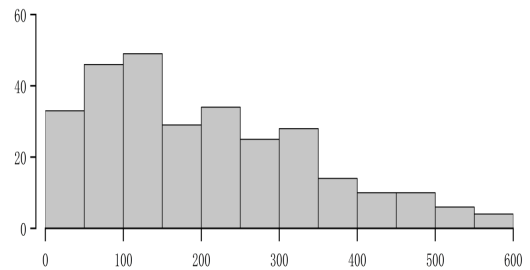
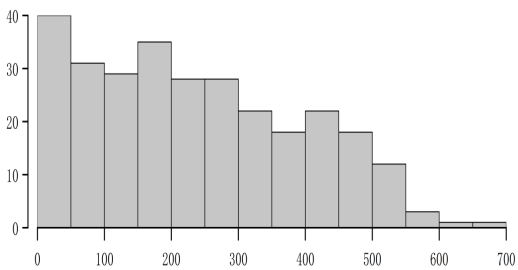
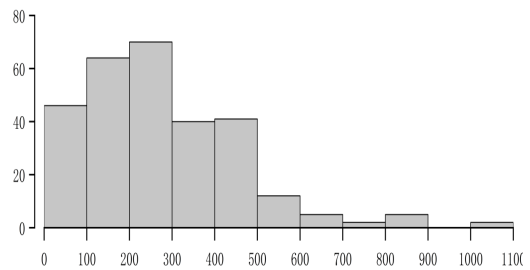
As can be observed from Table 14, the pairwise test results of the proposed MSHM model against all comparative models satisfy $p < 0.05$, which significantly rejects the null hypothesis. The results demonstrate that the performance improvement of the MSHM model over existing hybrid forecasting models does not stem from random experimental errors. Instead, it represents a statistically significant advantage brought by multi-scale decomposition, dual-branch differentiated feature extraction and adaptive parameter optimization mechanism of the model, which further verifies the effectiveness, robustness and superiority of the proposed framework.

Taking the data of July 2022 as an example: observing the error distribution of other comparable models, as shown in Fig. 16 (The horizontal axis denotes the magnitude of absolute errors, and the vertical axis represents the frequency of error samples), it can be seen that the MSHM model gradually reduces the absolute error of the predicted values with the increase of the error range, which is better than other models. The cumulative distribution function plot (as shown in Fig. 17) of the observation experience shows that the hybrid model converges quickly in the range of 200, close to 1, which fully indicates that the hybrid model has a higher probability of error concentration in a small range, and the prediction effect is better than other models. In short, the MSHM model has high prediction accuracy.

Table 14: Wilcoxon signed-rank test.

Comparison Model	<i>p</i> -value
MSHM vs. AO-VMD-GRU-BiLSTM	<0.05*
MSHM vs. AO-VMD-BiLSTM	<0.05*
MSHM vs. AO-VMD-LSTM-BiLSTM	<0.05*
MSHM vs. AO-VMD-COA-BiLSTM	<0.05*
MSHM vs. VMD-GRU-COA-BiLSTM	<0.05*
MSHM vs. VMD-GRU-BiLSTM	<0.05*
MSHM vs. VMD-BiLSTM	<0.05*
MSHM vs. VMD-LSTM-BiLSTM	<0.05*
MSHM vs. VMD-COA-BiLSTM	<0.05*

Note: *indicates that the *p*-value for MSHM are <0.05, and are statistically significant.

**(a)** AO-VMD-GRU-BiLSTM Absolute Errors**(b)** AO-VMD-GRU-MSCOA-BiLSTM Absolute Errors**(c)** AO-VMD-LSTM-BiLSTM Absolute Errors**(d)** AO-VMD-COA-BiLSTM Absolute Errors**(e)** VMD-GRU-COA-BiLSTM Absolute Errors**(f)** VMD-GRU-BiLSTM Absolute Errors**(g)** VMD-LSTM-BiLSTM Absolute Errors**(h)** VMD-COA-BiLSTM Absolute Errors**Figure 16:** Error distribution.

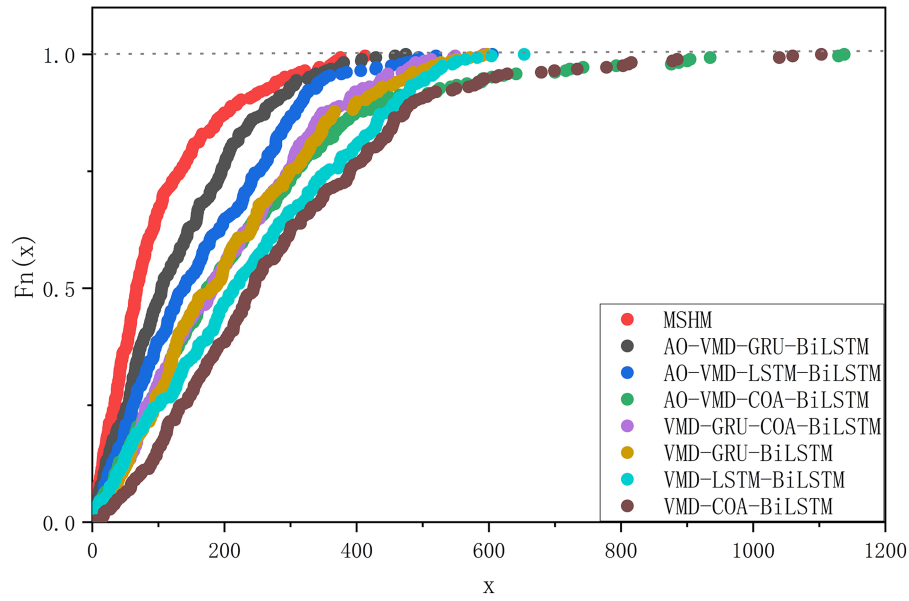


Figure 17: Empirical cumulative distribution function.

4.4 Interpretability of the Model

In the context of predictive modeling, interpretability refers to the capacity of a model to be articulated in plain language and comprehended by human users. In this paper, interpretability encompasses the following aspects:

The nonlinear effects between meteorological factors and energy demand are known to vary with temporal scale, and the coupling mechanisms among loads operate across multiple scales and processes. Unraveling the integration characteristics across cross-scale and mesoscale dimensions is essential for accurately characterizing the scale-dependent nonlinear relationships between meteorology and energy. Although urban electricity loads exhibit certain periodic regularities, the emergence of nonlinearities in load forecasting disrupts these patterns and gives rise to the internal mechanisms underlying load fluctuations and variations. From the perspective of global interpretability analysis, nonlinearity thus constitutes the primary factor affecting prediction accuracy.

First, the model introduces the Aquila optimizer algorithm to adaptively adjust the VMD decomposition parameters, automatically determining the optimal number of modes and the penalty coefficient. This enables the targeted separation of the cross-scale and meso-scale nonlinear components corresponding to the “M”-shaped peaks and “V”-shaped valleys in the load time series, precisely distinguishing effective features from random noise. In contrast, conventional comparison models typically employ fixed-parameter decomposition or no time-series decomposition at all, failing to adequately separate the non-stationary fluctuation components of the load, and thus remain susceptible to clutter interference that degrades prediction stability.

Second, the model constructs a dual-branch feature extraction architecture: the GRU is utilized to mine the inherent explicit evolutionary patterns within the load time series itself, while the Coati optimization algorithm-enhanced BiLSTM network is simultaneously employed to deeply integrate multi-dimensional meteorological factors—including precipitation, temperature, and humidity—thereby capturing the uncertain disturbances introduced by external meteorological variables. In contrast, benchmark models such as single GRU or LSTM rely solely on a single-stream network for information extraction, making it difficult to

simultaneously account for both the intrinsic temporal features of the load and the exogenous influences of meteorological factors, resulting in notable deficiencies in their ability to fit nonlinear coupled relationships.

Finally, this paper designs a dual-branch feature reconstruction and fusion mechanism that comprehensively integrates the intrinsic temporal features of the load with the meteorologically coupled features, completely characterizing the emergent patterns through which different meteorological variables influence electricity consumption. Most existing comparison models adopt lightweight single-structure architectures, lacking the collaborative framework of intelligent optimization-based parameter tuning, multi-scale decomposition, and dual-branch parallel feature extraction. Consequently, they are unable to adequately parse the complex coupling characteristics of power load, resulting in significantly higher prediction errors than those achieved by the MSHM model proposed in this paper. The synergistic effect of the above multi-level technical mechanisms endows the proposed hybrid framework with superior nonlinear fitting capability and robustness against uncertainty across different time scales.

5 Conclusion and Future Outlook

5.1 Conclusion

Accurate load forecasting can not only enhance the feasibility and security of power grid planning, but also lay a theoretical foundation for the technological advancement of power systems. Compared with conventional single recurrent neural networks and fixed-parameter signal decomposition forecasting frameworks, the proposed MSHM prediction model integrates adaptive signal decomposition, intelligent hyperparameter optimization and meteorological feature coupling, and its superiority is elaborated from multiple dimensions as follows:

(1) Most existing short-term load prediction methods merely fit overall load series without distinguishing multi-scale features, failing to quantify the nonlinear coupling between components of different time scales. This paper excavates cross-scale and mesoscale evolution rules of short-term power loads centered on peak-valley fluctuation characteristics, and quantitatively characterizes nonlinear interactions among multi-scale load components. The empirical results verify obvious discrepancies in electricity consumption modes between workdays and weekends: power consumption peaks on workdays while the overall load declines distinctly on Sundays, which is dominated by residents' daily work and living rhythms. Such refined scale division and behavioral interpretation enable the model to capture essential electricity consumption rules hidden in raw data, which fundamentally narrows forecasting errors and outperforms conventional holistic modeling strategies.

(2) Traditional fixed-parameter VMD is prone to modal aliasing and frequency overlap when coping with the strong nonlinearity, non-stationarity and multi-modal attributes of power load series. In contrast, the AO-optimized VMD adopted in this paper adaptively searches for the optimal decomposition number constrained by the physical periodic prior of load data, achieving clear separation of trend components and meteorology-induced high-frequency fluctuation components. On this basis, the GRU-BiLSTM dual-branch framework is constructed to learn long-period baseline trends and short-term random disturbances separately. This collaborative architecture effectively decouples intertwined nonlinear variations and stochastic uncertainties, delivering more precise feature extraction than standalone LSTM, GRU or BiLSTM models.

(3) Few existing load forecasting frameworks systematically incorporate meteorological coupling relationships into cross-scale temporal modeling. The MSHM model fuses deep learning with multi-dimensional meteorological covariates to jointly depict cross-scale nonlinear laws of short-term load, intricate coupling effects of meteorological factors and time-varying random disturbances, and harmonizes urban electricity consumption characteristics and actual power dispatching requirements. Quantitative verification based on

five universal evaluation indices (MAE, MAPE, RMSE, MSE and R2) demonstrates that the proposed model achieves consistent performance superiority over all comparative algorithms. The prominent prediction accuracy contributes to the reliable, economical and sustainable operation of power grids.

(4) In terms of computational efficiency, the extra computational overhead of the proposed framework is concentrated entirely on the offline training phase, covering AO-driven adaptive VMD decomposition, iterative hyperparameter optimization and joint training of the dual-branch network; no iterative calculation is required during online prediction. For practical power dispatching scenarios, periodic offline weight updating on a daily or weekly basis can be adopted to eliminate real-time delay. After offline training is completed, online prediction only executes lightweight feedforward inference, whose inference speed fully satisfies the timeliness requirements of short-term load forecasting. Although the multi-module coupled structure moderately increases offline computational complexity, the trade-off between high prediction precision and acceptable offline cost renders the model more applicable and valuable for practical power grid prediction tasks than conventional models with simple structures yet limited accuracy.

5.2 Future Outlook

This study has certain limitations, including the relatively high computational complexity of offline training and the need to further validate the model's generalization performance across multi-regional and multi-type power grid scenarios. Future work will focus on model lightweighting and the collection of multi-regional datasets to address these limitations. Meanwhile, based on the existing research on optimizing VMD and BiLSTM to improve load forecasting accuracy, this paper further expands the research scope and explores the application of large models in short-term and ultra-short-term power load forecasting scenarios. In future work, we will leverage the advantages of large models with massive parameters and complex computational architectures, combine multi-dimensional parameter features and embodied intelligence methods to enhance the capability of statistical feature mining, deeply explore the dynamic evolution rules of residents' power consumption behavior, and further improve the precision and applicability of power load forecasting models.

Acknowledgement: This work was funded by the Key Research Project in Universities of Henan Province (No. 25A450004).

Funding Statement: Guo-Feng Fan thanks the support from the project grants: Key Research Project in Universities of Henan Province (No. 25A450004).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Li-Ling Peng, Tong Li, Guo-Feng Fan; methodology, Guo-Feng Fan, Xin-Yu Yang; software, Tong Li, Xin-Yu Yang; validation, Li-Ling Peng, Tong Li; formal analysis, Guo-Feng Fan, Xin-Yu Yang; investigation, Li-Ling Peng, Tong Li, Guo-Feng Fan, Xin-Yu Yang; resources, Guo-Feng Fan; data curation, Li-Ling Peng, Tong Li, Xin-Yu Yang; writing-original draft preparation, Li-Ling Peng, Guo-Feng Fan; writing-review and editing, Wei-Chiang Hong; visualization, Tong Li, Xin-Yu Yang; supervision, Wei-Chiang Hong, Guo-Feng Fan; project administration, Guo-Feng Fan; funding acquisition, Guo-Feng Fan. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The power load dataset utilized in this study was obtained from the official portal of the TIPDM Cup Data Mining Challenge (<https://tipdmcup.cn/tzsList/index.jhtml>). The original data were sampled at 15-min intervals, from which we extracted 2016 consecutive time points (corresponding to 21 days) to examine the evolutionary patterns of power load. Beyond the load time series, the dataset also incorporates synchronously recorded meteorological observations, encompassing five critical variables—maximum temperature, minimum temperature, average temperature, relative humidity, and precipitation—thus furnishing a comprehensive empirical basis for exploring the nonlinear effects of meteorological conditions on electricity demand. All data are

publicly accessible via the aforementioned platform, and no modifications have been made to the original data, thereby guaranteeing the transparency and reproducibility of our findings.

Ethics Approval: Not applicable.

Conflicts of Interest: Given his role as Guest Editor of this Special Issue, Wei-Chiang Hong had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no other conflicts of interest.

References

1. Wang Y, Chen T, Liang Y, Sun H, Zhu Y. A novel cooling and power cycle based on the absorption power cycle and booster-assisted ejector refrigeration cycle driven by a low-grade heat source: energy, exergy and exergoeconomic analysis. *Energy Convers Manag.* 2020;204(9):112321. doi:10.1016/j.enconman.2019.112321.
2. Peña-Guzmán C, Rey J. Forecasting residential electric power consumption for Bogotá Colombia using regression models. *Energy Rep.* 2020;6:561–6. doi:10.1016/j.egy.2019.09.026.
3. Zhao H, Guo S. An optimized grey model for annual power load forecasting. *Energy.* 2016;107(11):272–86. doi:10.1016/j.energy.2016.04.009.
4. Zeng L, Liu C, Wu WZ. A novel discrete GM(2, 1) model with a polynomial term for forecasting electricity consumption. *Electr Power Syst Res.* 2023;214(2):108926. doi:10.1016/j.epsr.2022.108926.
5. Zhou W, Li H, Zhang Z. A novel seasonal fractional grey model for predicting electricity demand: a case study of Zhejiang in China. *Math Comput Simul.* 2022;200(7):128–47. doi:10.1016/j.matcom.2022.04.004.
6. Zhang B, Wu JL, Chang PC. A multiple time series-based recurrent neural network for short-term load forecasting. *Soft Comput.* 2018;22(12):4099–112. doi:10.1007/s00500-017-2624-5.
7. Gasparin A, Lukovic S, Alippi C. Deep learning for time series forecasting: the electric load case. *CAAI Trans Intel Tech.* 2022;7(1):1–25. doi:10.1049/cit2.12060.
8. Aguilar Madrid E, Antonio N. Short-term electricity load forecasting with machine learning. *Information.* 2021;12(2):50. doi:10.3390/info12020050.
9. Li K, Mu Y, Yang F, Wang H, Yan Y, Zhang C. A novel short-term multi-energy load forecasting method for integrated energy system based on feature separation-fusion technology and improved CNN. *Appl Energy.* 2023;351(5):121823. doi:10.1016/j.apenergy.2023.121823.
10. Zahid M, Ahmed F, Javaid N, Abbasi RA, Zainab Kazmi HS, Javaid A, et al. Electricity price and load forecasting using enhanced convolutional neural network and enhanced support vector regression in smart grids. *Electronics.* 2019;8(2):122. doi:10.3390/electronics8020122.
11. Zheng Y. Performance assessment of LSTM networks for short-term load forecasting based on temporal feature engineering. *Informatica.* 2026;50(9):171–86. doi:10.31449/inf.v50i9.9317.
12. Muzaffar S, Afshari A. Short-term load forecasts using LSTM networks. *Energy Proc.* 2019;158:2922–7. doi:10.1016/j.egypro.2019.01.952.
13. Mustaqeem IM, Kwon S. Short-term energy forecasting framework using an ensemble deep learning approach. *IEEE Access.* 2021;9:94262–71. doi:10.1109/ACCESS.2021.3093053.
14. Jiang Y, Liu M, Li Y, Liu Y, Zhang J, Liu Y, et al. Enhanced neighborhood node graph neural networks for load forecasting in smart grid. *Int J Mach Learn Cybern.* 2024;15(1):129–48. doi:10.1007/s13042-023-01796-8.
15. Ren Y, Xia L, Wang Y. An improved GM(1, 1) forecasting model based on *Aquila* Optimizer for wind power generation in Sichuan Province. *Soft Comput.* 2024;28(15):8785–805. doi:10.1007/s00500-023-09007-w.
16. Liu J, Hou Z, Yin T. Short-term power load forecast using OOA optimized bidirectional long short-term memory network with spectral attention for the frequency domain. *Energy Rep.* 2024;12:4891–908. doi:10.1016/j.egy.2024.10.050.
17. Chen Z, Jin T, Zheng X, Liu Y, Zhuang Z, Mohamed MA. An innovative method-based CEEMDAN–IGWO–GRU hybrid algorithm for short-term load forecasting. *Electr Eng.* 2022;104(5):3137–56. doi:10.1007/s00202-022-01533-4.

18. Goswami K, Kandali AB. Machine learning algorithms for predicting electrical load demand: an evaluation and comparison. *Sādhanā*. 2024;49(1):40. doi:10.1007/s12046-023-02354-2.
19. Li W, Zhang W. Hybrid GRU-EKAN model for short-term load forecasting of industrial air compressors. *Measurement*. 2026;267(21):120631. doi:10.1016/j.measurement.2026.120631.
20. Fan GF, Han YY, Li JW, Peng LL, Yeh YH, Hong WC. A hybrid model for deep learning short-term power load forecasting based on feature extraction statistics techniques. *Expert Syst Appl*. 2024;238:122012. doi:10.1016/j.eswa.2023.122012.
21. Manjili S, Ghouschi SJ, Maghami MR, Mohamed M. Interpretable AI hybrid model for electricity demand forecasting: combining TFT and XGBoost in smart grid data. *Comput Model Eng Sci*. 2026;147(1):1–10. doi:10.32604/cmescs.2026.076217.
22. Niu S, Zhang Z, Zhou H, Chen XB. Short-term power load prediction based on cluster analysis and temporal convolutional networks of attention mechanism. *Concurr Comput*. 2025;37(9–11):e70082. doi:10.1002/cpe.70082.
23. Yıldız C. An integrated model based on deep kernel extreme learning machine and variational mode decomposition for day-ahead electricity load forecasting. *Neural Comput Appl*. 2023;35(25):18763–81. doi:10.1007/s00521-023-08702-x.
24. Ribeiro MHD, da Silva RG, Moreno SR, Canton C, Larcher JHK, Stefenon SF, et al. Variational mode decomposition and bagging extreme learning machine with multi-objective optimization for wind power forecasting. *Appl Intell*. 2024;54(4):3119–34. doi:10.1007/s10489-024-05331-2.
25. Wu K, Peng X, Chen Z, Su H, Quan H, Liu H. A novel short-term household load forecasting method combined BiLSTM with trend feature extraction. *Energy Rep*. 2023;9(7):1013–22. doi:10.1016/j.egyrs.2023.05.041.
26. Yu Y, Wang F, Yang M, Zhang Y, Wang Z. Deep learning short-term load forecasting method based on adaptive mode decomposition. *Electr Power Syst Res*. 2026;252(1):112450. doi:10.1016/j.epsr.2025.112450.
27. Huang Y, Huang Z, Yu J, Dai X, Li Y. Short-term load forecasting based on IPSO-DBiLSTM network with variational mode decomposition and attention mechanism. *Appl Intell*. 2023;53(10):12701–18. doi:10.1007/s10489-022-04174-z.
28. Chen Z, Zhang D, Jiang H, Wang L, Chen Y, Xiao Y, et al. Load forecasting based on LSTM neural network and applicable to loads of “replacement of coal with electricity”. *J Electr Eng Technol*. 2021;16(5):2333–42. doi:10.1007/s42835-021-00768-8.
29. Deng M, Zhang Z, Zhou H, Chen X. Short-term electricity load forecasting based on T-CFSFDP clustering and stacking-BiGRU-CBAM. *Comput Mater Contin*. 2025;84(1):1189–202. doi:10.32604/cmcc.2025.064509.
30. Segura A, Meléndez J. Requirements and constraints of forecasting algorithms required in local flexibility markets. *Comput Model Eng Sci*. 2025;145(1):649–72. doi:10.32604/cmescs.2025.070954.
31. Liu M, Wang J, Jiang H, Hu C. Aperiodically intermittent control strategy and adaptive synchronization of neural networks with inertial and memristive terms. *IEEE Access*. 2023;11:93077–89. doi:10.1109/ACCESS.2023.3310473.
32. Neda OM. A multi-source data-driven framework for intelligent diagnostics and prognostics of power transformers. *Measurement*. 2026;286:122354. doi:10.1016/j.measurement.2026.122354.
33. Kim D, Lee D, Nam H, Joo SK. Short-term load forecasting for commercial building using convolutional neural network (CNN) and long short-term memory (LSTM) network with similar day selection model. *J Electr Eng Technol*. 2023;18(6):4001–9. doi:10.1007/s42835-023-01660-3.
34. Kiruthiga D, Manikandan V. Levy flight-particle swarm optimization-assisted BiLSTM+ dropout deep learning model for short-term load forecasting. *Neural Comput Appl*. 2023;35(3):2679–700. doi:10.1007/s00521-022-07751-y.
35. Fan GF, Zhang RT, Cao CC, Yeh YH, Hong WC. Applications of empirical wavelet decomposition, statistical feature extraction, and antlion algorithm with support vector regression for resident electricity consumption forecasting. *Nonlinear Dyn*. 2023;111(21):20139–63. doi:10.1007/s11071-023-08922-9.
36. Wang L, Tian T, Xu H, Tong H. Short-term power load forecasting model based on t-SNE dimension reduction visualization analysis, VMD and LSSVM improved with chaotic sparrow search algorithm optimization. *J Electr Eng Technol*. 2022;17(5):2675–91. doi:10.1007/s42835-022-01101-7.

37. Fan C, Li G, Xiao L, Yi L, Nie S. Short-term power load forecasting in city based on ISSA-BiTCN-LSTM. *Cogn Comput*. 2025;17(1):39. doi:10.1007/s12559-024-10401-1.
38. Neda OM. A hybrid LSTM-XGBoost model for multi-horizon short-term load forecasting in smart power distribution systems. *E Prime*. 2026;17(2):201211. doi:10.1016/j.eprime.2026.201211.
39. Li X, Wang F, Gu W. Short-term power load forecasting based on VMD-SPOA-BiLSTM. In: *Proceedings of the UNified Conference of DAMAS, IncoME and TEPEN Conferences (UNified 2025)*. Cham, Switzerland: Springer Nature; 2026. p. 181–91. doi:10.1007/978-3-032-00968-5_16.
40. Huang J, Wang S, Chen S, Ye P, Xu H, Wu Z, et al. Research on hybrid deep learning modelling for short-term electricity load forecasting. *Energies*. 2026;19(4):1019. doi:10.3390/en19041019.
41. Dragomiretskiy K, Zosso D. Variational mode decomposition. *IEEE Trans Signal Process*. 2014;62(3):531–44. doi:10.1109/tsp.2013.2288675.
42. Yu Z, Niu Z, Tang W, Wu Q. Deep learning for daily peak load forecasting—a novel gated recurrent neural network combining dynamic time warping. *IEEE Access*. 2019;7:17184–94. doi:10.1109/access.2019.2895604.
43. Gong J, Qu Z, Zhu Z, Xu H. Parallel TimesNet-BiLSTM model for ultra-short-term photovoltaic power forecasting using STL decomposition and auto-tuning. *Energy*. 2025;320:135286. doi:10.1016/j.energy.2025.135286.