

Article

IMVMD-MADNet: A Hybrid Framework for Multi-Scale Prediction of Chiller Energy Consumption

Ronghao Cheng, Xiaoqin Wen * and Yinghao Li

School of Automation Science and Engineering, South China University of Technology, Guangzhou 510641, China; 202421017695@mail.scut.edu.cn (R.C.); zhizhuxin@outlook.com (Y.L.)

* Correspondence: xqwen@scut.edu.cn

Abstract

Accurate prediction of chiller energy consumption is crucial for the efficient operation and intelligent management of heating, ventilation, and air conditioning (HVAC) systems in large buildings. However, such prediction remains challenging due to the multi-scale temporal coupling and nonstationary dynamics of chiller systems. Therefore, an IMVMD-MADNet hybrid framework integrating Improved Multivariate Variational Mode Decomposition (IMVMD) and a Multi-scale Aggregation Decomposition Network (MADNet) is proposed for chiller energy consumption prediction. To avoid information leakage and capture multi-scale features, a rolling local decomposition strategy with adaptive mode selection is employed. First, IMVMD performs stepwise decomposition within a sliding window, and the optimal number of modes is determined using envelope entropy. Then, sample entropy is used to reconstruct the multivariate modes into high-, medium-, and low-frequency components. Subsequently, a dual-branch MADNet combining wavelet-domain time-frequency modeling (WDP) and time-domain causal dependency modeling (TDP) predicts each component, and the results are aggregated to generate the final prediction. Bayesian optimization is employed to optimize the key hyperparameters. One year of real industrial chiller data from a plant in Huizhou, China, is used to evaluate the proposed model against 11 forecasting models. Results show that the dual-branch architecture outperforms single-branch models. The proposed model achieves the best performance across all forecasting horizons, with its advantage becoming more pronounced as the forecasting horizon increases, demonstrating stable predictive performance under the same-plant setting.

Keywords: prediction of chiller energy consumption; multivariate variational mode decomposition; multi-scale aggregation decomposition network; rolling decomposition; sample entropy reconstruction



Academic Editor: Maria Vicidomini

Received: 11 June 2026

Revised: 24 July 2026

Accepted: 28 July 2026

Published: 3 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Chiller energy consumption prediction aims to estimate future power demand and its temporal variations. This objective is achieved by analyzing historical operational data, weather conditions, and other relevant system parameters. The building sector accounts for approximately one-third of global energy consumption and nearly 40% of global greenhouse gas emissions [1,2]. Within buildings, HVAC systems are the primary energy consumers, accounting for 40–50% of total building energy consumption. As the core components of central air-conditioning systems, chiller units can account for more than 40% of HVAC energy consumption [3]. Therefore, accurate prediction of chiller energy

consumption is essential for improving building energy efficiency, supporting demand response strategies, and enabling intelligent operation and maintenance [4].

Mechanistic and data-driven methods constitute the two primary categories of chiller energy prediction. Mechanistic approaches rely on simulation platforms such as EnergyPlus [5] and TRNSYS [6]. These models characterize chiller systems based on thermodynamic and heat transfer principles. Semi-empirical models based on ASHRAE performance curves reduce the number of required parameters [7]. However, mechanistic methods require extensive building-specific parameterization. Moreover, they are computationally expensive and exhibit limited generalizability across different operating conditions. Statistical approaches, including ARIMA [8] and its variants, have also been applied to chiller performance prediction. However, these methods assume linear and stationary time series. Consequently, they are unable to capture the nonlinear and non-stationary characteristics of chiller energy data. Traditional machine learning methods, such as Random Forest [9], GBDT [10], and XGBoost [11], have likewise been applied to cooling load and chiller power prediction. Although these methods can model nonlinear relationships, their performance remains highly dependent on domain knowledge and manual feature engineering. Deep learning approaches, including LSTM [12], GRU [13], 1D-CNN [14], TCN [15], and Transformer [16], have demonstrated superior capability in modeling temporal dependencies. However, individual models remain insufficient for accurately characterizing the complex dynamics of chiller systems. Specifically, they struggle to simultaneously capture the multi-scale temporal coupling, nonstationary dynamics, and complex multivariable interactions inherent in chiller energy data.

Hybrid models that combine signal decomposition with neural networks have been widely adopted to address the limitations of standalone models. These models typically follow a “decomposition–prediction” paradigm. Specifically, EMD or VMD is employed to decompose nonstationary sequences into multiple frequency components. Each component is predicted independently, and the resulting predictions are subsequently aggregated. For example, ref. [17] proposed a VMD-LSTM model for power grid load forecasting. Similarly, ref. [18] developed a wavelet-LSTM hybrid model for short-term chiller energy prediction. These studies demonstrate that decomposition-based preprocessing effectively extracts multiscale features from nonstationary sequences. Multivariate VMD (MVMD) further extends this framework to multivariate systems. By processing multichannel signals simultaneously, MVMD extracts shared frequency components [19]. Another widely adopted strategy is the “decomposition–reconstruction–prediction” paradigm. In this paradigm, sample entropy (SE) is employed to reconstruct the decomposed modes into a smaller number of components with distinct frequency characteristics. For instance, ref. [20] proposed a CEEMDAN-SE-Transformer-BiGRU model that classifies modes into high- and low-frequency groups. Similarly, ref. [21] developed a CEEMDAN-SE-LSTM model for ultra-short-term load forecasting. Furthermore, ref. [22] introduced a dual-stage VMD-SE-CEEMDAN-BO-CNNGRU model. However, these methods still suffer from several important limitations. First, the number of decomposition modes is typically determined manually, without an adaptive selection mechanism. Second, most studies perform global decomposition on the entire dataset, which may introduce information leakage and limit practical applicability.

In parallel, advanced deep learning architectures have been proposed for time series forecasting. One research direction focuses on multiscale temporal modeling. Models such as TimesNet [23], Pyraformer [24], TimeMixer [25], and WPMixer [26] follow a common design principle. These models decompose input sequences into multiple temporal scales, learn representations at each scale, and subsequently fuse the resulting multiscale features for prediction. Another research direction emphasizes joint time–frequency modeling.

FEDformer [27] extends the Transformer architecture by incorporating frequency-domain attention through Fourier transforms. Wavelet-enhanced architectures have likewise been applied to chiller-related forecasting, facilitating multiresolution time–frequency analysis [18]. A third research direction focuses on channel-wise attention mechanisms. Representative examples include SE blocks, which perform adaptive feature recalibration [28], and iTransformer, which encodes each variable independently before attention-based fusion [29]. Despite these advances, most existing architectures remain general-purpose frameworks. Consequently, they lack domain-specific adaptations for chiller systems. Moreover, they do not jointly model the mode dimension introduced by signal decomposition and the channel dimension associated with multivariate inputs. Furthermore, time-domain causal modeling and frequency-domain feature extraction are not integrated within a unified complementary framework.

Two critical challenges remain in chiller energy prediction. The first is the information leakage introduced by global decomposition. When decomposition is performed on the entire dataset before data partitioning, future information is implicitly embedded into the training samples, leading to overly optimistic prediction results. Existing studies based on CEEMDAN [30], EMD [31], and wavelet decomposition [18] commonly adopt this strategy, which is impractical for real-world deployment. The second challenge is the lack of a prediction architecture tailored to chiller systems. Existing models cannot effectively capture the multiscale temporal coupling, nonstationary dynamics, and complex multivariable dependencies of chiller energy consumption, nor do they fully exploit decomposition-derived mode information or jointly model temporal and frequency-domain features, resulting in accumulated prediction errors over longer forecasting horizons.

To overcome these challenges, a hybrid forecasting framework, termed IMVMD-MADNet (Improved Multivariate Variational Mode Decomposition–Multi-scale Aggregation Decomposition Network), is proposed. The framework consists of three main components. The first component is a rolling IMVMD module. Information leakage is prevented through stepwise local decomposition within a sliding window, while the number of modes is adaptively determined using envelope entropy. The second component is a sample entropy-based reconstruction strategy, in which the multivariate modes are grouped into three predefined frequency components: high-, medium-, and low-frequency. The third component is the MADNet predictor. It integrates three submodules: a Mode–Channel Aggregator (MCA) for feature recalibration across both the mode and channel dimensions, a Wavelet Decomposition Predictor (WDP) for time–frequency structural modeling, and a Time-Domain Decomposition Predictor (TDP), which combines multiscale mixing, TCN, and BiGRU to capture temporal causal dependencies.

Specifically, The main contributions are summarized as follows:

1. A rolling IMVMD decomposition method is proposed to perform stepwise local decomposition within a sliding window, thereby avoiding information leakage and adaptively determining the optimal number of modes.
2. A sample entropy-based reconstruction strategy is proposed to reorganize multivariate components into three predefined frequency groups, thereby standardizing the input dimensionality and reducing complexity.
3. A dual-branch MADNet architecture is developed, integrating MCA, WDP, and TDP to jointly model mode, channel, temporal, and frequency information.
4. Extensive experiments on a real industrial chiller dataset demonstrate that the proposed method consistently outperforms 11 mainstream forecasting models across all forecasting horizons.

The rest of this study is arranged in five sections. Section 2 reviews related methodologies. Section 3 details the prediction framework, experimental data, and design. Section 4

reports multiple results—decomposition analysis, comparative evaluation, ablation studies, long-term prediction assessment, and interpretability analysis. Section 5 provides the conclusion. The abbreviations used in this paper are summarized in Table 1.

Table 1. List of abbreviations used in this paper.

Abbreviation	Full Name
IMVMD	Improved Multivariate Variational Mode Decomposition
MADNet	Multi-scale Aggregation Decomposition Network
MCA	Mode-Channel Aggregator
WDP	Wavelet Decomposition Predictor
TDP	Time-Domain Decomposition Predictor
RBB	Resolution Branch Block
MDM	Multi-scale Decomposition Mixing
TCN	Temporal Convolutional Network
BiGRU	Bidirectional Gated Recurrent Unit
SE	Squeeze-and-Excitation
BO	Bayesian Optimization
DM	Diebold-Mariano
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
SMAPE	Symmetric Mean Absolute Percentage Error
PeakMAE	Peak Mean Absolute Error
RampMAE	Ramp Mean Absolute Error
HVAC	Heating, Ventilation, and Air Conditioning
ADMM	Alternating Direction Method of Multipliers
DWT	Discrete Wavelet Transform
SHAP	SHapley Additive exPlanations

2. Methods

2.1. Multivariate Variational Mode Decomposition

Designed for analyzing univariate signals, the conventional Variational Mode Decomposition (VMD) algorithm decomposes a single-channel signal into several intrinsic mode functions (IMFs). Its main limitation lies in its inability to directly process multivariate signals or capture oscillatory modes shared across multiple channels. To address this issue, Rehman et al. [32] proposed the Multivariate Variational Mode Decomposition (MVMD) algorithm in 2019. This algorithm treats multi-channel signals as an integrated whole and performs modal decomposition in a multidimensional space, extracting frequency components common to all channels. It supports aligned decomposition across multiple channels and exhibits stronger anti-interference capability and improved stability.

MVMD extends VMD through the decomposition of a C-channel multivariate input $x(t)$ into K multivariate components $u_k(t)$.

$$x(t) = [x_1(t), x_2(t), \dots, x_C(t)]^T \quad (1)$$

$$u_k(t) = [u_{k,1}(t), u_{k,2}(t), \dots, u_{k,C}(t)]^T, \quad k = 1, 2, \dots, K \quad (2)$$

The core of MVMD centers around solving a constrained variational optimization problem. The associated components can be jointly derived through the following optimization framework:

$$\min_{\{u_{k,c}\}, \{\omega_k\}} \left\{ \sum_{k=1}^K \sum_{c=1}^C \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_{k,c}(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \quad (3)$$

$$\text{subject to } \sum_{k=1}^K u_{k,c}(t) = x_c(t), \quad \forall c = 1, 2, \dots, C, \quad (4)$$

where $u_{k,c}$ denotes the k -th component of the c -th channel, and ω_k represents the center frequency of the k -th component. Here, all channels share the same set of center frequencies $\{\omega_k\}$, which ensures that components with the same index across different channels exhibit identical frequency-band characteristics, thereby achieving cross-channel modal alignment.

To solve this constrained problem, the Alternating Direction Method of Multipliers (ADMM) is introduced for iterative solution. The relevant steps are as follows:

1. Initialize all $\{u_{k,c}^1\}$, $\{\omega_k^1\}$, and the Lagrange multipliers $\{\lambda_c^1(t)\}$.
2. Iteratively update $\hat{u}_{k,c}$, ω_k , $\hat{\lambda}$.

$$\hat{u}_{k,c}^{n+1}(\omega) = \frac{\hat{x}_c(\omega) - \sum_{i \neq k} \hat{u}_{i,c}^n(\omega) + \frac{\hat{\lambda}_c^n(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k^n)^2} \quad (5)$$

$$\omega_k^{n+1} = \frac{\sum_{c=1}^C \int_0^\infty \omega \left| \hat{u}_{k,c}^{n+1}(\omega) \right|^2 d\omega}{\sum_{c=1}^C \int_0^\infty \left| \hat{u}_{k,c}^{n+1}(\omega) \right|^2 d\omega} \quad (6)$$

$$\hat{\lambda}_c^{n+1}(\omega) = \hat{\lambda}_c^n(\omega) + \tau \left(\hat{x}_c(\omega) - \sum_{k=1}^K \hat{u}_{k,c}^{n+1}(\omega) \right), \quad (7)$$

where τ is the update parameter.

3. Check for convergence:

$$\frac{\sum_{c,k} \left\| \hat{u}_{k,c}^{n+1} - \hat{u}_{k,c}^n \right\|_2^2}{\sum_{c,k} \left\| \hat{u}_{k,c}^n \right\|_2^2} < \epsilon, \quad (8)$$

where ϵ is the convergence tolerance.

4. Repeat steps 2 to 5 until the convergence condition is satisfied or the maximum number of iterations is reached.

The number of modes k is the key hyperparameter of the MVMD algorithm. An appropriate choice of k enables the decomposition results to accurately capture the signal's essential characteristics while avoiding information redundancy. Therefore, this study introduces the concept of envelope entropy to enhance the objectivity and rationality of hyperparameter selection.

Envelope entropy measures the complexity of the IMF envelope. A lower envelope entropy indicates that the corresponding modal component contains more informative content. For the multivariate IMFs obtained from MVMD, the overall average envelope entropy over all modes and variables can be computed.

2.2. Sample Entropy

Sample entropy is a metric for quantifying the complexity and regularity of a time series. Each group of IMFs obtained from MVMD decomposition is evaluated and reconstructed according to its sample entropy, thereby effectively characterizing the information complexity of the time series. The steps to compute sample entropy are as follows [33]:

1. For the input signal component $\{x_i\} = \{x_1, x_2, \dots, x_N\}$, the embedding dimension m is chosen as 2, while the similarity tolerance r is selected as 0.2 multiplied by the standard deviation of the component. Subsequently, m -dimensional vectors X_i^m are constructed.

$$X_i^m = [x(i), x(i+1), \dots, x(i+m-1)] \quad (i = 1, 2, \dots, N-m+1) \quad (9)$$

- Calculate the similarity between vectors, i.e., the distance $d[X_i^m, X_j^m]$ between two m -dimensional vectors X_i^m and X_j^m .

$$d[X_i^m, X_j^m] = \max_{k=0,1,\dots,m-1} |x(i+k) - x(j+k)| \quad (10)$$

- For each m -dimensional vector X_i^m , count the number of vector pairs satisfying $d[X_i^m, X_j^m] < r$ (where $i \neq j$), denoted as $B_m^i(r)$. Then calculate the average proportion, i.e., the mean of $B_m^i(r)$, denoted as $B^m(r)$.

$$B_i^m(r) = \frac{1}{N-m} \times \text{num}\{D_m(X_i, X_j) < r\} \quad (11)$$

$$B^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} B_i^m(r) \quad (12)$$

- Update the dimension from m to $m+1$ and re-execute steps 1-3 to get $B^{m+1}(r)$.

$$B^{m+1}(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} B_i^{m+1}(r) \quad (13)$$

- Calculate the sample entropy:

$$\text{SampEn}(m, r, N) = -\ln\left(\frac{B^{m+1}(r)}{B^m(r)}\right) \quad (14)$$

The modal components obtained through MVMD decomposition contain a channel dimension, thus requiring integrated processing along the channel dimension when calculating sample entropy. Specifically, the sample entropy is first computed separately for each channel of each mode, resulting in a two-dimensional sample entropy matrix that includes both channel and mode dimensions. The average is then taken along the channel dimension to obtain the mean sample entropy corresponding to each mode.

Accordingly, all modes are ranked in descending order based on their mean sample entropy values. Using predefined thresholds, the modes are regrouped into three predefined frequency components: high-, medium-, and low-frequency. This process standardizes the input feature dimensions and mode counts across samples, thereby improving model stability and reliability.

2.3. Squeeze-and-Excitation Block

The Squeeze-and-Excitation (SE) Block [34] is a lightweight channel attention module. It explicitly models interdependencies among channels, enabling the network to adaptively recalibrate channel-wise feature responses. As a result, informative features are enhanced, whereas less informative ones are suppressed. The SE Block consists of three sequential operations: Squeeze, Excitation, and Scale, which collectively enable adaptive feature weighting. Only a brief overview of the SE Block is presented below, while detailed theoretical and mathematical formulations can be found in [34].

In multivariate time series forecasting tasks, the SE Block is often extended to scenarios involving 1D convolution or sequence modeling. It is used to capture inter-variable relationships and their relative importance for the prediction task. This module has proven successful in several areas, e.g., blood pressure prediction [35], bearing RUL prediction [36], and real-world vehicle charging duration prediction [37].

In this study, the introduction of MVMD decomposition adds a mode dimension, thereby transforming the input data from a conventional one-dimensional time series structure into a dual-dimensional structure encompassing both mode and channel dimensions. Building on this, the SE Block can be adapted to process 2D tensors. It models importance weights along both the mode and channel dimensions simultaneously, enabling the model to focus on both critical modes and critical channels. The overall architecture is illustrated in Figure 1.

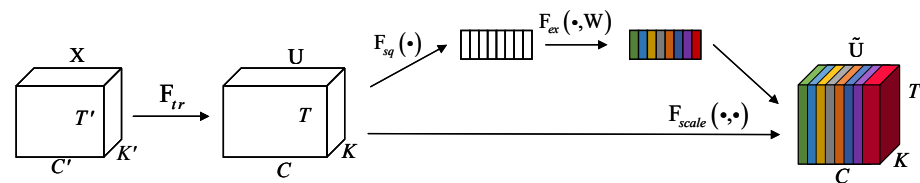


Figure 1. The structure of SE Block.

2.4. Multi-Level Discrete Wavelet Transform

The multi-level Discrete Wavelet Transform (DWT) iteratively decomposes time series data using a pair of high-pass and low-pass filters to extract local features at different frequency bands. Unlike Fourier transform, which only provides global frequency information, DWT possesses localization capabilities in both the time and frequency domains, offering distinct advantages for processing non-stationary data.

Given a discrete signal $x[n]$, a single-level wavelet decomposition obtains the approximation coefficient A_1 and the detail coefficient D_1 through convolution and downsampling operations [38]:

$$A_1[k] = \sum_n x[n] \cdot h[2k - n] \quad (15)$$

$$D_1[k] = \sum_n x[n] g[2k - n], \quad (16)$$

where $h[2k - n]$ is the scaling filter and $g[2k - n]$ is the wavelet filter.

Subsequently, DWT can recursively decompose the approximation component to achieve a multi-resolution representation across levels:

$$A_j[k] = \sum_n A_{j-1}[n] h[2k - n] \quad (17)$$

$$D_j[k] = \sum_n A_{j-1}[n] g[2k - n], \quad (18)$$

here, $A_j[k]$ represents the j -th level approximation coefficient, capturing the lowest-frequency trend information, while $D_j[k]$ denotes the j -th level detail coefficient, reflecting the high-frequency details.

Finally, a complete set of wavelet coefficients $\{A_j, D_j, D_{j-1}, \dots, D_1\}$, consisting of multi-scale approximations and details, is obtained. The original signal can be reconstructed by performing the inverse transform through upsampling and convolution at each layer, which can be expressed by the specific formula as follows:

$$A_{j-1}[n] = \sum_k A_j[k] h[n - 2k] + \sum_k D_j[k] g[n - 2k], \quad j = J, \dots, 1 \quad (19)$$

$$x[n] = A_0[n], \quad (20)$$

where A_{j-1} is the reconstructed approximation coefficient at level $j - 1$, and D_{j-1} is the reconstructed detail coefficient at level $j - 1$. When $j = 1$, the reconstructed $A_0[n]$ corresponds exactly to the original input signal $x[n]$.

2.5. Multi-Scale Decomposable Mixing (MDM) Block

Time series data generally exhibit two complementary temporal patterns: coarse-scale slow variations and fine-scale rapid fluctuations. These complementary patterns jointly characterize the overall dynamic structure of the time series. The MDM Block [39] explicitly models these two types of features by constructing a hierarchical multi-scale representation: the sequence is first decomposed layer by layer according to time scales, and then fused progressively from coarse to fine scales, thereby enhancing the expressive power of temporal features across scales.

Multi-scale decomposition: Given a single channel raw time series $\mathbf{x} \in \mathbb{R}^{1 \times L}$, it is directly treated as the first layer (finest-scale) pattern τ_1 . Subsequently, average pooling is applied recursively to the pattern of the previous layer for downsampling, obtaining coarser-scale sequential representations. The i -th scale decomposition can be expressed as:

$$\tau_i = \text{AvgPooling}(\tau_{i-1}), \forall i \in \{2, \dots, h\}, \quad (21)$$

Here, d is the downsampling rate, and h denotes the total downsampling steps. Once the $(i-1)$ -th downsampling is completed, the length of the i -th layer pattern τ_i becomes $\left\lfloor \frac{L}{d^{i-1}} \right\rfloor$. This process constructs a multi-scale sequence $\{\tau_1, \tau_2, \dots, \tau_h\}$ from high to low resolution, progressively suppressing detail while reinforcing smooth trends, thereby explicitly separating patterns at different temporal scales.

Cross-scale mixing: In this stage, the coarsest-scale pattern τ_h is designated as the top level cross-scale mixing information ξ_h . To inject global information from coarse-scale patterns into finer-scale patterns layer by layer, the Multi-Layer Perceptron (MLP) is used to non-linearly transform and upsample the coarse-scale pattern, aligning its length with that of the corresponding finer-scale pattern. It is then added to the finer-scale pattern via a residual connection, achieving top-down cross-scale feature enhancement. The cross-scale mixing information ξ_i at the i -th layer can be expressed as:

$$\xi_i = \tau_i + \text{MLP}(\xi_{i+1}) \quad (22)$$

This module, combining multiscale decomposition and cross-scale mixing, preserves the local variation patterns of fine scales while incorporating the long-term dependencies of coarse scales. This enables the model to learn a more comprehensive multiscale temporal representation.

2.6. Bidirectional Gated Recurrent Unit

The Gated Recurrent Unit (GRU) optimizes the hidden layer computation of Recurrent Neural Networks (RNNs). Compared to the Long Short-Term Memory (LSTM) network, GRU keeps only two gating components: the reset gate and the update gate, resulting in a simplified architecture. Its forward computation can be expressed as follows [40]:

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (23)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (24)$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h (r_t \odot h_{t-1})) \quad (25)$$

$$h_t = (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1}, \quad (26)$$

Here, z_t represents the update gate, r_t denotes the reset gate, $\tilde{h}_t$ is the candidate hidden state, and h_t is the final hidden state. Definitions of the remaining parameters can be found in reference [40].

By means of the update gate, which dictates how much historical memory is kept, and the reset gate, which regulates the contribution of past states to the candidate hidden state, GRU effectively overcomes the vanishing and exploding gradient challenges typical of conventional RNNs. This improves memory stability during the modeling of long sequences.

However, The unidirectional GRU has an obvious shortcoming: each time step's state is based only on past sequence patterns, with no ability to use future trends to better represent the current state. In practical time-series modeling scenarios, this often leads to semantic bias or incomplete feature extraction.

To address this drawback, this study introduces the Bidirectional Gated Recurrent Unit (BiGRU). The model combines a forward GRU and a backward GRU, then concatenates their hidden states $\vec{h}_t$ and $\overleftarrow{h}_t$ at each time step t to form the final feature representation:

$$h_t = [\vec{h}_t \parallel \overleftarrow{h}_t] \quad (27)$$

This bidirectional architecture allows the model to integrate contextual information from both directions, resulting in stronger feature extraction capabilities. The structure of BiGRU is illustrated in Figure 2.

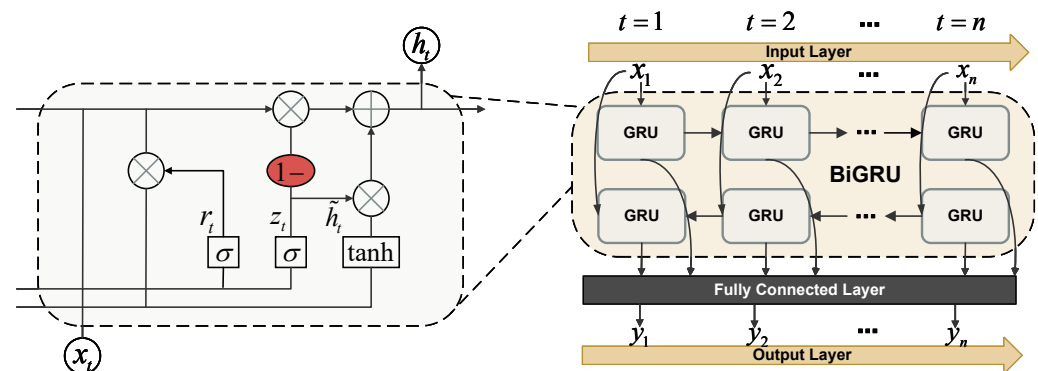


Figure 2. The structure of BiGRU.

2.7. Temporal Convolutional Network

Temporal Convolutional Network (TCN) refers to an enhanced sequence modeling method built upon conventional convolutional neural networks. It aims to strengthen the model's capability in capturing long-range temporal dependencies. This is achieved through dilated causal convolutions and residual connections, without violating the causality constraint. The structure of TCN is illustrated in Figure 3.

Dilated causal convolution ensures that each output depends only on the current and previous inputs, thereby preventing information leakage from future inputs. Moreover, it expands the receptive field without increasing the number of model parameters, thereby enabling the network to capture longer-range temporal dependencies. Its mathematical formulation is expressed as [41]:

$$F(x_t) = (X *^d F)(x_t) = \sum_{k=1}^K f_k \cdot x_{t-d(K-k)}, \quad (28)$$

where F denotes the filter, X is the input sequence, $*$ represents the convolution operation, d is the dilation factor, and K is the filter size.

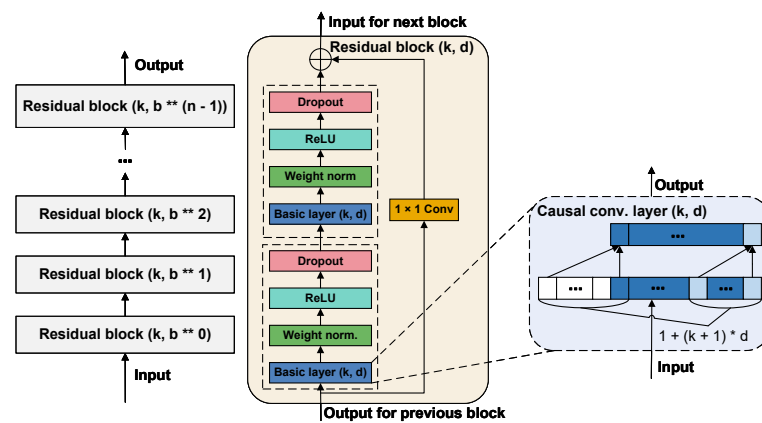


Figure 3. The structure of TCN.

Subsequently, to alleviate the vanishing gradient problem common in deep networks, TCN introduces a residual connection in each layer. When the feature dimensions of the input and output do not match, a 1×1 convolution is applied to the input for linear transformation, aligning its dimension with the output. The transformed input is then added to the convolutional output, yielding the final layer output as follows:

$$o_t^{(l)} = y_t^{(l)} + \text{Conv}_{1 \times 1} \left(x_t^{(l-1)} \right), \quad (29)$$

Here, $\text{Conv}_{1 \times 1}(\cdot)$ is the 1×1 convolution operation used to unify feature dimensions of the input and output, and $o_t^{(l)}$ is the l -th layer's final output.

2.8. Bayesian Optimization

Traditional hyperparameter tuning methods generally suffer from low search efficiency and a tendency to converge to local optima. To achieve more stable model performance, Bayesian optimization is employed to automatically optimize the key hyperparameters in each module. Bayesian optimization constructs a probabilistic surrogate model of the relationship between hyperparameters and model performance, which is iteratively updated using previous evaluation results. This approach enables efficient exploration of the hyperparameter space, making it particularly suitable for high-dimensional optimization problems and computationally expensive model evaluations.

A core step in Bayesian optimization is constructing a surrogate model that describes the uncertainty of the objective function, used to predict the potential performance of unevaluated parameter points. Without directly computing the objective function, the surrogate model provides a probabilistic estimation over its range, thereby guiding the selection of the next evaluation point. A common surrogate model is the Gaussian process, whose basic form can be expressed as [42]:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')), \quad (30)$$

where $f(x)$ is the objective function, $m(x) = \mathbb{E}[f(x)]$ is the mean function, and $k(x, x') = \text{Cov}(f(x), f(x'))$ is the covariance kernel function.

Once the surrogate model is built, an acquisition function is used to decide the next evaluation point. This study adopts Expected Improvement (EI) as the acquisition strategy to balance exploration of unknown regions and exploitation of promising areas. Its expression is:

$$\alpha_{\text{EI}}(\theta) = (f_{\text{best}} - \mu(x)) \Phi \left(\frac{f_{\text{best}} - \mu(x)}{\sigma(x)} \right) + \sigma(x) \phi \left(\frac{f_{\text{best}} - \mu(x)}{\sigma(x)} \right), \quad (31)$$

where $\Phi(\cdot)$ stands for the standard normal cumulative distribution function, and $\phi(\cdot)$ for its probability density function.

The hyperparameter tuning process of Bayesian optimization is carried out based on the aforementioned core components, with specific steps detailed in Algorithm 1.

Algorithm 1 Bayesian Optimization for Hyperparameter Tuning.

Require: Objective function $f(x)$; search space $\mathcal{X}$; initial sample size n_{init} ; maximum iterations T

Ensure: Best hyperparameter x^*

```

1: Initialize dataset  $\mathcal{D} \leftarrow \emptyset$ 
2: Randomly draw  $n_{\text{init}}$  samples  $\{x_1, \dots, x_{n_{\text{init}}}\}$  from  $\mathcal{X}$ 
3: for  $i = 1$  to  $n_{\text{init}}$  do
4:    $y_i \leftarrow f(x_i)$  {Evaluate the model}
5:    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(x_i, y_i)\}$ 
6: end for
7: Fit Gaussian Process surrogate model GP using  $\mathcal{D}$ 
8: for  $t = n_{\text{init}} + 1$  to  $T$  do
9:   Define acquisition function  $\alpha(x \mid \text{GP})$ 
10:   $x_{\text{next}} \leftarrow \arg \max_{x \in \mathcal{X}} \alpha(x)$ 
11:   $y_{\text{next}} \leftarrow f(x_{\text{next}})$  {Evaluate the objective function}
12:   $\mathcal{D} \leftarrow \mathcal{D} \cup \{(x_{\text{next}}, y_{\text{next}})\}$ 
13:  Update GP using the updated dataset  $\mathcal{D}$ 
14: end for
15:  $x^* \leftarrow \arg \min_{(x, y) \in \mathcal{D}} y$ 
16: return  $x^*$ 

```

2.9. Rolling IMVMD Decomposition Module

The conventional MVMD approach typically performs a single decomposition on the entire time series, which may incorporate information from future observations, thereby causing information leakage in forecasting tasks. Moreover, global decomposition may smooth out important local time–frequency characteristics. To address these shortcomings, an improved MVMD (IMVMD) method is proposed. By introducing a rolling decomposition mechanism, the input sequence fed into the prediction model is independently decomposed within a selected time window at each step. This avoids the premature use of future information and enhances the ability to capture local non-stationary characteristics [43].

Regarding parameter setting, IMVMD adopts the minimum envelope entropy as the optimization objective and adaptively determines the optimal number of modes K based on pre-training data samples. This effectively avoids the subjectivity associated with manually set parameters. Furthermore, to achieve consistent frequency-band partitioning of multivariate components, this study introduces the average sample entropy as a complexity measure [44]. Specifically, the sample entropy is calculated separately for each channel of the same mode, and the mean value is taken. This strategy reduces the influence of anomalous fluctuations that may exist in individual channels and ensures stable frequency-band division at the multivariate level.

Based on the relative magnitude of the average sample entropy, empirical thresholds are set to reconstruct the obtained multivariate components. Specifically, using thresholds of 0.7 and 0.2, the components are divided into three categories: high-frequency, medium-frequency, and low-frequency. The thresholds of 0.7 and 0.2 were selected because they provide a good balance between frequency discrimination and the distribution of modes across the three frequency bands, leading to stable decomposition and prediction performance. Consequently, this strategy enables a unified multi-scale representation and standardized input dimensions. Finally, the rolling IMVMD decomposition framework for chiller energy consumption data is illustrated in Figure 4.

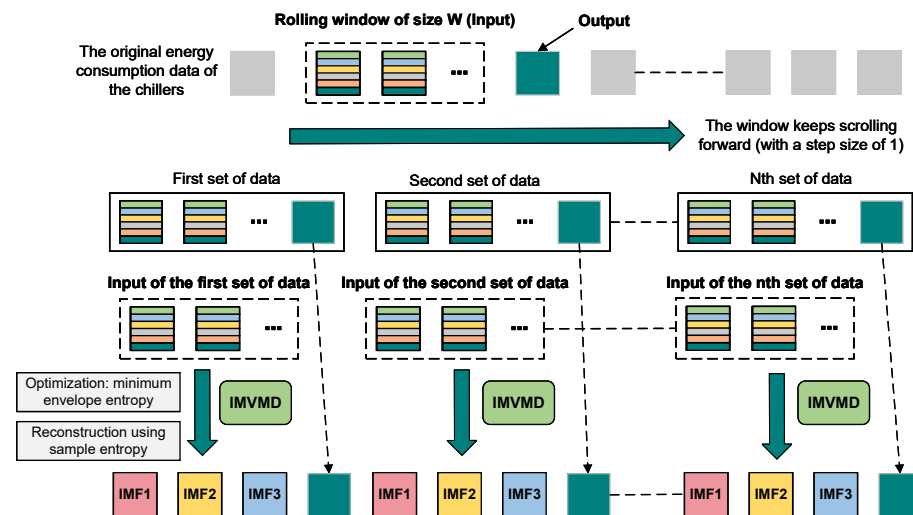


Figure 4. The rolling IMVMD decomposition framework for energy consumption data of chiller.

To further clarify the online prediction procedure and the data flow within each sliding window, the complete workflow of IMVMD-MADNet is described as follows. Let the original multivariate time series be $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$, where $\mathbf{x}_t \in \mathbb{R}^C$ is the observation of C channels at time t . Given a sliding window length L and forecasting horizon h , at any time step t :

1. Extract the current window $\mathcal{W}_t = \{\mathbf{x}_{t-L+1}, \dots, \mathbf{x}_t\}$, which uses only data up to time t .
2. Perform MVMD decomposition on $\mathcal{W}_t$ to obtain multivariate modes $\{\mathbf{u}_k^{(t)}\}_{k=1}^K$, where each $\mathbf{u}_k^{(t)} \in \mathbb{R}^{C \times L}$. This decomposition depends only on $\mathcal{W}_t$.
3. Compute the sample entropy of each mode and aggregate them into three frequency bands, using only the current window information.
4. MADNet takes the reconstructed features as input and outputs the prediction $\hat{\mathbf{x}}_{t+1:t+h}$.
5. The window advances to $\mathcal{W}_{t+1} = \{\mathbf{x}_{t-L+2}, \dots, \mathbf{x}_{t+1}\}$ and steps 2–4 are repeated.

At every time step t , the decomposition, reconstruction, and prediction steps all use only data within $\mathcal{W}_t$, which is a subset of $\{\mathbf{x}_1, \dots, \mathbf{x}_t\}$, ensuring that no future information beyond time t is used. The corresponding pseudocode has been added to the revised manuscript as Algorithm 2.

Algorithm 2 IMVMD-MADNet Framework.

Require: Multivariate time series $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$; window length L ; forecast horizon h ; mode number K ; sample entropy thresholds θ_H and θ_L

Ensure: Predictions $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_{L+1}, \dots, \hat{\mathbf{y}}_T\}$

- 1: **for** $t = L$ to $T - h$ **do**
 - 2: Extract the historical window $W_t \leftarrow \{\mathbf{x}_{t-L+1}, \dots, \mathbf{x}_t\}$
 - 3: Perform rolling MVMD decomposition $\{u_1, \dots, u_K\} \leftarrow \text{MVMD}(W_t, K)$
 - 4: **for** $k = 1$ to K **do**
 - 5: $SE_k \leftarrow \text{SampleEntropy}(u_k)$
 - 6: **end for**
 - 7: Reconstruct the high-, medium-, and low-frequency components $\{HF, MF, LF\} \leftarrow \text{Reconstruct}(\{u_k\}, \{SE_k\}, \theta_H, \theta_L)$
 - 8: Generate multi-scale prediction $\hat{\mathbf{y}}_{t+1:t+h} \leftarrow \text{MADNet}(HF, MF, LF)$
 - 9: **end for**
 - 10: **return** $\hat{\mathbf{Y}}$
-

2.10. Multi-Scale Aggregation Decomposition Network

Chiller energy-related variables exhibit significant nonstationarity under actual operating conditions and are characterized by multiscale temporal coupling and complex

dynamic evolution. To address these issues, a Multi-scale Aggregation Decomposition Network (MADNet) is proposed to capture the multiscale variation patterns of chiller energy consumption through hierarchical feature modeling. The network consists of an aggregation module and two structurally complementary prediction branches. These components deeply analyze the time-frequency characteristics of the energy consumption signal from the mode, channel, and time dimensions, thereby accomplishing short-term energy consumption prediction of chiller.

The MADNet is the core predictor of the proposed IMVMD-MADNet framework. The overall model comprises three core modules: the Mode-Channel Aggregator (MCA), the Wavelet Decomposition Predictor (WDP), and the Time-Domain Decomposition Predictor (TDP). The architecture of MADNet is presented in Figure 5.

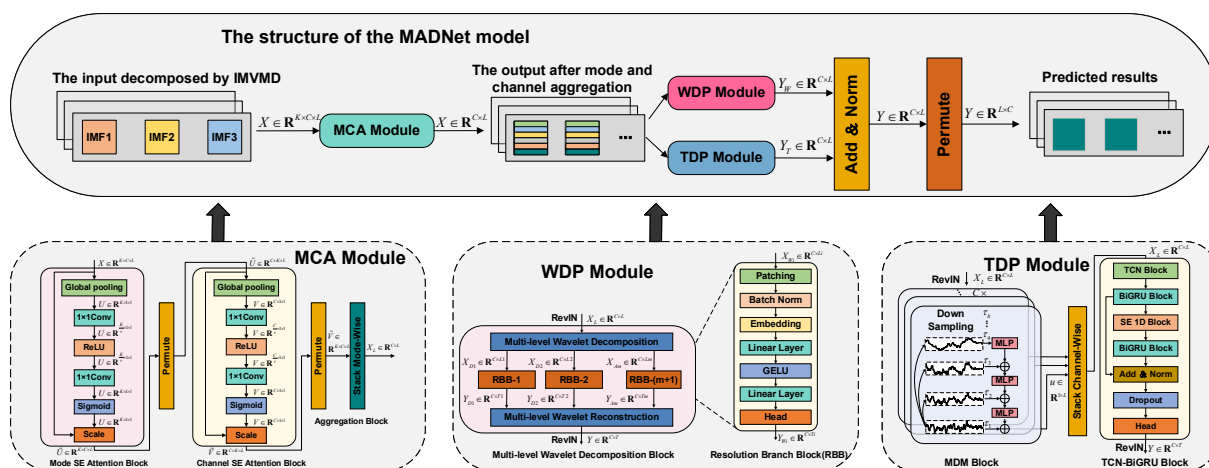


Figure 5. The network structure of MADNet.

2.10.1. MCA Module

The MCA module primarily focuses on multi-scale structure screening and information recalibration. Through IMVMD decomposition, the original chiller energy consumption parameters are decomposed into three groups of modal signals: the low-frequency component reflects the long-term load variation trend of the chiller system, the medium-frequency component captures system fluctuations caused by operational adjustments and control strategy changes, and the high-frequency component corresponds to anomalous disturbances and measurement noise.

Since the underlying operating mechanisms governing chiller energy consumption vary under different operating conditions, the relative contributions of different frequency bands and input variables to prediction also change accordingly. Therefore, the designed MCA module consists of two cascaded SE blocks. The first SE block operates from the mode dimension, adaptively learning the importance weights of different frequency bands under the current operating condition, thereby highlighting key temporal scales. The second SE block further models the relative contribution of system-related parameters from the channel dimension, enhancing core features and suppressing redundant information. After the two-level recalibration, a weighted aggregation is performed along the mode dimension, fusing the multi-modal signals into a single multi-scale feature sequence. This aggregated structure preserves the multi-scale structural information of the signal, significantly improving the discriminative power and physical interpretability of the feature representation.

2.10.2. WDP Module

The WDP module focuses on modeling the patterns of energy consumption variation from the perspectives of time-frequency structure and local patterns. First, multi-level

wavelet decomposition is performed on the aggregated feature sequence, mapping the energy consumption signal into different frequency sub-bands to explicitly separate long-term trend components from local fluctuation features [26].

Subsequently, the Resolution Branch Block (RBB) is employed to further model each frequency-band component. The sequence is partitioned along the temporal dimension into fixed-length patches, enabling the network to better capture local operating patterns and stage-wise variations in chiller operating conditions. These patches are then projected into a high-dimensional embedding space, where feature mixing is performed to enable information interaction across different temporal segments and feature combinations. Finally, the fused representation is mapped back to the time domain, providing important complementary time-frequency predictive information to the overall model.

2.10.3. TDP Module

The TDP module focuses on capturing the temporal causal dependencies and dynamic evolution of the chiller system. Chiller energy consumption is not merely a variation at the signal level, but rather the output of a dynamic system influenced by equipment inertia, control strategies, and multi-parameter coupling. To compensate for the limitations of wavelet-domain modeling in terms of causality and temporal dependency, this study further designs the TDP module, forming a complementary modeling structure with the WDP module.

The MDM Block is first employed to learn multiscale temporal representations and interactions, thereby expanding the temporal receptive field. The resulting features are then fed into a TCN block to capture short-term causal dependencies, thereby modeling the lagged impact of control actions on energy consumption. Subsequently, a BiGRU block is incorporated to further capture long-term dependencies across time scales, enhancing the model's ability to represent the operational periodicity and state inertia of the chiller. Finally, a 1D SE Block is applied to refine the output features. This time-domain decomposition module interprets the evolution of chiller energy consumption from a system dynamics perspective, strengthening the overall model's capability in temporal modeling.

3. Research Framework

Within the proposed forecasting framework, IMVMD is first applied to perform rolling decomposition, yielding three frequency-band components. Subsequently, Bayesian optimization is employed to adjust the hyperparameters of MADNet, followed by model training to generate the predictions. Specifically, the complete workflow of the IMVMD-MADNet method for short-term chiller energy consumption forecasting is illustrated in Figure 6.

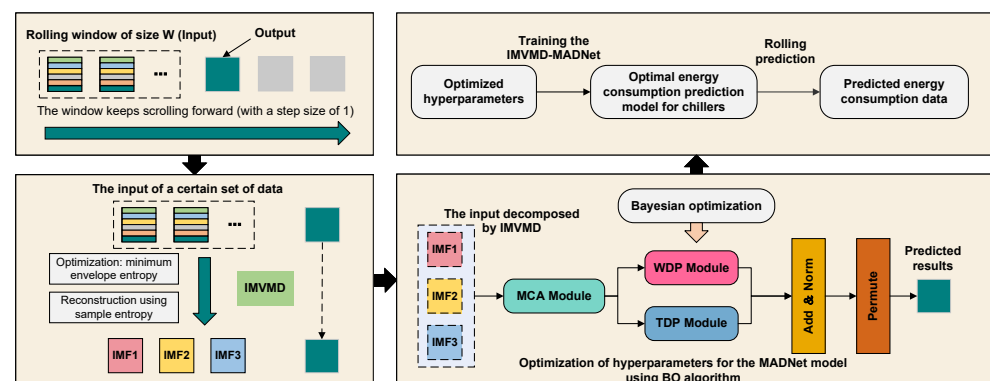


Figure 6. The process of the IMVMD-MADNet method for prediction of chiller energy consumption.

3.1. Data Description

This study focuses on an industrial plant located in Huizhou City, Guangdong Province, China, with particular emphasis on the energy consumption characteristics of the chiller units within its central air-conditioning system under actual operating conditions. The plant covers an area of approximately 60,000 square meters. Its cooling system employs 2 centrifugal chillers as the primary cold source, complemented by 3 chilled water pumps, 3 cooling water pumps, and 2 cooling towers, forming a complete central air-conditioning cooling loop. Data generated during system operation are collected in real-time by an on-site data acquisition system and stored centrally, providing a reliable foundation for subsequent analysis and modeling.

3.2. Data Preparation

3.2.1. Feature Selection

This study utilizes the original operational data from the plant throughout the entire year of 2024 as the research sample. The data sampling interval is 1 min, resulting in a total of 527,040 data points. The dataset comprises 13 measurement parameters closely related to chiller operating status and energy consumption variation, with the system parameters selected representing the operational information of the main pipelines. Chiller energy consumption is expressed in the form of power, and its distribution over the annual time scale is illustrated in Figure 7.

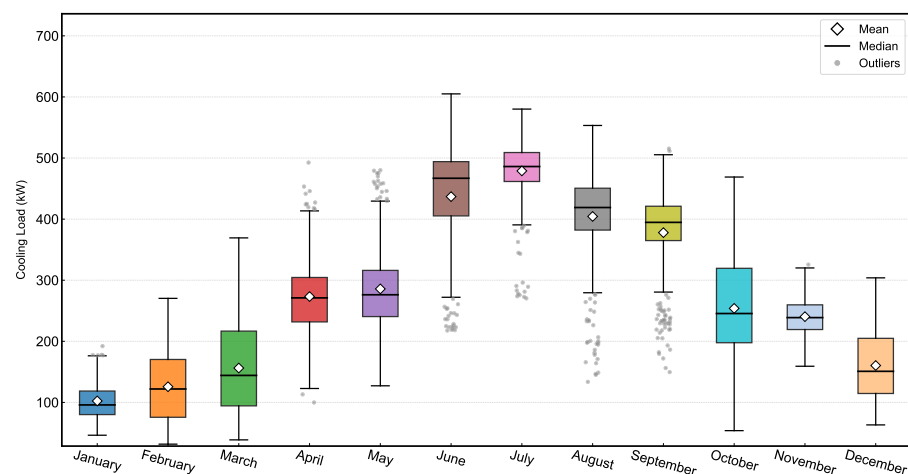


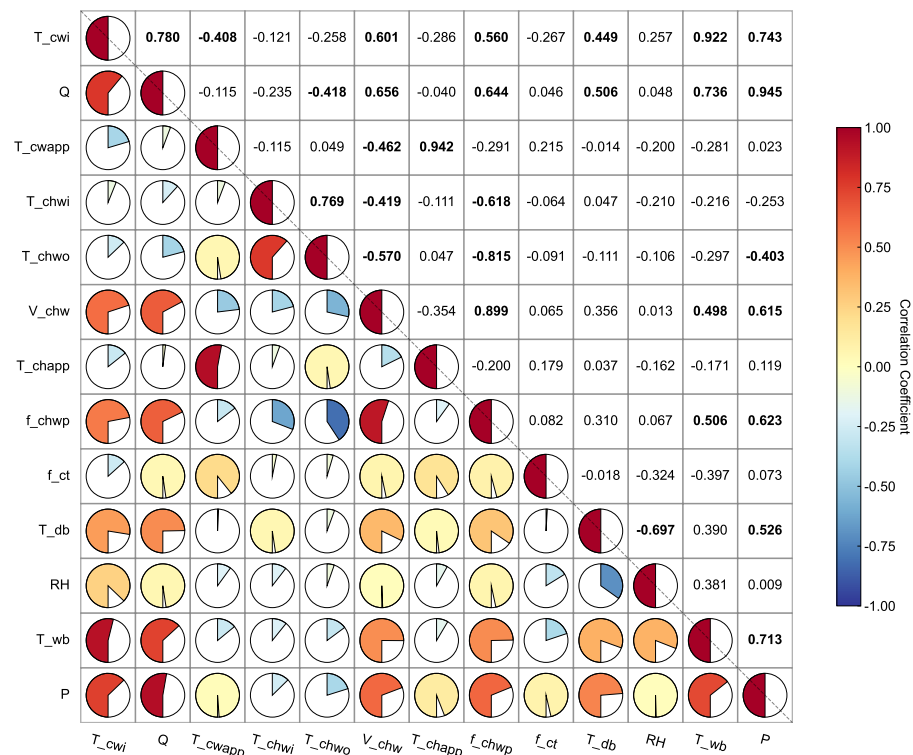
Figure 7. The distribution of chiller energy consumption across the annual time scale.

The input features are selected based on Pearson correlation analysis with a threshold of 0.4 for weak-correlation elimination. This method is chosen because it quantifies the linear relationship between each candidate variable and chiller power, allowing us to retain physically relevant variables that most directly influence the target while reducing input redundancy. The Pearson correlation coefficients were computed using only the training set to avoid data leakage. Figure 8 reports the results.

From the correlation analysis results, six exogenous variables exhibiting high correlation with chiller power were selected. Together with the chiller power itself, these form the input features for the prediction model. The output feature of the model is the chiller power. Finally, according to the Augmented Dickey–Fuller (ADF) test results in Table 2, all features are stationary at the 5% significance level and can be directly used in subsequent time series modeling [45,46].

Table 2. ADF test results of variables.

Variable Name	Symbolic	Units	ADF Statistic	<i>p</i> -Value
Cooling water inlet temperature	T_{cwi}	°C	−2.9622	0.0386
Cooling load	Q	kW	−6.6265	0
Chilled water outlet temperature	T_{chwo}	°C	−6.257	0
Chilled water flow rate	V_{chw}	m ³ /h	−8.698	0
Outdoor dry-bulb temperature	T_{db}	°C	−5.7929	0
Outdoor wet-bulb temperature	T_{wb}	°C	−3.3989	0.011
Power (Energy consumption)	P	kW	−6.6671	0

**Figure 8.** The correlation analysis results.

3.2.2. Abnormal Data Cleaning

In the real-time operational data of chiller systems, in addition to common anomalies caused by sensor failures or data transmission losses, structural abnormalities due to switching of operating conditions are also prevalent. The system's operating states are classified into four typical conditions: dual-chiller operation, single-chiller 1 operation, single-chiller 2 operation, and no-chiller operation. Taking chiller power as an example, it typically exhibits abrupt transients and short-term fluctuations during chiller start-up, shutdown, and operating-condition transitions. Such data do not represent stable operating behavior and may degrade prediction performance if included in model training. Therefore, the transient segments associated with operating-condition transitions are identified and excluded during data preprocessing.

To this end, this study determines the start–stop status of chillers by setting working thresholds for chiller power, constructs corresponding start–stop flags, and consequently identifies the system's operational condition at every time step. According to the condition identification results, only data under stable single-chiller or dual-chiller operation are retained as valid samples. The transition segments during start–stop switching and the segments with no-chiller operation are treated as abnormal data. Specifically, data from no-chiller operation segments are directly removed, while the start–stop transition segments

and abnormal data caused by sensor faults are replaced via interpolation to ensure the quality of the data. The effect of the specific abnormal data cleaning is illustrated in Figure 9.

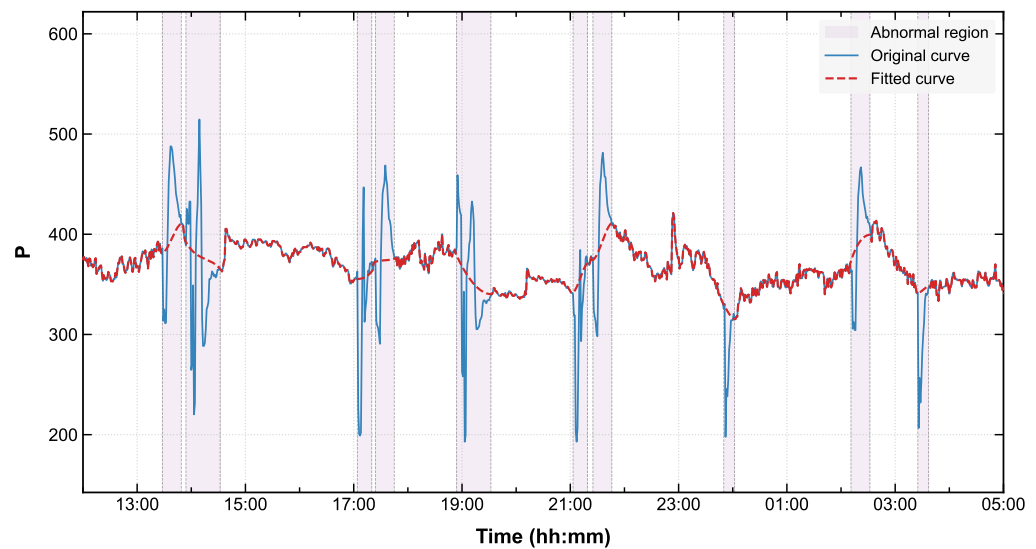


Figure 9. The effect of the specific abnormal data cleaning.

The original dataset contains 527,040 data points (full-year 2024, 1-min sampling interval). Approximately 12% of the data was affected by the cleaning process: about 3% (~15,811 data points) correspond to no-chiller operation segments that were directly removed, and about 9% (~47,434 data points) correspond to chiller switching transition segments that were interpolated to maintain temporal continuity. After removing the no-chiller segments, approximately 511,229 valid data points remain.

3.2.3. Normalization Processing

To meet the training requirements of the chiller energy consumption prediction model, the anomaly-cleaned data are further subjected to normalization. Features such as chiller power, temperature, and flow rate differ significantly in units and value ranges; if input directly into the model, they would hinder effective feature representation and model convergence. By applying a unified scale transformation to each input feature so that their value distributions lie in similar intervals, the model can establish a more accurate nonlinear mapping from multi-source operational parameters to chiller energy consumption. Normalization is performed using the training set's statistical parameters and is consistently applied to the validation and test sets. This prevents information leakage and ensures the reliability of model evaluation results. This study adopts Min-Max normalization, expressed as follows:

$$\tilde{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \quad (32)$$

where $\tilde{x}$ is the normalized feature value, x is the original feature value, and $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values of that feature in the training set, respectively.

3.2.4. Evaluation Indicators

To comprehensively evaluate the prediction performance of the chiller energy consumption model, this study employs multiple evaluation metrics for quantitative analysis of the forecasting results. The corresponding definitions and formulas are listed in Table 3.

Table 3. Model evaluation indicators.

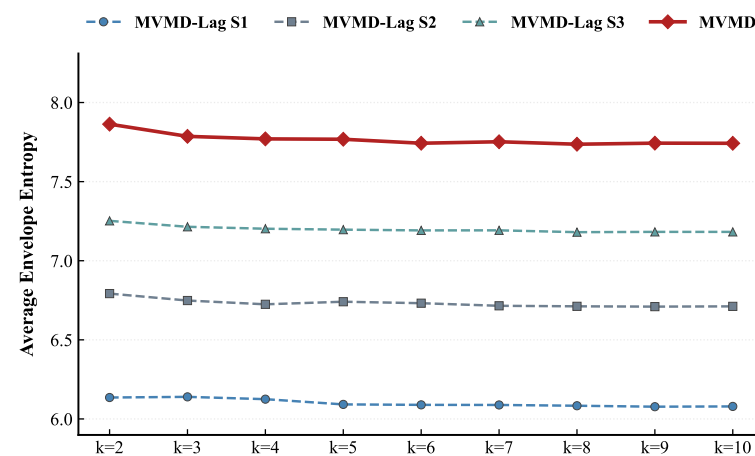
Indicator	Definition	Equation ¹
MAE	Mean Absolute Error	$MAE = \frac{1}{N} \sum_{t=1}^N y_t - \hat{y}_t $
RMSE	Root Mean Squared Error	$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (y_t - \hat{y}_t)^2}$
SMAPE	Symmetric Mean Absolute Percentage Error	$SMAPE = \frac{100\%}{N} \sum_{t=1}^N \frac{ y_t - \hat{y}_t }{(y_t + \hat{y}_t)/2}$
R ²	R-squared	$R^2 = 1 - \frac{\sum_{t=1}^N (\hat{y}_t - \bar{y})^2}{\sum_{t=1}^N (y_t - \bar{y})^2}$
PeakMAE	Peak Mean Absolute Error	$PeakMAE = \frac{1}{N_p} \sum_{t \in \mathcal{P}} y_t - \hat{y}_t $
RampMAE	Ramp Mean Absolute Error	$RampMAE = \frac{1}{N_r} \sum_{t \in \mathcal{R}} y_t - \hat{y}_t $

¹ Note: y_t denotes the actual value, $\hat{y}_t$ denotes the predicted value, $\bar{y} = \frac{1}{N} \sum_{t=1}^N y_t$ denotes the mean of the actual values. $\mathcal{P}$ denotes the set of peak-load time indices, $N_p = |\mathcal{P}|$ is the number of peak samples. $\mathcal{R}$ denotes the set of ramp-event time indices, and $N_r = |\mathcal{R}|$ is the number of ramp samples. MAE, RMSE, PeakMAE and RampMAE are expressed in kW, while SMAPE is expressed as a percentage.

4. Case Study

4.1. Experiment 1: The Decomposition Performance of the IMVMD Module

In the rolling decomposition experiments, three decomposition window lengths (48, 96, and 192) were evaluated and denoted as Lag S1, Lag S2, and Lag S3, respectively. Subsequently, these rolling decomposition schemes were compared with the conventional full-sequence decomposition method using a long test sequence. For each window setting, the penalty factor α was fixed at 2000, and the average envelope entropy was computed for mode numbers K ranging from 2 to 10 to evaluate the decomposition quality under different parameter combinations. The result is illustrated in Figure 10.

**Figure 10.** The average envelope entropy corresponding to different modal numbers k .

As the mode number K increases, the average envelope entropy decreases rapidly at first and then gradually approaches a stable value across all decomposition window sizes. When k is small, the decomposition is insufficient, leading to noticeable mode mixing among the components. As k increases, the main signal constituents are progressively separated, and the average envelope entropy decreases significantly. Once k exceeds a certain threshold, further increasing the mode number does not bring a noticeable reduction in entropy; in some cases, a slight increase is even observed, indicating that over-decomposition introduces redundant modes. Based on the above analysis, and by comprehensively considering both the lower values of average envelope entropy and the inflection points in its variation, $k = 6$ is ultimately selected as the optimal mode number.

Further comparison of the results under different decomposition windows reveals that, compared with global decomposition, small-scale window decomposition overall yields lower average envelope entropy, and its entropy- k curve shows a clearer declining

trend. This rolling local decomposition approach not only effectively avoids data leakage but also enhances the concentration of information within each local time window, thereby producing decomposition components with stronger structure and lower complexity. These components provide more favorable input features for the subsequent prediction model.

Following IMVMD decomposition, average sample entropy is introduced to reconstruct the decomposed components, thereby standardizing the model inputs. Specifically, the decomposed components are grouped into high-, medium-, and low-frequency components according to three sample entropy intervals: greater than 0.7, between 0.2 and 0.7, and less than 0.2. Figure 11 illustrates the reconstruction results for the Lag S3 window size, where different color shades represent the dynamic complexity of each frequency component over time. For chiller power, the low-frequency component mainly captures its long-term operational trend, with a stable dynamic structure that is minimally affected by random disturbances. The medium-frequency component reflects periodic variations associated with load adjustment and operating-state transitions, whereas the high-frequency component primarily captures short-term fluctuations and transient disturbances. The reconstructed inputs effectively separate different frequency characteristics, demonstrating a clear hierarchical structure and good interpretability.

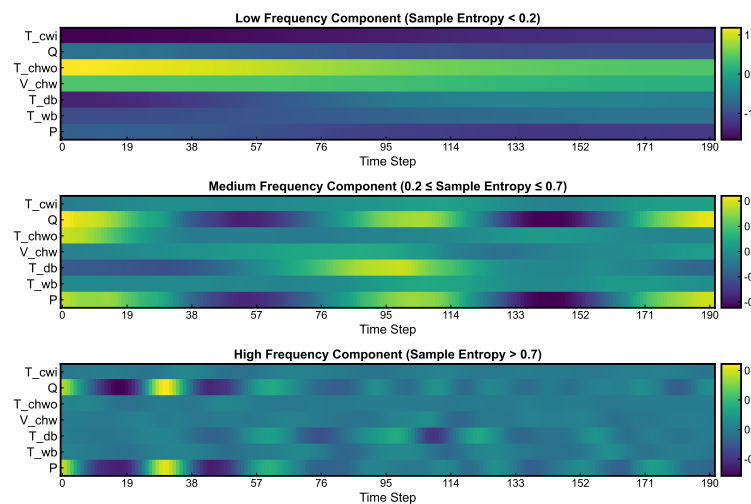


Figure 11. The reconstruction results for Lag S3 window size.

4.2. Experiment 2: Optimization of Hyperparameters Using the BO Algorithm

To ensure consistency between the sliding process of the rolling decomposition window and the training process of the prediction model, the length of the model input sequence was set equal to the rolling decomposition window length, corresponding to three time-step settings: Lag S1, Lag S2, and Lag S3. Given the time-series characteristics of the forecasting task, the dataset was partitioned strictly in chronological order. The original data was sequentially divided into training, validation and test sets at a fixed ratio of 0.7:0.1:0.2, which eliminates random splitting. The fixed validation set (10% of total data) is used for early stopping and the low-data sensitivity analysis.

It is worth noting that prior to dataset splitting, a continuous one-week period in May was extracted from the original data to serve as an independent test sample for subsequent specific experiments. This period falls within a seasonal transition phase, during which chiller operation gradually shifts from relatively stable to higher-load conditions. It contains both steady operating characteristics and certain dynamic fluctuations, making it highly representative. Moreover, this period does not correspond to extreme high- or low-load operating ranges, which helps avoid condition bias and enables a more objective evaluation of the model's predictive capability in typical operational scenarios.

To address model parameter configuration, Bayesian optimization was employed to automatically search for model hyperparameters, aiming to enhance overall performance in the chiller energy consumption forecasting task. The hyperparameter optimization was performed exclusively on the training data using a hold-out validation strategy, where 70% of the training data were used for model training and the remaining 30% served as the validation subset for performance evaluation. The threshold sensitivity analysis presented in the following subsection is based on this internal validation subset, which corresponds to approximately 21% of the total dataset. No independent test data were involved in the hyperparameter optimization process. For a given forecasting horizon, 100 independent training runs were conducted on multiple candidate hyperparameter combinations, with prediction error on the validation subset used as the optimization objective to obtain the corresponding optimal hyperparameter set. All competing models were optimized using exactly the same Bayesian Optimization procedure as IMVMD-MADNet, with identical search spaces for shared hyperparameters (sequence length, batch size, hidden dimension, dropout rate, learning rate, and optimizer). Figure 12 illustrates the optimization results when the forecasting horizon is set to 12. The optimal hyperparameter combination obtained includes: input sequence length = 48, batch size = 32, hidden layer dimension = 120, dropout rate = 0.4, learning rate = 0.0004, with Adam selected as the optimizer. Detailed configurations of the main model parameters are provided in Table 4.

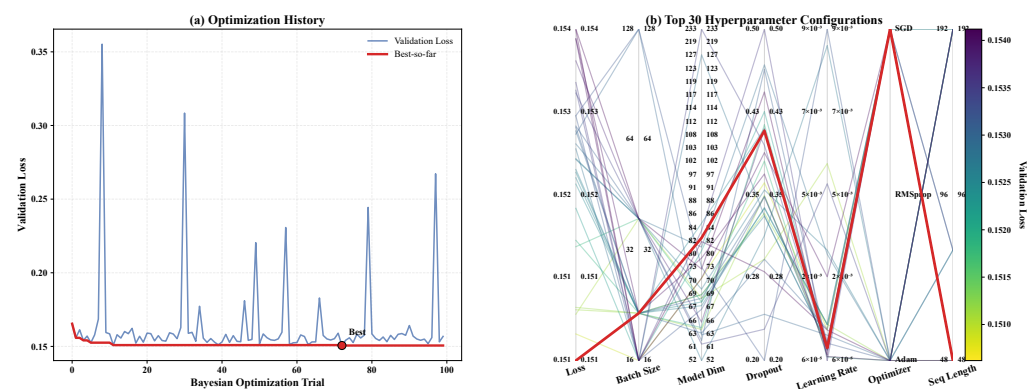


Figure 12. The optimized result of the BO algorithm.

To provide full transparency on the model architecture, we analyze the tensor flow and trainable parameters using a representative configuration (batch size = 32, channels = 7, Input sequence length = 48, forecasting horizon = 12, model dimension = 120). Table 5 summarizes the input/output tensor shapes, trainable parameters, and their proportions.

Table 4. Detailed configurations of the main model parameters.

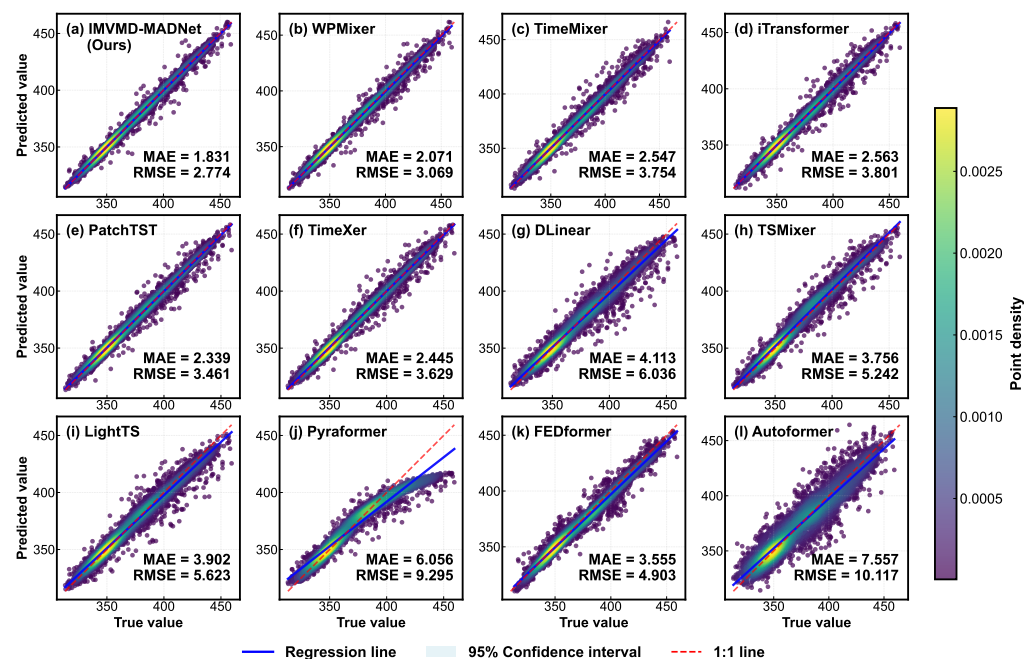
Model	Parameter Settings
IMVMD	Reconstructed mode number: 3, Decomposition window: {48, 96, 192}.
MCA	Number of SE Block layers: 2.
WDP	Wavelet function: db2, Wavelet decomposition scale: 2, Expansion factor: 5, Patch length: 16, patch stride: 8.
TDP	Number of TCN layers: 1, Number of BiGRU layers: 2, Number of SE Block layers: 1, Number of downsampling layers: 3, Downsampling window size: 2.
Other	Epoch: 50, Loss function: RMSE, Sequence length: {48, 96, 192}, Learning rate: [0.0001, 0.01], Batch size: {16, 32, 64, 128}, Dropout rate: [0.1, 0.5], Optimizer: {Adam, rmsprop, sgd}, Hidden layer dimension: [50, 256].

Table 5. Module tensor shape and trainable parameter statistics of IMVMD-MADNet.

Module	Input Tensor Shape	Output Tensor Shape	Trainable Parameters	Proportion of Total Parameters
Rolling IMVMD	$R^{32 \times 7 \times 48}$	$R^{32 \times 3 \times 7 \times 48}$	—	0%
MCA Module	$R^{32 \times 3 \times 7 \times 48}$	$R^{32 \times 7 \times 48}$	139	<0.1%
WDP Wavelet Branch	$R^{32 \times 7 \times 48}$	$R^{32 \times 80 \times 48}$	271,280	32.10%
TDP Time Series Branch	$R^{32 \times 7 \times 48}$	$R^{32 \times 240 \times 48}$	559,215	66.20%
Dual-branch Fusion and Prediction Layer	$R^{32 \times 120 \times 48}, R^{32 \times 240 \times 48}$	$R^{32 \times 12}$	14,052	1.70%
IMVMD-MADNet (Total)	—	—	844,686	100%

4.3. Experiment 3: Comparison of Predictive Effectiveness with Other Models

To further validate the overall performance of IMVMD-MADNet, this study conducted comparative experiments between the proposed model and several existing mainstream forecasting models (WPMixer [26], TimeMixer [25], iTransformer [29], PatchTST [47], TimeXer [48], DLinear [49], TSMixer [50], LightTS [51], Pyraformer [24], FEDformer [27], Autoformer [42]). Taking single-step prediction as an example, Figure 13 presents density scatter plots of the predictions of the 12 models versus the true chiller power values. The scatter points of IMVMD-MADNet cluster closely around the ideal prediction line $y = x$, and its regression line almost overlaps the ideal line. The results reveal that the model outputs match well with the true values and perform best in capturing dynamic variations of chiller power. Furthermore, most mainstream models (WPMixer, TimeMixer, iTransformer, PatchTST, TimeXer, DLinear, TSMixer, LightTS, FEDformer) also exhibit certain predictive capabilities, with their scatter points relatively concentrated in the medium-value range. However, their regression lines still deviate from the ideal prediction line to varying degrees. In comparison, data points of Pyraformer and Autoformer show obvious dispersion. Meanwhile, their regression lines drift far from the ideal line, demonstrating poorer fitting capability when predicting chiller power.

**Figure 13.** Density scatter plots of the predictions of the 12 models versus the true chiller power values.

The reported results are not based on a single run. In our experiments, each deep learning model was independently trained and evaluated using five different random seeds (10, 20, 30, 40, and 50), and the reported metrics are the average across these five runs. In addition, the statistical significance of the performance differences was assessed using the Diebold-Mariano (DM) test [52]. All DM statistics are negative and the corresponding p-values are well below 0.001, confirming that the performance advantages of IMVMD-MADNet over all baseline models are statistically significant and not due to random chance.

Table 6 summarizes the quantitative evaluation results of IMVMD-MADNet and the 11 comparison models under different forecasting horizons, with the best and second-best metrics highlighted. Figures 14 and 15 illustrate the visualization of relevant evaluation metrics. The forecasting horizons were set to 1, 6, 12, 24, 48, and 96 steps. The average values of each metric were also computed to fully evaluate the overall prediction capability of the models. It should be noted that to ensure fairness in the comparative experiments, the results for all models are reported as their best performance among three input sequence length settings: 48, 96, and 192.

Table 6. The quantitative evaluation results of IMVMD-MADNet and the 11 comparison models under different forecasting horizons.

Model	Metric (kW)	Time Interval						Avg
		1	6	12	24	48	96	
IMVMD-MADNet (Ours)	MAE	1.831	4.828	6.566	8.476	10.572	12.850	7.521
	RMSE	2.774	7.560	9.936	12.322	14.841	17.513	10.824
WPMixer	MAE	2.071	5.273	6.716	8.570	10.812	13.536	7.830
	RMSE	3.069	8.224	10.144	12.415	15.242	18.581	11.279
TimeMixer	MAE	2.547	5.312	6.866	8.611	10.949	13.681	7.994
	RMSE	3.754	8.269	10.256	12.411	15.325	18.676	11.449
iTransformer	MAE	2.563	5.281	6.772	8.683	10.962	13.301	7.927
	RMSE	3.801	8.129	10.102	12.368	15.151	18.037	11.265
PatchTST	MAE	2.339	5.197	6.729	8.566	10.876	13.519	7.871
	RMSE	3.461	8.016	10.036	12.362	15.239	18.450	11.261
TimeXer	MAE	2.445	5.446	6.752	8.665	10.832	13.148	7.881
	RMSE	3.629	8.157	9.942	12.323	14.990	17.913	11.159
DLinear	MAE	4.113	5.956	7.255	8.939	11.503	14.403	8.695
	RMSE	6.036	8.854	10.543	12.694	15.850	19.439	12.236
TSMixer	MAE	3.756	6.385	7.870	9.540	12.158	14.876	9.098
	RMSE	5.242	9.103	10.888	12.825	15.914	19.186	12.193
LightTS	MAE	3.902	7.274	8.361	10.019	12.563	15.255	9.562
	RMSE	5.623	10.271	11.759	13.871	17.004	20.394	13.154
Pyraformer	MAE	6.056	8.682	10.749	11.755	13.720	17.544	11.418
	RMSE	9.295	12.350	15.183	16.045	18.645	23.661	15.863
FEDformer	MAE	3.555	6.027	7.254	10.927	13.427	16.923	9.686
	RMSE	4.903	8.719	10.350	14.648	17.634	21.883	13.023
Autoformer	MAE	7.557	11.264	11.156	11.634	12.855	15.809	11.713
	RMSE	10.117	14.933	14.884	15.785	17.300	20.981	15.667

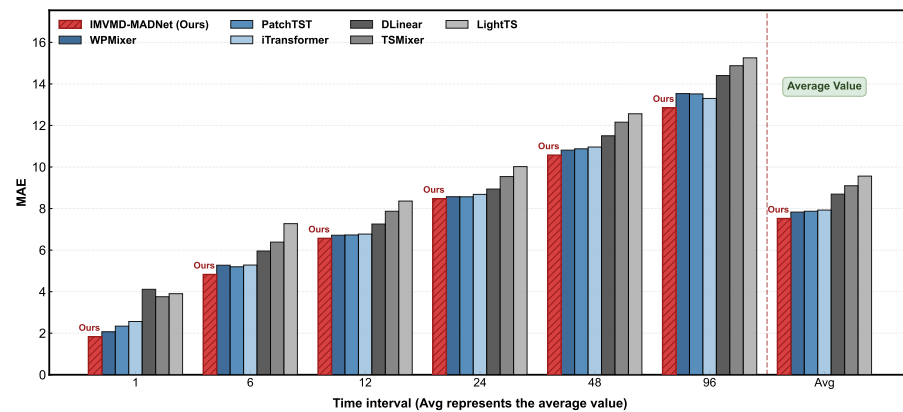


Figure 14. The corresponding visualization of the MAE.

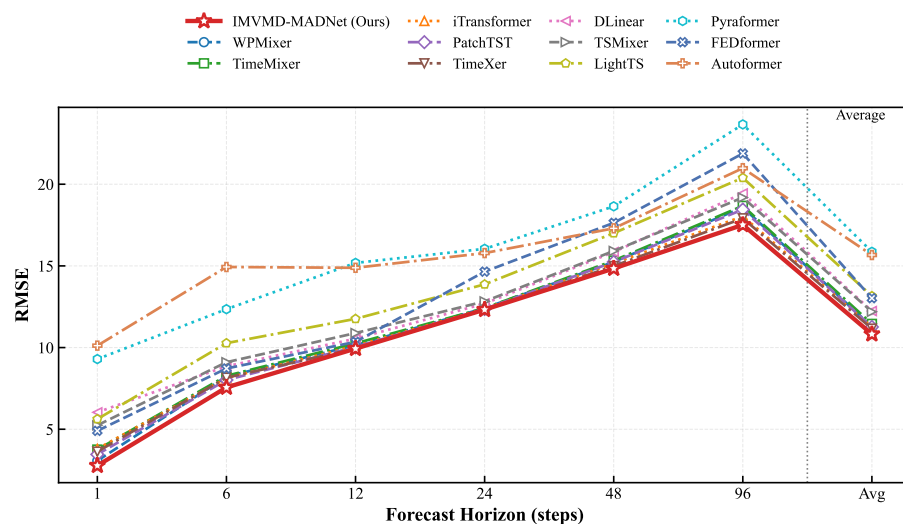


Figure 15. The corresponding visualization of the RMSE.

From the perspective of average metric values, Pyraformer and Autoformer exhibit relatively poor overall predictive performance, with MAE values of 11.418 and 11.713 and RMSE values of 15.863 and 15.667, respectively. Among the mainstream comparison models, WPMixer performs relatively well, achieving an MAE of 7.830 and an RMSE of 11.279. In comparison, IMVMD-MADNet achieves the best predictive performance, with an MAE of 7.521 and an RMSE of 10.824. According to the improvement rate defined in Equation (33), IMVMD-MADNet reduces the MAE by 34.1%, 35.8%, and 3.95% compared with Pyraformer, Autoformer, and WPMixer, respectively, while reducing the RMSE by 31.8%, 30.9%, and 4.03%, respectively.

$$\text{Improvement (\%)} = \frac{M_b - M_p}{M_b} \times 100\%, \quad (33)$$

where M_b and M_p denote the evaluation metric of the baseline model and the proposed model, respectively. Since MAE and RMSE are error-based metrics, a larger improvement value indicates better predictive performance.

A further analysis of predictive performance under different forecasting horizons reveals that IMVMD-MADNet achieves the best results under all six forecasting horizons. The corresponding MAE values are [1.831, 4.828, 6.566, 8.476, 10.572, 12.850], and RMSE values are [2.774, 7.560, 9.936, 12.322, 14.841, 17.513], indicating strong predictive stability across different time scales. Especially in medium- and long-term forecasting tasks, while the prediction errors of some comparison models increase significantly as the forecasting horizon lengthens, the growth rates of MAE and RMSE for IMVMD-MADNet remain rela-

tively small. This demonstrates its stronger capability in modeling long-term dependencies and multi-scale dynamic features, effectively mitigating the error accumulation caused by increasing forecast horizons.

Moreover, the choice of forecasting horizon also influences model performance to some extent. Taking PatchTST and TimeXer as examples, when the forecast horizon is 6, PatchTST achieves MAE and RMSE values of 5.197 and 8.016, respectively, outperforming TimeXer's 6.385 and 9.103 and showing better short-term predictive ability. However, when the forecast horizon increases to 96, PatchTST yields MAE and RMSE values of 13.536 and 18.581, while TimeXer obtains 13.148 and 17.913, indicating that TimeXer possesses a relative advantage in long-term forecasting tasks.

4.4. Experiment 4: Ablation Experiments for the Model's Ability

To investigate the contribution of each functional module in the IMVMD-MADNet model for chiller power prediction, an additional continuous one-week dataset from May was reserved as an independent test set and excluded from model training, and an ablation study was conducted on this dataset. Considering the practical engineering requirement for short-term prediction accuracy, the forecasting horizon in this experiment was set to 6, and comparisons were performed under three input sequence lengths: Lag S1, Lag S2, and Lag S3 (corresponding to 48, 96, and 192 steps, respectively). Here, IMVMD-MCA represents the integrated decomposition-aggregation module, WDP denotes the wavelet-based prediction branch, and TDP refers to the time-domain-based prediction branch.

Table 7 summarizes the prediction evaluation results of different models under the three input sequence lengths. As baseline models, WPMixer and PatchTST achieved relatively better predictive performance under the Lag S1 condition. Specifically, WPMixer obtained MAE, RMSE, R^2 , and SMAPE values of 5.3633, 8.1427, 0.9224, and 1.3686%, respectively, while PatchTST obtained 5.2032, 7.9033, 0.9270, and 1.3277%. Furthermore, the two single-branch prediction modules, WDP and TDP, also performed best under Lag S1, with WDP achieving MAE, RMSE, R^2 , and SMAPE of 5.3752, 8.1711, 0.9219, and 1.3712%, and TDP achieving 5.4102, 8.1511, 0.9223, and 1.3803%. This indicates that in the chiller power prediction task, a single decomposition-prediction branch, while structurally simpler, can still achieve prediction accuracy close to that of the baseline models and maintain relatively stable performance across different time scales.

After further introducing the IMVMD-MCA module, both IMVMD-MCA-WDP and IMVMD-MCA-TDP showed improved prediction performance compared to their corresponding single-branch counterparts under all three input sequence lengths. Under Lag S1, IMVMD-MCA-WDP achieved MAE, RMSE, R^2 , and SMAPE of 5.3807, 8.1111, 0.9231, and 1.3731%, while IMVMD-MCA-TDP achieved 5.4062, 8.1128, 0.9230, and 1.3798%. These results suggest that the decomposition-aggregation mechanism can more fully exploit multi-scale temporal features from multivariate chiller operation data, providing more discriminative feature representations for subsequent prediction, thereby alleviating feature aliasing issues under complex operating conditions to some extent.

Building on this, the dual-branch structure combining WDP and TDP (WDP-TDP) further enhanced prediction performance across all three input sequence lengths, with MAE values of 5.0749, 5.2106, and 5.2752; RMSE values of 7.6219, 7.7407, and 7.8546; R^2 values of 0.9321, 0.9299, and 0.9279; and SMAPE values of 1.2934%, 1.3303%, and 1.3474%, respectively. This result further verifies the complementarity of frequency-domain and time-domain decomposition in chiller power prediction, effectively compensating for the limitations of single-branch modeling in capturing complex dynamic features.

Finally, the complete IMVMD-MADNet achieved the best performance under Lag S1, Lag S2, and Lag S3 conditions, with optimal MAE (4.9720, 5.0394, 5.1367), RMSE (7.5669, 7.6331, 7.6992), R^2 (0.9330, 0.9319, 0.9307), and SMAPE (1.2690%, 1.1287%, 1.3123%), outperforming all comparison models.

Table 7. The prediction evaluation results of different models under the three input sequence lengths.

Lag Steps	Model	Metric			
		MAE (kW)	RMSE (kW)	R^2	SMAPE (%)
Lag S1	WPMixer	5.3633	8.1427	0.9224	1.3686%
	PatchTST	5.2032	7.9033	0.9270	1.3277%
	WDP	5.3752	8.1711	0.9219	1.3712%
	TDP	5.4102	8.1511	0.9223	1.3803%
	IMVMD-MCA-WDP	5.3807	8.1111	0.9231	1.3731%
	IMVMD-MCA-TDP	5.4062	8.1128	0.9230	1.3798%
	WDP-TDP	5.0749	7.6219	0.9321	1.2934%
	IMVMD-MADNet (Ours)	4.9720	7.5669	0.9330	1.2690%
Lag S2	WPMixer	5.4658	8.2352	0.9224	1.3947%
	PatchTST	5.4509	8.0825	0.9236	1.3933%
	WDP	5.4025	8.1275	0.9228	1.3786%
	TDP	5.4486	8.1424	0.9225	1.3900%
	IMVMD-MCA-WDP	5.3986	8.1303	0.9227	1.3778%
	IMVMD-MCA-TDP	5.4149	8.1233	0.9230	1.3819%
	WDP-TDP	5.2106	7.7407	0.9299	1.3303%
	IMVMD-MADNet (Ours)	5.0394	7.6331	0.9319	1.1287%
Lag S3	WPMixer	5.4463	8.1470	0.9224	1.3902%
	PatchTST	5.4041	8.0199	0.9248	1.3801%
	WDP	5.4689	8.1698	0.9220	1.3953%
	TDP	5.4792	8.2217	0.9225	1.3983%
	IMVMD-MCA-WDP	5.3909	8.1193	0.9229	1.3758%
	IMVMD-MCA-TDP	5.4305	8.1658	0.9220	1.3859%
	WDP-TDP	5.2752	7.8546	0.9279	1.3474%
	IMVMD-MADNet (Ours)	5.1367	7.6992	0.9307	1.3123%

Figure 16 presents a comparison of the prediction results for five representative models under the three input sequence lengths. Overall, IMVMD-MADNet maintained stable and consistent performance advantages across different input sequence length settings, indicating that the proposed rolling IMVMD decomposition module, the MCA aggregation module, and the dual-branch structure composed of WDP and TDP exhibit good synergy and strong generalization capability in the chiller power prediction task.

Furthermore, the choice of input sequence length (decomposition window length) has a significant impact on chiller power prediction accuracy. Shorter input sequences facilitate the capture of short-term dynamic changes during chiller operation, whereas longer input sequences provide more historical information for modeling medium- and low-frequency characteristics and long-term operational trends. The experimental results show that IMVMD-MADNet maintained relatively stable prediction performance under different input sequence lengths, suggesting that the proposed rolling decomposition and multi-scale modeling structure mitigates, to some extent, the impact of input length variation on model performance. Therefore, in practical applications, the input sequence length should be selected appropriately based on the time scale of the forecasting task to strike a balance prediction accuracy and network complexity.

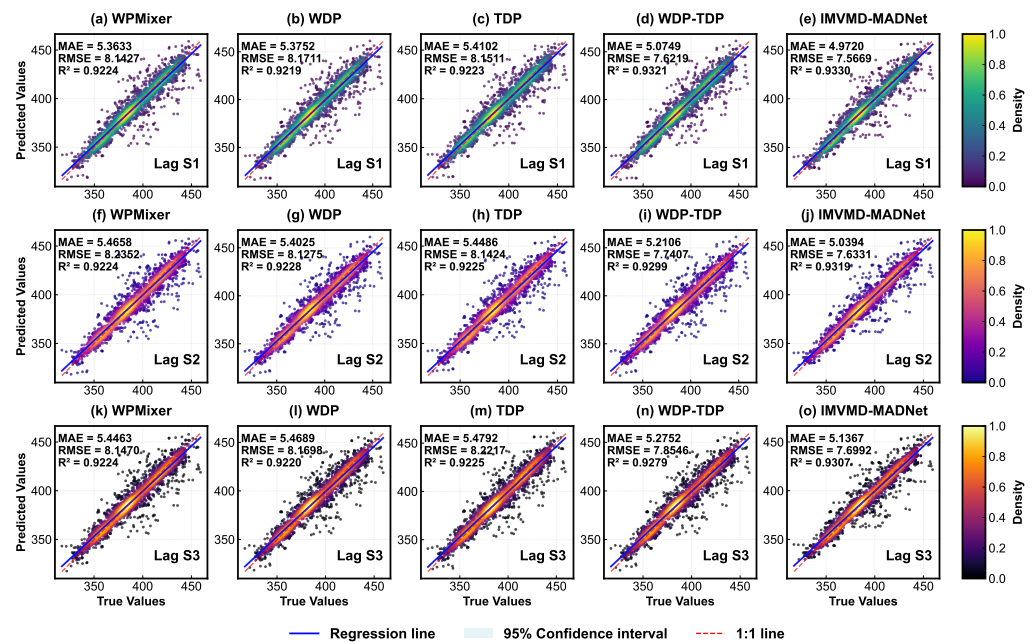


Figure 16. The comparison of the prediction results for five representative models under the three input sequence lengths.

4.5. Experiment 5: The Potential of Long-Term Prediction by the Model

To evaluate the long-term forecasting capability of IMVMD-MADNet, WPMixer, TimeMixer, and TimeXer, which achieved the most stable performance in the previous comparative experiments, were selected as baseline models. Multi-step forecasting experiments were conducted using the held-out one-week chiller operation dataset from May. The forecasting horizons were set to 1, 6, 12, 24, 48, 96, and 192 steps. For each model, the best performance obtained with the three input sequence lengths (48, 96, and 192) was selected to ensure a fair comparison. The evaluation results are summarized in Table 8.

Table 8. Model evaluation results of multi-step experiments.

Model	Metric	Time Interval						
		1	6	12	24	48	96	192
IMVMD-MADNet (Ours)	MAE (kW)	1.919	4.975	6.710	8.817	11.076	12.835	13.543
	RMSE (kW)	2.827	7.557	9.829	12.468	15.167	16.730	17.670
	SMAPE (%)	0.491	1.270	1.713	2.254	2.839	3.295	3.473
WPMixer	MAE (kW)	2.166	5.363	7.038	8.969	11.313	14.036	16.487
	RMSE (kW)	3.244	8.143	10.154	12.480	15.191	18.363	21.636
	SMAPE (%)	0.553	1.369	1.795	2.229	2.901	3.603	4.225
TimeMixer	MAE (kW)	2.437	5.493	7.082	8.967	11.225	14.132	16.645
	RMSE (kW)	3.520	8.304	10.234	12.494	15.095	18.373	21.643
	SMAPE (%)	0.622	1.401	1.806	2.291	2.877	3.363	4.255
TimeXer	MAE (kW)	2.538	5.278	6.876	8.872	11.143	13.332	15.431
	RMSE (kW)	3.639	7.908	9.895	12.318	14.942	17.360	20.081
	SMAPE (%)	0.650	1.347	1.757	2.227	2.860	3.425	3.954

As shown in the trends of MAE versus forecasting horizon in Figure 17 and RMSE versus forecasting horizon in Figure 18, the prediction errors of all models were relatively close under short-term forecasting scenarios (horizons of 1 and 6). This indicates that when chiller operating conditions change relatively smoothly and short-term dependencies dom-

inate, mainstream time-series forecasting models can all reasonably capture the immediate variations in chiller power. With the forecast horizon extending to 12, 24, and 48 steps, the MAE values of WPMixer, TimeMixer, and TimeXer show a more pronounced increasing trend, reflecting their gradually limited ability to characterize the multi-scale variation features of chiller power during medium-term forecasting. In contrast, the increase in prediction error for IMVMD-MADNet is relatively smaller, gradually demonstrating the advantage of its multi-scale decomposition and aggregation structure in mitigating error propagation in medium-term predictions.

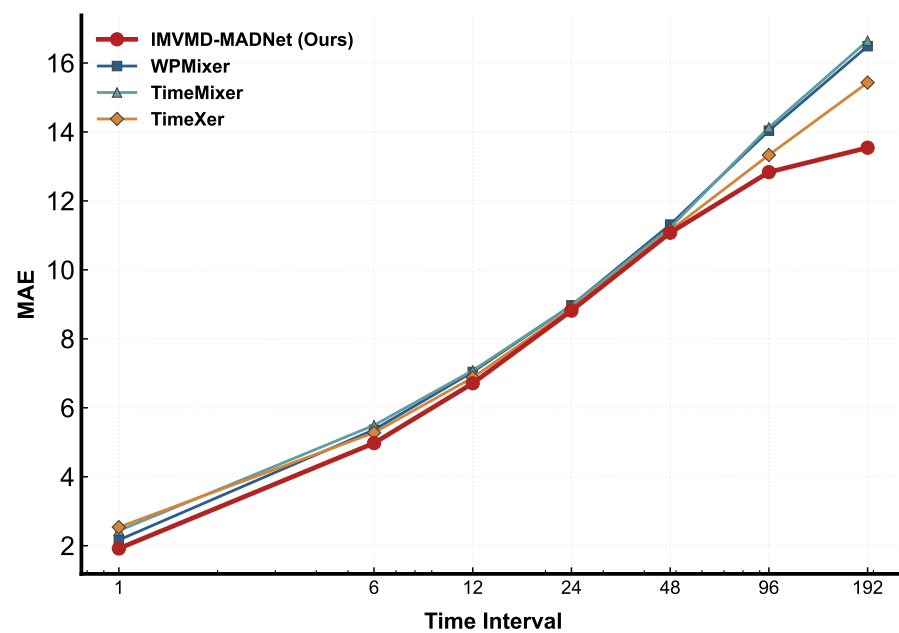


Figure 17. The trend of MAE versus forecasting horizon.

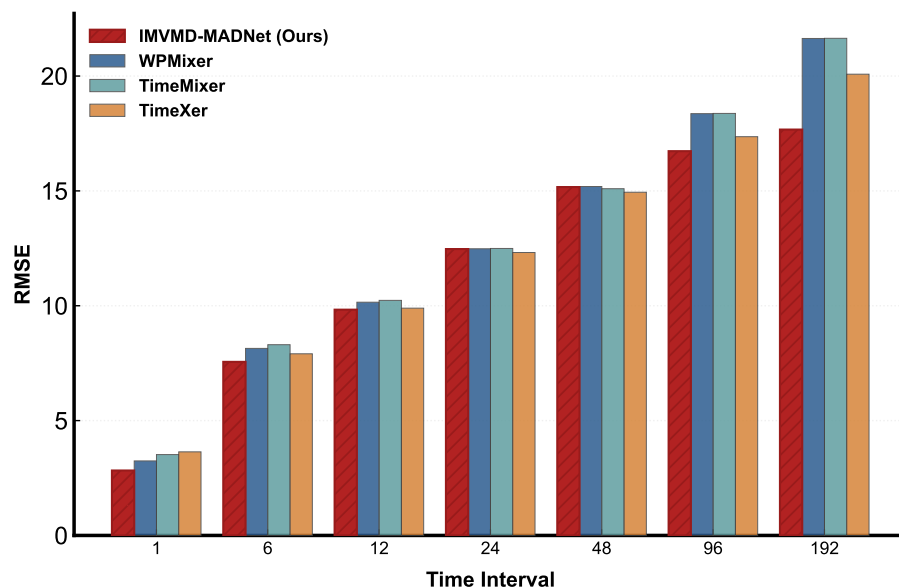


Figure 18. The trend of RMSE versus forecasting horizon.

When the forecast horizon further increases to 96 and 192 steps, the MAE of the baseline models exhibits an accelerated rise, indicating certain deficiencies in modeling the long-term evolutionary trends and complex operational state changes of chiller power. Comparatively, IMVMD-MADNet maintains a relatively gentle error growth trend, suppressing the gradual accumulation of errors to some extent during long-term forecasting,

thereby exhibiting more robust modeling capabilities for the medium- to low-frequency variation patterns and long-term dependencies of chiller power.

In summary, IMVMD-MADNet demonstrates good prediction stability across different forecasting horizons. Its relative advantage becomes increasingly evident as the forecast horizon extends, making it particularly suitable for medium- and long-term chiller power forecasting scenarios. This validates the model's potential value in practical engineering applications.

4.6. Experiment 6: Sensitivity Analysis

To evaluate the robustness of the proposed model under different seasonal operating conditions, the model was trained using data from January to June and tested on three independent periods with a forecasting horizon of 12 steps: August (summer, representing high-load conditions), October (transition season, representing medium-load conditions), and December (winter, representing low-load conditions). The corresponding results are presented in Tables 9–11.

Table 9. Evaluation results on summer (high-load) test samples.

Model	MAE (kW)	RMSE (kW)	SMAPE (%)	PeakMAE (kW)	RampMAE (kW)
IMVMD-MADNet (Ours)	7.081	10.179	1.833	8.981	12.410
WPMixer	7.203	10.344	1.867	9.027	13.071
TimeMixer	7.292	10.445	1.888	9.062	13.106
ARIMA	7.554	10.410	1.954	11.058	12.886
XGBoost	7.260	10.381	1.880	9.502	12.924

Table 10. Evaluation results on transition season (medium-load) test samples.

Model	MAE (kW)	RMSE (kW)	SMAPE (%)	PeakMAE (kW)	RampMAE (kW)
IMVMD-MADNet (Ours)	6.671	9.603	1.762	7.846	11.476
WPMixer	6.786	9.753	1.794	8.380	12.033
TimeMixer	6.853	9.831	1.810	8.166	11.925
ARIMA	6.968	9.740	1.843	10.137	11.614
XGBoost	6.720	9.682	1.777	8.192	12.088

Table 11. Evaluation results on winter (low-load) test samples.

Model	MAE (kW)	RMSE (kW)	SMAPE (%)	PeakMAE (kW)	RampMAE (kW)
IMVMD-MADNet (Ours)	6.817	10.009	1.782	8.066	12.210
WPMixer	6.994	10.262	1.830	8.414	12.773
TimeMixer	7.090	10.366	1.852	8.598	12.708
ARIMA	7.437	10.545	1.943	11.677	12.487
XGBoost	7.078	10.361	1.850	8.465	12.562

Based on Tables 9–11, our method (IMVMD-MADNet) consistently achieves the lowest errors across all five metrics in summer, transition, and winter conditions, demonstrating reliable performance across different load regimes.

In summer (high load), frequent chiller switching between multiple units creates complex load dynamics, and our method shows a notable advantage in PeakMAE (8.981), with improvements of 0.046–2.077 over the baselines, indicating better handling of extreme cooling demand events. In the transition season (medium load), the system typically operates with a single chiller under stable conditions, resulting in smoother load patterns, and all models achieve better results than in summer. Our method further reduces the MAE to 6.671 and PeakMAE to 7.846, with the largest advantage over ARIMA reaching 2.291 in PeakMAE. In winter (low load), the system experiences prolonged shutdown

periods and intermittent low-load operation, and our method still leads with the lowest MAE (6.817) and SMAPE (1.782), suggesting robust adaptation to reduced cooling loads. ARIMA consistently yields the highest PeakMAE (10.137–11.677) across all three seasons, confirming its limitation in capturing abrupt load changes regardless of load conditions.

Notably, the performance gap between our method and the baselines narrows in winter compared to high-load and medium-load conditions, suggesting that the more challenging load dynamics under high-load and transition conditions better differentiate model capability in chiller plant forecasting.

To quantitatively assess the impact of threshold selection on prediction performance, we conducted a sensitivity analysis on the validation set used in the Bayesian optimization hold-out strategy under a unified experimental setup (Lag S1 decomposition window, forecasting horizon = 12). The low-frequency threshold (low/medium boundary) was varied over 0.1, 0.15, 0.2, 0.25, 0.3, and the high-frequency threshold (medium/high boundary) was varied over 0.5, 0.6, 0.7, 0.8, 0.9. The results are shown in Table 12.

Table 12. Sensitivity analysis of sample entropy thresholds.

Low-Frequency Threshold	High-Frequency Threshold	MAE (kW)	RMSE (kW)	Relative MAE Increase vs. Optimal ¹	Relative RMSE Increase vs. Optimal ¹
0.2	0.5	6.623	10.017	0.861%	0.809%
0.2	0.6	6.582	9.961	0.243%	0.251%
0.2	0.7	6.566	9.936	Optimal	Optimal
0.2	0.8	6.594	9.952	0.425%	0.161%
0.2	0.9	6.645	10.038	1.189%	1.016%
0.1	0.7	—	—	—	—
0.15	0.7	6.587	9.953	0.319%	0.171%
0.25	0.7	6.602	9.974	0.545%	0.381%
0.3	0.7	6.658	10.049	1.382%	1.125%

¹ Note: Positive values in the last two columns indicate the percentage increase in prediction error relative to the optimal configuration (low = 0.2, high = 0.7).

As shown in Table 12, the combination (low = 0.1, high = 0.7) produced no valid results, because a threshold of 0.1 is too small to extract low-frequency components in certain windows, causing input format mismatch. Across all tested combinations, the maximum increase in MAE relative to the optimal configuration was only 1.38%, indicating that the proposed framework is not highly sensitive to the exact threshold values. The combination (low = 0.2, high = 0.7) achieved the best overall performance, further confirming the suitability of the literature-recommended thresholds for the chiller dataset used in this study.

To further verify whether the model's fitting capacity exceeds the effective data size, we conducted low-data sensitivity experiments by training IMVMD-MADNet on 10%, 25%, 50%, 70%, and 85% of the training data. The evaluation results on the validation set (forecasting horizon = 12) are shown in Table 13.

Table 13. Low-data sensitivity experiment results for IMVMD-MADNet.

Training Data Proportion (%)	MAE (kW)	RMSE (kW)	SMAPE (%)
10	7.455	10.577	2.032
25	7.038	10.093	1.916
50	6.975	10.066	1.902
70	6.803	9.854	1.855
85	6.807	9.881	1.857

The results show that even with only 10% of the training data, the model does not collapse or exhibit obvious overfitting. As the data proportion increases, errors generally

follow a decreasing trend, reaching the optimum at 70% training data. The overall trend follows a standard U-shaped curve, demonstrating that the model does not suffer from excessive fitting capacity. Even with strong sliding-window overlap that reduces the number of effective independent samples, the model learns generalizable load fluctuation patterns rather than memorizing redundant noise. These validation set results confirm that the proposed method avoids excessive fitting capacity.

4.7. Experiment 7: Computational Cost Analysis

To further evaluate the computational efficiency and deployment feasibility of the proposed IMVMD-MADNet, additional experiments were conducted to assess its computational complexity and inference performance. Specifically, the number of trainable parameters and the end-to-end inference latency under different lag settings were evaluated. The total number of trainable parameters in IMVMD-MADNet is 844,686. The inference latency under different lag settings is reported in Tables 14 and 15.

All measurements were conducted on a testbed with a 12th Gen Intel Core i7-12700H (2.30 GHz) CPU and an NVIDIA GeForce RTX 3060 GPU. The IMVMD decomposition and sample entropy reconstruction are executed on the CPU, while the MADNet backbone inference runs on the GPU.

As shown in Table 14, the IMVMD decomposition time increases with window length, from 2.635 ms (Lag S1) to 5.887 ms (Lag S3), approximately a $2.2\times$ increase, because the ADMM iteration complexity scales linearly with window length. The sample entropy and frequency-band reconstruction is the computational bottleneck: under Lag S3, it takes 52.237 ms, accounting for 88.26% of the total time, due to the $O(N^2)$ complexity of pairwise distance calculations. The MADNet backbone is very lightweight, requiring only 1.058 ms even under Lag S3 (1.79% of total time), confirming the high efficiency of the neural network component.

Table 14. Computational efficiency of IMVMD-MADNet and its sub-modules.

Forecast Horizon	Component	Batch Latency (ms/Batch)	Sample Latency (ms/Window)	Percentage (%)
Lag S1	IMVMD	84.320	2.635	33.998
	Sample entropy & reconstruction	141.344	4.417	56.990
	MADNet inference	22.350	0.698	9.012
	Total pipeline	248.014	7.750	100.000
Lag S2	IMVMD	119.008	3.719	21.586
	Sample entropy & reconstruction	406.240	12.695	73.684
	MADNet inference	26.077	0.815	4.730
	Total pipeline	551.325	17.229	100.000
Lag S3	IMVMD	188.384	5.887	9.947
	Sample entropy & reconstruction	1671.584	52.237	88.264
	MADNet inference	33.868	1.058	1.788
	Total pipeline	1893.836	59.182	100.000

Table 15 evaluates real-time deployment feasibility at a 1-minute (60,000 ms) sampling interval. We define the safety factor as the ratio of the sampling interval to the single-step inference time. Under all three window settings, the inference times are well within the sampling interval. Even under Lag S3, the safety factor is 1031.2, meaning the inference time occupies only 0.097% of the sampling budget. These results confirm that IMVMD-MADNet fully meets the real-time deployment requirements of building automation systems with a 1-minute sampling interval.

Table 15. Real-time deployment assessment of IMVMD-MADNet.

Metric	Lag S1	Lag S2	Lag S3
Sampling Interval (ms)	60,000	60,000	60,000
Sample Latency (ms/window)	7.750	17.229	59.182
Time Utilization (%)	0.013	0.029	0.099
Remaining Time (ms)	59,992.250	59,982.771	59,940.818
Safety Factor	7741.935	3482.500	1013.820

For real-time deployment, the end-to-end inference latency is the key concern. Table 16 compares the inference cost of our method with representative baseline models. An important distinction is that the additional latency in our pipeline comes from the parameter-free signal processing steps (IMVMD + sample entropy), not from extra neural network capacity. The MADNet backbone alone (0.698–1.058 ms) is within the same order of magnitude as the baseline models. The signal processing overhead is the cost of incorporating physically meaningful time-frequency decomposition, which we believe is justified by the consistent performance gains demonstrated across all evaluation scenarios.

Table 16. Single-window inference latency of IMVMD-MADNet and baseline models.

Model	MADNet-Only Latency (ms/Window)	Total Pipeline Latency (ms/Window)
IMVMD-MADNet (Ours)	0.698	7.750
WPMixer	0.153	0.153
TimeMixer	0.455	0.455
iTransformer	0.186	0.186

4.8. Experiment 8: Analysis of the Interpretability of Model Input Variables

To explore the physical interpretability and reasonableness of the input variables in the IMVMD-MADNet model for chiller power prediction, this study uses SHAP (SHapley Additive exPlanations) to quantitatively assess how each input variable contributes to the prediction outcomes under a single-step forecasting setting. SHAP values can characterize the influence of features on model outputs from both global and local perspectives, helping to reveal the consistency between the model's decision-making process and the actual operational mechanisms.

The input variables selected in this study include: chiller power P , cooling load Q , chilled-water outlet temperature T_{chw} , chilled-water flow rate V_{chw} , cooling-water inlet temperature T_{cwi} , outdoor dry-bulb temperature T_{db} , and outdoor wet-bulb temperature T_{wb} . A one-day continuous operation segment from the test set is extracted as the analysis sample. The SHAP analysis yields the results presented in Figure 19, and the corresponding average SHAP values for each input variable are 0.0095, 0.2639, 0.0335, 0.0440, 0.0083, 0.0303, and 0.3792, respectively.

As an endogenous variable, chiller power P attains the highest average SHAP value, indicating its most significant contribution to the model predictions. This aligns with the actual operating characteristics of chillers: due to factors such as compressor inertia, continuity of control strategies, and system thermal inertia, chiller power typically exhibits strong autocorrelation along the time series.

Among the exogenous variables, cooling load Q is the second most important feature, making it the most influential predictor after chiller power. Cooling load directly determines the chiller's cooling demand and compressor workload and is therefore strongly coupled with chiller power. Chilled-water flow rate V_{chw} ranks third, as its variation affects evaporator heat-exchange efficiency and system operating conditions, thereby significantly influencing chiller energy consumption. Chilled-water outlet temperature T_{chw} ranks

fourth, influencing the chiller's cooling performance and operating state. In comparison, outdoor wet-bulb temperature T_{wb} , cooling-water inlet temperature T_{cwi} , and outdoor dry-bulb temperature T_{db} exhibit relatively low average SHAP values. This is because, in short-term power prediction, environmental variables primarily influence chiller power indirectly or with temporal delays, making their contributions less significant than those of load- and operation-related variables.

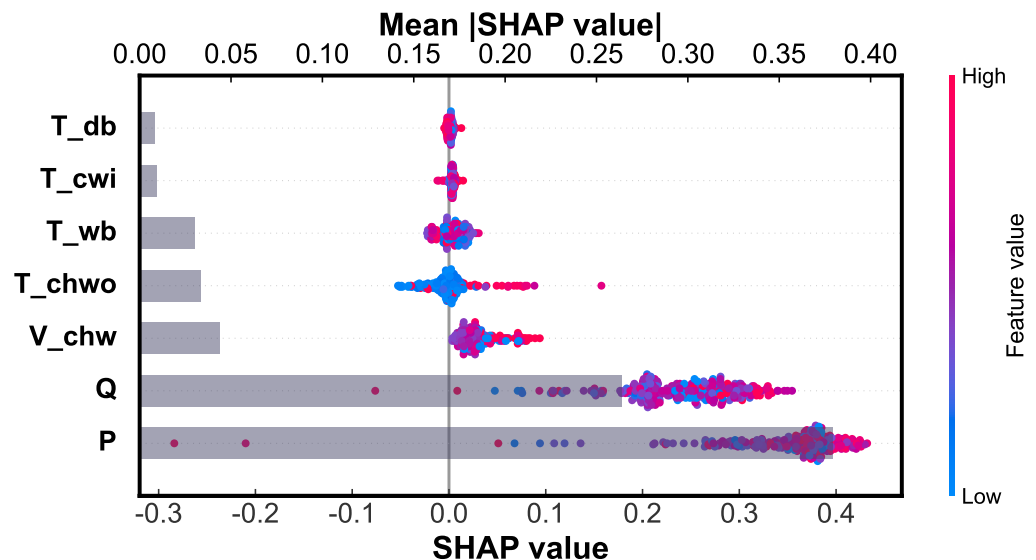


Figure 19. The results of the shap analysis.

In summary, IMVMD-MADNet reasonably allocates importance among different input variables during chiller power prediction: treating the endogenous power variable as the core driving factor, while fully utilizing information from cooling load and key operational parameters, and maintaining appropriate sensitivity to environmental variables. This feature-contribution distribution not only verifies the physical plausibility of the model's predictions but also demonstrates the model's effectiveness and reliability from an interpretability perspective.

5. Conclusions

A novel hybrid learning framework, IMVMD-MADNet, is proposed for short-term chiller energy consumption prediction. The framework integrates improved multivariate variational mode decomposition (IMVMD), sample entropy reconstruction, and the Multi-scale Aggregation Decomposition Network (MADNet). The framework addresses two key challenges in chiller energy forecasting: the information leakage introduced by global decomposition and the absence of an architecture capable of jointly modeling the mode, channel, and temporal dimensions. A rolling IMVMD module is employed to perform stepwise local decomposition, while envelope entropy is utilized for adaptive mode selection, thereby preventing the premature use of future information. A sample entropy-based reconstruction strategy reorganizes multivariate components into predefined frequency groups. The MADNet predictor adopts a dual-branch architecture that combines wavelet-domain time–frequency modeling (WDP) and time-domain causal dependency modeling (TDP). The two branches are coordinated by a cascaded Mode–Channel Aggregator (MCA). The framework is validated using one year of real industrial chiller operational data collected from a plant in Huizhou, Guangdong Province, China, and its performance is further benchmarked against 11 mainstream forecasting models. All experiments are conducted on the cleaned dataset under stable single-chiller or dual-chiller operating conditions. The proposed model is not intended to predict chiller start/stop

transients, as those segments were intentionally removed during preprocessing. The main findings are summarized below.

1. IMVMD-MADNet consistently outperforms all baseline models across six forecasting horizons (1–96 steps). It achieves an average MAE of 7.521 and an average RMSE of 10.824. Compared with WPMixer, TimeMixer, and iTransformer, MAE is reduced by 3.95%, 5.92%, and 5.12%, while RMSE is reduced by 4.03%, 5.46%, and 3.91%, respectively. The performance advantage becomes more pronounced at longer forecasting horizons, indicating improved robustness against error accumulation.
2. The rolling IMVMD module eliminates information leakage while adaptively determining the optimal number of modes ($K = 6$). The sample entropy reconstruction strategy generates components with clearer frequency separation. Ablation studies demonstrate the effectiveness of each module and their complementary contributions, with the complete IMVMD-MADNet achieving the best performance ($R^2 = 0.9330$, 0.9319 , and 0.9307 under three input sequence lengths).
3. The effectiveness of the dual-branch MADNet is attributed to the intrinsic phase coupling and thermal inertia of chiller systems. By jointly modeling these characteristics, the proposed architecture improves forecasting accuracy. These physical characteristics remain present under stable operating conditions, and the model captures them without relying on the removed start/stop transients.
4. Long-term forecasting experiments (up to 192 steps) show that IMVMD-MADNet maintains relatively slow error growth, whereas baseline models exhibit faster error accumulation. This confirms its ability to capture multiscale load variations under stable operating conditions. SHAP analysis further identifies chiller power and cooling load as the most influential features, demonstrating the reliability and physical plausibility of the proposed model.

Despite its promising performance, several limitations should be noted. First, the current evaluation is limited to one year of data collected from a single chiller plant, which may limit the generalizability of the proposed framework. Second, although data cleaning is necessary to improve data quality, it may yield more optimistic performance estimates than those achievable under real-world deployment with raw data. Third, only shared hyperparameters were optimized across models, and the rolling IMVMD decomposition and sample entropy computation incur additional computational overhead compared with simpler baseline methods.

Future work will focus on four directions: (i) validating the framework across different chiller types, climate zones, and building types to further assess its generalizability; (ii) improving the computational efficiency of rolling IMVMD decomposition and sample entropy computation through warm-start initialization and vectorized computation; (iii) incorporating K-fold Time Series Split cross-validation into Bayesian optimization for more robust hyperparameter selection; and (iv) establishing practical guidelines for recalibrating sample entropy thresholds when transferring the framework to new datasets.

Despite these limitations, the physically motivated multiscale decomposition architecture of IMVMD-MADNet shows strong potential for cross-domain generalization, as the extracted frequency-band characteristics are expected to be transferable across different chiller systems.

Author Contributions: Conceptualization, R.C. and X.W.; methodology, R.C.; software, R.C. and Y.L.; validation, R.C. and X.W.; formal analysis, R.C. and X.W.; investigation, R.C., X.W. and Y.L.; data curation, R.C. and Y.L.; writing—original draft preparation, R.C. and X.W.; writing—review and editing, R.C. and X.W.; visualization, R.C.; supervision, X.W.; project administration, X.W.; funding acquisition, R.C. and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be made available on request.

Acknowledgments: The authors are deeply obliged to the anonymous reviewers and editor for their insightful comments and constructive feedback.

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

References

1. Peng, Y.; Rysanek, A.; Nagy, Z.; Schlüter, A. Using machine learning techniques for occupancy-prediction-based cooling control in office buildings. *Appl. Energy* **2018**, *211*, 1343–1358.
2. Hong, T.; Wang, Z.; Luo, X.; Zhang, W. State-of-the-art on research and applications of machine learning in the building life cycle. *Energy Build.* **2020**, *212*, 109831.
3. Tang, P.; Yuan, Z.; Yu, Y.; Gao, B. Monotonicity-constrained physics-informed deep neural network for chiller energy consumption prediction: Extrapolation performance and constraint Improvement. *Energy Build.* **2026**, *359*, 117299.
4. Zhao, D.; Watari, D.; Ozawa, Y.; Taniguchi, I.; Suzuki, T.; Shimoda, Y.; Onoye, T. Data-driven online energy management framework for HVAC systems: An experimental study. *Appl. Energy* **2023**, *352*, 121921.
5. Sang, Y.; Zhao, J.R.; Sun, J.; Chen, B.; Liu, S. Experimental investigation and EnergyPlus-based model prediction of thermal behavior of building containing phase change material. *J. Build. Eng.* **2017**, *12*, 259–266.
6. Rashad, M.; Żabnieńska-Góra, A.; Norman, L.; Jouhara, H. Analysis of energy demand in a residential building using TRNSYS. *Energy* **2022**, *254*, 124357.
7. Rahmanparast, A.; Milani, M.; Camci, M.; Karakoyun, Y.; Dalkilic, A.S. Parameter reduction for PMV prediction via data driven approaches using the ASHRAE global thermal comfort database II and Chinese dataset. *Appl. Therm. Eng.* **2025**, *279*, 127553.
8. Ho, W.; Yu, F. Predicting chiller system performance using ARIMA-regression models. *J. Build. Eng.* **2021**, *33*, 101871.
9. Rizi, B.S.; Pavlak, G.; Cushing, V.; Heidarinejad, M. Predicting uncertainty of a chiller plant power consumption using quantile random forest: A commercial building case study. *Energy* **2023**, *283*, 129112.
10. Zhang, W.; Yu, J.; Zhao, A.; Zhou, X. Predictive model of cooling load for ice storage air-conditioning system by using GBDT. *Energy Rep.* **2021**, *7*, 1588–1597.
11. Wang, H.; Chen, Y.; Kang, J.; Ding, Z.; Zhu, H. An XGBoost-Based predictive control strategy for HVAC systems in providing day-ahead demand response. *Build. Environ.* **2023**, *238*, 110350.
12. Seo, B.; Yoon, Y.; Lee, K.H.; Cho, S. Comparative analysis of ANN and LSTM prediction accuracy and cooling energy savings through AHU-DAT control in an office building. *Buildings* **2023**, *13*, 1434.
13. Gao, J.; Ye, X.; Lei, X.; Huang, B.; Wang, X.; Wang, L. A multichannel-based CNN and GRU method for short-term wind power prediction. *Electronics* **2023**, *12*, 4479.
14. Yan, K.; Zhou, X. Chiller faults detection and diagnosis with sensor network and adaptive 1D CNN. *Digit. Commun. Netw.* **2022**, *8*, 531–539.
15. Lin, J.; Lin, W.; Lin, W.; Liu, T.; Wang, J.; Jiang, H. Multi-objective cooling control optimization for air-liquid cooled data centers using TCN-BiGRU-Attention-based thermal prediction models. In *Proceedings of the Building Simulation*; Springer: Berlin/Heidelberg, Germany, 2024; Volume 17, pp. 2145–2161.
16. Li, L.; Su, X.; Bi, X.; Lu, Y.; Sun, X. A novel Transformer-based network forecasting method for building cooling loads. *Energy Build.* **2023**, *296*, 113409.
17. Lv, L.; Wu, Z.; Zhang, J.; Zhang, L.; Tan, Z.; Tian, Z. A VMD and LSTM based hybrid model of load forecasting for power grid security. *IEEE Trans. Ind. Inform.* **2021**, *18*, 6474–6482.
18. Wang, Y.; Cheng, H.; Chen, H.; Ye, M.; Ren, Y.; Yang, C. A hybrid model based on wavelet decomposition and LSTM for short-term energy consumption prediction of chillers. *J. Build. Eng.* **2025**, *99*, 111539.
19. Xie, B.; Shi, S.; Liu, W. Integrated forecasting of marine renewable power: An adaptively Bayesian-optimized MVMD-LSTM framework for wind-solar-wave energy. *Sustain. Energy Technol. Assessments* **2025**, *84*, 104753.
20. Wang, S.; Shi, J.; Yang, W.; Yin, Q. High and low frequency wind power prediction based on Transformer and BiGRU-Attention. *Energy* **2024**, *288*, 129753.
21. Li, K.; Huang, W.; Hu, G.; Li, J. Ultra-short term power load forecasting based on CEEMDAN-SE and LSTM neural network. *Energy Build.* **2023**, *279*, 112666.

22. Cui, M.; Hou, E.; Hou, P. VMD-SE-CEEMDAN-BO-CNNGRU: A Dual-Stage Mode Decomposition Hybrid Deep Learning Model for Microseismic Time Series Prediction. *Mathematics* **2025**, *13*, 2121.
23. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv* **2022**, arXiv:2210.02186.
24. Liu, S.; Yu, H.; Liao, C.; Li, J.; Lin, W.; Liu, A.X.; Dustdar, S. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–8 May 2021.
25. Wang, S.; Wu, H.; Shi, X.; Hu, T.; Luo, H.; Ma, L.; Zhang, J.; Zhou, J. Timemixer: Decomposable multiscale mixing for time series forecasting. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024; Volume 2024, pp. 38626–38652.
26. Murad, M.M.N.; Aktukmak, M.; Yilmaz, Y. Wpmixer: Efficient multi-resolution mixing for long-term time series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 19581–19588.
27. Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; Jin, R. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; PMLR; pp. 27268–27286.
28. Hua, J.; Zou, J.; Rao, J.; Yin, H.; Chen, J. ECG signals deep compressive sensing framework based on multiscale feature fusion and SE block. *IEEE Access* **2023**, *11*, 104359–104372.
29. Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; Long, M. itransformer: Inverted transformers are effective for time series forecasting. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024; Volume 2024, pp. 11116–11140.
30. Karijadi, I.; Chou, S.Y. A hybrid RF-LSTM based on CEEMDAN for improving the accuracy of building energy consumption prediction. *Energy Build.* **2022**, *259*, 111908.
31. Huang, X.; Han, Y.; Yan, J.; Zhou, X. Hybrid forecasting model of building cooling load based on EMD-LSTM-Markov algorithm. *Energy Build.* **2024**, *321*, 114670.
32. Ur Rehman, N.; Aftab, H. Multivariate variational mode decomposition. *IEEE Trans. Signal Process.* **2019**, *67*, 6039–6052.
33. Sun, X.; Liu, H. Multivariate short-term wind speed prediction based on PSO-VMD-SE-ICEEMDAN two-stage decomposition and Att-S2S. *Energy* **2024**, *305*, 132228.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
35. Zhang, G.; Choi, D.; Jung, J. Development of continuous cuffless blood pressure prediction platform using enhanced 1-D SENet-LSTM. *Expert Syst. Appl.* **2024**, *242*, 122812.
36. Yang, S.; Liu, Y.; Liao, Y.; Su, K. A new method of bearing remaining useful life based on life evolution and SE-ConvLSTM neural network. *Machines* **2022**, *10*, 639.
37. Chen, Y.; Zhang, S.; Tang, X.; LIU, H. Real-world vehicle charging duration prediction based on transfer learning with SENet-CNN-Transformer Model. *Green Energy Intell. Transp.* **2026**, *5*, 100372.
38. Li, D.; Ma, J.; Rao, K.; Wang, X.; Li, R.; Yang, Y.; Zheng, H. Prediction of rainfall time series using the hybrid DWT-SVR-Prophet model. *Water* **2023**, *15*, 1935.
39. Hu, Y.; Liu, P.; Zhu, P.; Cheng, D.; Dai, T. Adaptive multi-scale decomposition framework for time series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 17359–17367.
40. Abumohsen, M.; Owda, A.Y.; Owda, M. Electrical load forecasting using LSTM, GRU, and RNN algorithms. *Energies* **2023**, *16*, 2283.
41. Bai, J.; Zhu, W.; Liu, S.; Ye, C.; Zheng, P.; Wang, X. A temporal convolutional network–bidirectional long short-term memory (TCN-BiLSTM) prediction model for temporal faults in industrial equipment. *Appl. Sci.* **2025**, *15*, 1702.
42. Chen, J.; Peng, T.; Qian, S.; Ge, Y.; Wang, Z.; Nazir, M.S.; Zhang, C. An error-corrected deep Autoformer model via Bayesian optimization algorithm and secondary decomposition for photovoltaic power prediction. *Appl. Energy* **2025**, *377*, 124738.
43. Xu, A.; Chen, J.; Li, J.; Chen, Z.; Xu, S.; Nie, Y. Multivariate rolling decomposition hybrid learning paradigm for power load forecasting. *Renew. Sustain. Energy Rev.* **2025**, *212*, 115375.
44. Tan, J.; Zhu, H.; Zhang, J.; Liu, H. Multi-stage wind speed prediction with CEEMDAN-SE-IDBO-LSTM based on rolling decomposition. *Energy* **2025**, *338*, 138741.
45. Huang, X.; Zhou, X.; Yan, J.; Huang, X. Cooling load forecasting method for central air conditioning systems in manufacturing plants based on iTransformer-BiLSTM. *Appl. Sci.* **2025**, *15*, 5214.
46. Wu, S.; Cai, H. Short-term power load prediction of VMD-LSTM based on ISSA optimization. *Appl. Sci.* **2025**, *15*, 5037.
47. Huang, X.; Tang, J.; Shen, Y. Long time series of ocean wave prediction based on PatchTST model. *Ocean Eng.* **2024**, *301*, 117572.

48. Wang, Y.; Wu, H.; Dong, J.; Qin, G.; Zhang, H.; Liu, Y.; Qiu, Y.; Wang, J.; Long, M. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 469–498.
49. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are transformers effective for time series forecasting? In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 11121–11128.
50. Ekambaram, V.; Jati, A.; Nguyen, N.; Sinthong, P.; Kalagnanam, J. Tsmixer: Lightweight mlp-mixer model for multivariate time series forecasting. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 6–10 August 2023; pp. 459–469.
51. Campos, D.; Zhang, M.; Yang, B.; Kieu, T.; Guo, C.; Jensen, C.S. Lightts: Lightweight time series classification with adaptive ensemble distillation. *Proc. Acm Manag. Data* **2023**, *1*, 1–27.
52. Liu, S.; Xu, T.; Du, X.; Zhang, Y.; Wu, J. A hybrid deep learning model based on parallel architecture TCN-LSTM with Savitzky-Golay filter for wind power prediction. *Energy Convers. Manag.* **2024**, *302*, 118122.

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