energies

MDPI

Article

Short-Term Power Load Forecasting Based on Interpretable Time-Frequency Dual-Domain Fusion

Hengyu Xu [1] and Fengyong Li [1,2,*]

[1] Faculty of Artificial Intelligence, Shanghai University of Electric Power, Shanghai 201306, China; hengyuxu01@gmail.com
[2] Engineering Research Center of Offshore Wind Technology Ministry of Education (Shanghai University of Electric Power), Yangpu District, Shanghai 200090, China
* Correspondence: fyli@shiep.edu.cn

Abstract

Real-world power load series exhibit pronounced non-stationarity and complex spatiotemporal dependencies, while the black-box nature of deep learning limits its credibility in industrial dispatching. This paper proposes a probabilistic short-term load forecasting model integrating multi-scale temporal modeling and learnable frequency-domain filtering to improve forecasting accuracy, uncertainty quantification, and interpretability. The model employs parallel temporal and frequency branches. The temporal branch uses multi-scale patches to capture local fluctuations and long-range dependencies, while the frequency branch applies a learnable complex-valued filter to extract dominant global seasonal periodicities. A dynamic gating mechanism adaptively fuses features from both domains, and quantile regression generates prediction intervals. On the Quanzhou and Australian datasets, the proposed model achieves mean absolute percentage errors of 0.49% and 0.77%, respectively. The corresponding prediction interval coverage probabilities are 79.18% and 80.65%, both close to the nominal coverage of 80%. A SHapley Additive exPlanations (SHAP)-based post hoc attribution framework further identifies key factors influencing predictions through feature contributions. The results suggest that the proposed model effectively combines local temporal fluctuations with global frequency-domain periodic information and provides a degree of post hoc interpretability, supporting accurate probabilistic load forecasting and a clearer understanding of the model's predictions.

Keywords: short-term power load forecasting; time-frequency fusion; transformer; quantile regression; SHAP explainability

Check for updates

Academic Editor: Marcin Kaminski

Received: 28 September 2026
Revised: 22 October 2026
Accepted: 24 October 2026
Published: 30 October 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) license.

1. Introduction

The transition toward carbon peaking and carbon neutrality, together with the development of new power systems, is increasing the complexity of electricity supply and demand [1,2]. Large-scale renewable energy integration and the growing use of interactive electrical devices have made load profiles more variable [3]. Meanwhile, the combined effects of weather conditions, calendar factors, and stochastic user behavior give rise to pronounced non-stationarity, multiple periodic patterns, and nonlinear load fluctuations [4]. Accurate load forecasting is therefore essential for secure grid operation and efficient dispatch, while complex and changing operating conditions place greater demands on forecasting methods [5].

Traditional statistical and shallow machine learning methods, including the autoregressive integrated moving average (ARIMA) model [6] and support vector machine (SVM) [7],

offer relatively simple structures and fast model fitting but have difficulty representing highly nonlinear and volatile load patterns. Deep learning methods, such as long short-term memory (LSTM) networks [8] and temporal convolutional networks (TCNs) [9], have improved the modeling of temporal dependencies. Nevertheless, limitations associated with local receptive fields and the retention of long-range information can restrict their ability to represent non-stationary load dynamics across multiple time scales.

Through matrix-based self-attention, Transformer models process all time steps of the input sequence in parallel during training, enabling efficient processing of large volumes of historical observations [10]. Reference [11] proposed a load forecasting method based on multi-frequency sequence feature analysis and multi-point modified FEDformer, which enhances the model's ability to capture complex periodic patterns. Reference [12] developed a short-term load forecasting model based on the residual fusion of Autoformer and Transformer. By dynamically integrating sequence decomposition features and global attention information through an adaptive mechanism, this model improves the modeling of complex load fluctuations. To overcome the drawback that point-wise attention may disrupt local semantics, Reference [13] proposed a secondary decoupling reconstruction load forecasting method based on the PatchTST architecture. Through patching, this method preserves the local semantics of load sequences while improving the efficiency of dynamic load feature extraction. Reference [14] proposed a load forecasting method based on modal decomposition and an improved iTransformer. By decoupling load sequences and employing an inverted-dimension architecture, this method improves the representation of global temporal features.

Multi-domain feature fusion has attracted increasing attention as a way to address the limitations of modeling from a single perspective. Reference [15] proposed an end-to-end medium-term load forecasting framework based on multi-channel techniques and time-frequency domain fusion. By simultaneously extracting temporal and frequency-domain features and employing a multi-channel collaborative mechanism, this framework jointly models the dynamic evolution patterns of sequences and learns multi-scale feature representations in the time-frequency space. Reference [16] proposed a multi-scale time-frequency complementary load forecasting method for integrated energy systems. By using hierarchical downsampling and seasonal-trend decoupling techniques, this method separately extracts local non-stationary fluctuations in the time domain and global periodic components in the frequency domain, thereby improving the integration of complementary multi-scale information across the two domains.

Although the aforementioned multi-domain fusion models have made significant progress in capturing complex temporal patterns, further improvements are still needed in the adaptive representation and interaction of local non-stationary fluctuations and global periodic information. First, in terms of feature extraction and interaction, some existing methods rely on predefined signal decomposition or fixed feature aggregation. Although these strategies can effectively incorporate time-frequency information, they do not explicitly combine learned spectral representations with input-dependent routing between temporal and frequency-domain features within a unified framework. From a physical perspective, power load sequences inherently involve both local abrupt variations and macroscopic periodic patterns [17]. Therefore, reducing reliance on manually specified assumptions and enabling input-dependent interaction between temporal and frequency-domain features are crucial for improving multi-scale forecasting accuracy. Second, from the perspective of model credibility and operational security, existing multi-domain fusion studies still face the dual challenge of quantifying extreme fluctuation risks and interpreting model predictions. On the one hand, most existing models mainly focus on deterministic point forecasting and provide limited information about forecasting uncertainty, making it difficult

to assess the risks associated with load fluctuations. On the other hand, the increasing complexity of time-frequency models makes it difficult to interpret the contributions of input features to model predictions. Therefore, uncertainty quantification and complementary interpretability analyses are needed to improve the interpretability and practical credibility of load forecasting models [18].

To address the above issues, this paper proposes an end-to-end probabilistic load forecasting model based on time-frequency dual-domain fusion. The model employs a multi-scale temporal branch and a learnable frequency-domain filtering branch in parallel to capture local fluctuations and global periodic patterns, respectively. Unlike the decomposition-based multi-channel framework in [15], the proposed model learns these representations directly from the input sequences without a separate VMD stage, thereby simplifying the preprocessing pipeline and avoiding additional decomposition parameter selection. The resulting temporal and spectral representations are then fused through a dynamic gating mechanism, which uses their joint context to generate feature-wise fusion weights for each input window. Compared with the cascaded multi-scale mixing and frequency-domain interpolation framework in [16], the proposed design applies temporal encoding and spectral filtering to the same normalized input in parallel, then uses both outputs to determine their relative contributions to the fused representation. This provides a flexible way to combine local fluctuations and periodic information under different load conditions. Finally, quantile regression is incorporated to generate prediction intervals, while SHAP-based post hoc analysis explains the model predictions in terms of input feature contributions. The main contributions of this paper are summarized as follows:

(1) A parallel time-frequency dual-domain framework is developed for short-term probabilistic load forecasting. The multi-scale patch-based temporal branch captures local fluctuations and long-range dependencies at different temporal resolutions, while the frequency-domain branch employs a learnable complex-valued filter to refine global periodic information. The architectural contribution lies in jointly learning multi-scale temporal and filtered spectral representations through parallel paths from the same normalized historical input, without requiring a separate decomposition stage. This makes spectral refinement part of the trainable feature-extraction process alongside multi-scale temporal encoding.

(2) An input-dependent gated fusion mechanism coordinates the temporal and spectral branches at the feature level. For each input window, the concatenated representations jointly determine a gating vector whose elements control the relative contributions of the two domains to the corresponding fused features. This design allows the temporal–spectral balance to vary across features and input windows, providing a flexible alternative to fixed-weight aggregation. Ablation results and the learned gating-value distributions support the effectiveness and input-dependent behavior of the fusion design.

(3) A unified probabilistic forecasting and post hoc interpretation framework is developed on the proposed dual-domain architecture. The fused time-frequency representation supports the joint estimation of median forecasts and interval bounds through multi-quantile regression, incorporating uncertainty quantification directly into the forecasting network. SHAP complements these probabilistic outputs by quantifying the contributions of historical input features to the central forecast, providing a feature-level explanation of the information underlying the model's prediction.

2. Problem Definition and Theoretical Foundation

Power load forecasting aims to extract the inherent patterns and periodic characteristics of load sequences from historical observations, thereby providing an estimate of future load

variations. In practical and complex dispatching scenarios of new power systems, dispatch centers require not only accurate reference values for short-term operational scheduling but also reasonable reserve capacity to cope with extreme load fluctuations. Therefore, this paper aims to construct an integrated forecasting framework that simultaneously considers high-accuracy deterministic point forecasting and reliable probabilistic interval forecasting. Given a multivariate time series containing D feature variables, at the current time step t, the data within a historical look-back window of length T are defined as the input feature matrix X, which is expressed as:

$$X = [x_{t-T+1}, x_{t-T+2}, \ldots, x_t]^\top \tag{1}$$

where x_i denotes the observed feature vector at the i-th historical time step. A traditional deterministic forecasting model learns a deterministic mapping and outputs a single point forecast sequence $\hat{Y}$ for the future L time steps. However, in a highly non-stationary power grid environment, a single point forecast provides limited warning information about extreme-value risks. To overcome this limitation, this paper designs a deep time-frequency fusion network f_θ parameterized by θ. Given a target quantile set $Q = \{q_1, q_2, \ldots, q_k\}$, the proposed model can simultaneously output, in parallel, a set of predicted sequences $\hat{Y}$ for the future L steps at different quantile levels:

$$\hat{Y} = f_\theta(X) = \{\hat{Y}^{q_i} \mid q_i \in Q\} \tag{2}$$

Specifically, for the future time step $t+h$, where $h \in \{1, 2, \ldots, L\}$, and a specific quantile level q, the predicted value $\hat{y}^q_{t+h}$ output by the model represents the q-th quantile of the probability distribution of the future actual load y_{t+h}. Under ideal calibration, the probability that the actual load value is less than or equal to the predicted value is equal to q:

$$P(y_{t+h} \le \hat{y}^q_{t+h} \mid X) = q \tag{3}$$

Based on the above multi-quantile outputs, the proposed model can support both point forecasting and interval forecasting. Taking the target quantile set $Q = \{0.1, 0.5, 0.9\}$ as an example, the median prediction sequence $\hat{Y}^{0.5}$ output by the model, namely $q = 0.5$, represents the most probable future load trajectory under the criterion of minimizing the expected absolute error. It can be directly used as the baseline load curve for conventional short-term operational scheduling and economic dispatch. The combination of the upper and lower quantile prediction sequences $[\hat{Y}^{0.1}, \hat{Y}^{0.9}]$ constructs a prediction interval with a confidence level of 80%.

2.1. Quantile Regression

Traditional point forecasting cannot quantify the uncertainty of load fluctuations. Quantile regression does not require strict assumptions about the residual distribution and can robustly estimate conditional quantiles at different probability levels.

For a target quantile $q \in (0, 1)$, the model uses the asymmetric Pinball Loss function to optimize the network parameters. Given the actual load y and the predicted value $\hat{y}^q$, the single-sample loss is defined as:

$$L_q(y, \hat{y}^q) = \begin{cases} q(y - \hat{y}^q), & y \ge \hat{y}^q \\ (1-q)(\hat{y}^q - y), & y < \hat{y}^q \end{cases} \tag{4}$$

Pinball Loss imposes asymmetric penalties on overestimation and underestimation through the weights q and $1-q$, thereby driving the model to output specific probabilistic boundaries. In the proposed multi-quantile architecture, for the target quantile set

$Q = \{q_1, \ldots, q_k\}$ and the prediction horizon L, the total objective function L_{total} is defined as the average loss over all forecasting steps and target quantile levels:

$$L_{\text{total}} = \frac{1}{L \times |Q|} \sum_{h=1}^{L} \sum_{q \in Q} L_q\left(y_{t+h}, \hat{y}_{t+h}^{q}\right) \tag{5}$$

By minimizing L_{total}, the model can simultaneously output the median point forecast ($q = 0.5$) and the upper and lower bounds of the probabilistic prediction interval for quantifying extreme-value risks.

2.2. SHAP-Based Post Hoc Explainability Theory

Deep learning models are often regarded as difficult-to-interpret "black boxes" in load forecasting. To quantify input feature contributions to the predictions, this study employs the SHAP (Shapley Additive Explanations) post hoc attribution framework.

SHAP treats the prediction process as a cooperative game among features, thereby quantifying the marginal contribution of each feature to the final output. For a feature set F, the Shapley value ϕ_i of the i-th feature is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [v(S \cup \{i\}) - v(S)] \tag{6}$$

where S denotes any feature subset that does not contain feature i, and $v(S)$ represents the expected model output when only the subset S is used. SHAP satisfies the important property of local accuracy, meaning that the sum of the attribution values of all features is strictly equal to the difference between the prediction value $f(x)$ of the current sample and the baseline expected output ϕ_0:

$$f(x) = \phi_0 + \sum_{i=1}^{|F|} \phi_i \tag{7}$$

For the dual-branch network proposed in this paper, a gradient-based explainer is adopted to perform approximate computation through backpropagation. Through this feature attribution mechanism, SHAP provides a solid quantitative basis for interpreting the model's predictions.

3. Time-Frequency Dual-Domain Load Forecasting Model

3.1. Overall Architecture Design

Unlike methods that rely on time-domain modeling alone or predefined static decomposition, the proposed model jointly models local fluctuations in the time domain and periodic structures in the frequency domain within an end-to-end framework and realizes adaptive fusion of dual-domain information through dynamic gating. As shown in Figure 1, the overall workflow of the model comprises the following four stages:

(1) Input and preprocessing stage: Given a multivariate time series matrix X containing historical load and meteorological covariates, the model first passes it through the reversible instance normalization (RevIN) module. This module independently calculates and caches the mean μ and standard deviation σ of each feature along the temporal dimension:

$$\mu = \frac{1}{T} \sum_{t=1}^{T} x_t, \quad \sigma = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (x_t - \mu)^2 + \epsilon} \tag{8}$$

(2) Parallel time-frequency dual-branch extraction stage: The normalized sequence is fed simultaneously into parallel temporal-domain and frequency-domain feature extraction branches to extract complementary representations. The temporal branch focuses on capturing local nonlinear abrupt changes and long-range contextual dependencies in the load sequence, while the frequency branch is dedicated to refining global periodic patterns and suppressing high-frequency random noise. Together, the two branches characterize complementary aspects of load dynamics.

(3) Dynamic gated fusion stage: An adaptive gating mechanism is introduced to perform nonlinear fusion of dual-domain features. This mechanism dynamically calculates fusion weights according to the time-frequency context of the input sequence and adaptively adjusts the model's attention to temporal local details or frequency-domain global periodic patterns. This produces a fused representation that combines local temporal information with global periodic features.

(4) Denormalized probabilistic forecasting stage: The fused feature vector is fed into parallel quantile regression heads. The model simultaneously computes the prediction results under the target quantile set, such as $q \in \{0.1, 0.5, 0.9\}$. Finally, the statistical mean μ and standard deviation σ cached by the RevIN module in the first stage are used to perform inverse scaling and shifting on the prediction results:

$$\hat{y}^q_{t+h} = \frac{\hat{y}^q_{t+h} - \beta}{\gamma} \odot \sigma + \mu \tag{9}$$

Thus, the load interval forecasting sequence $\hat{Y}$ is restored to its original physical scale.

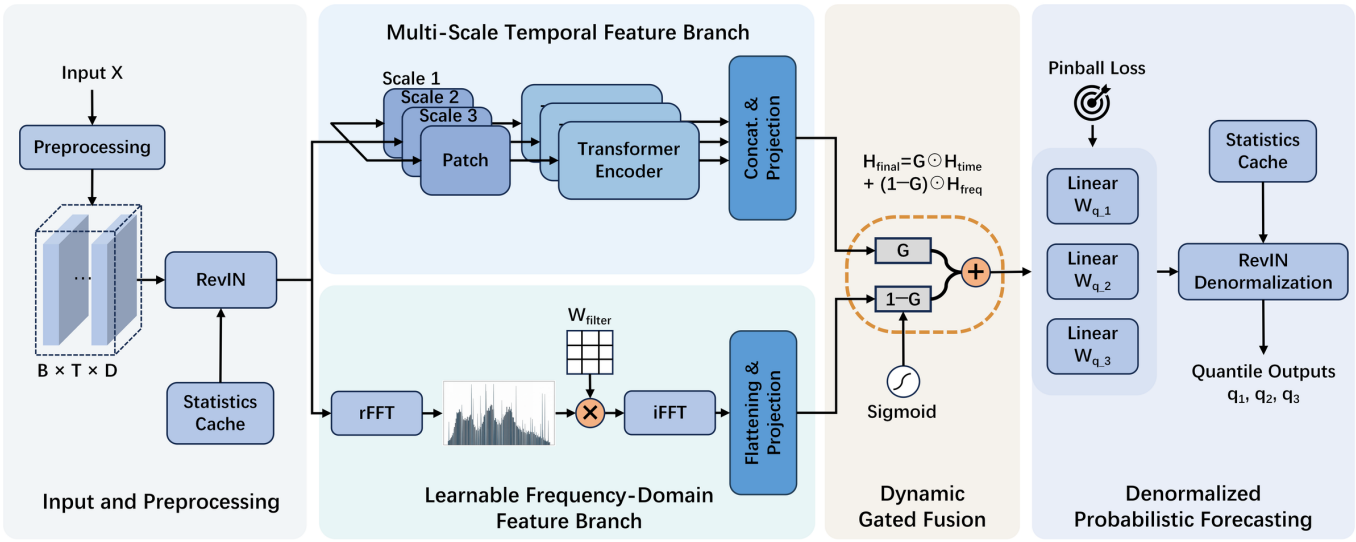


Figure 1. Overall architecture of the proposed model.

3.2. Multi-Scale Temporal Feature Branch

The traditional point-wise attention mechanism is computationally demanding and may disrupt the local semantics of time series. To address these limitations, this branch introduces multi-scale patching and the Transformer architecture to efficiently capture local nonlinear abrupt changes and long-range contextual dependencies in load sequences. As illustrated in Figure 2, given the original input sequence X, the model adopts multiple patch lengths in parallel, such as $P \in \{4, 8, 12\}$, to divide the sequence into non-overlapping local segments. After linear projection and positional encoding, these segments are mapped into high-dimensional token sequences. Then, the multi-head self-attention mechanism of the Transformer encoder is used to conduct deep feature interaction within a global receptive

field, producing the temporal representation $Z^{(P)}$. Subsequently, the final token is extracted from the output sequence at each scale:

$$z_{\text{last}}^{(P)} = Z^{(P)}[N, :] \tag{10}$$

After self-attention processing, this token summarizes the global historical context within the look-back window. Finally, the tail tokens extracted from all scales are concatenated along the feature dimension and compressed through the multi-scale projection layer W_{ms}, yielding the final temporally aggregated feature:

$$H_{\text{time}} = W_{\text{ms}}\text{Concat}\left(z_{\text{last}}^{(P_1)}, \ldots, z_{\text{last}}^{(P_m)}\right) \tag{11}$$

The feature H_{time} can characterize load dynamics and evolution trends from different scale perspectives.

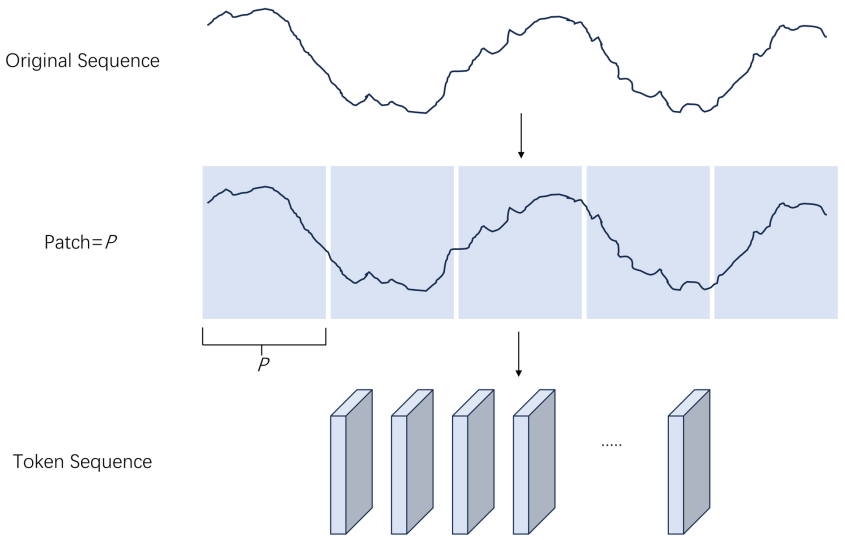


Figure 2. Schematic diagram of sequence segmentation with Patch = P.

3.3. Learnable Frequency-Domain Feature Extraction Branch

Power load sequences exhibit strong periodic patterns driven by human daily routines. To capture these global periodic patterns, this branch introduces a learnable frequency-domain filtering mechanism based on the discrete Fourier transform. The module first applies the real fast Fourier transform (rFFT) to the normalized sequence. By exploiting the conjugate symmetry of real-valued signals, only the first $\lfloor T/2 \rfloor + 1$ independent valid frequency components are retained to obtain the frequency-domain representation F, thereby reducing redundant computational overhead. Subsequently, a learnable complex-valued filter W_{filter} is introduced. After expressing F as a real-valued tensor containing its real and imaginary components, the model performs element-wise multiplicative modulation:

$$F = F \odot W_{\text{filter}} \tag{12}$$

This operation uses learned spectral weights to refine dominant periodic components and suppress high-frequency noise. The filter W_{filter} is optimized jointly with the forecasting network during training and shared across input samples. Its parameters remain fixed during inference, although the filtered representation varies with the input spectrum. Finally, the inverse fast Fourier transform (iFFT) is applied to restore the representation to the time-domain space, and a linear projection produces the global periodic feature representation H_{freq}. This global periodic representation complements the local abrupt-change feature H_{time} extracted by the multi-scale temporal module. Input-dependent fusion is implemented by the subsequent gating module in Section 3.4, which generates fusion weights for each input window.

3.4. Dynamic Gated Fusion Mechanism

To account for variations in the relative importance of global periodicity and local abrupt changes, the model uses a dynamic gated fusion mechanism based on sigmoid activation to adaptively balance time-frequency dual-domain features. Given the multi-scale temporal feature H_{time} and the frequency-domain periodic feature H_{freq}, the model first concatenates them along the feature dimension to construct a joint representation $H_{\text{concat}} = \left[H_{\text{time}}; H_{\text{freq}}\right]$. Then, the joint representation is passed through a single-layer linear transformation and the sigmoid activation function $\sigma(\cdot)$ to generate a dynamic gating vector G, whose elements lie in $(0,1)$:

$$G = \sigma\left(W_{\text{gate}} H_{\text{concat}} + b_{\text{gate}}\right) \tag{13}$$

where W_{gate} and b_{gate} are learnable parameters of the gating network. Finally, the weight G is used to perform element-wise weighted combination of the dual-domain features, producing the final fused representation H_{final}:

$$H_{\text{final}} = G \odot H_{\text{time}} + (1 - G) \odot H_{\text{freq}} \tag{14}$$

Through this soft selection mechanism, the model realizes implicit feature routing and enhances its robustness in complex non-stationary environments.

3.5. Multi-Quantile Prediction Head and Objective Function

To support both point forecasting and interval forecasting, parallel multi-quantile regression heads are applied to the fused representation H_{final}. For the predefined target quantile set $Q = \{q_1, \ldots, q_k\}$, such as $Q = \{0.1, 0.5, 0.9\}$, the model uses a separate linear mapping for each quantile to produce the prediction sequence $\hat{Y}^q$ for the future L steps on the normalized scale:

$$\hat{Y}^q = W_q H_{\text{final}} + b_q, \quad \forall q \in Q \tag{15}$$

where W_q and b_q are the weight and bias corresponding to quantile q, respectively. Then, using the local mean μ, standard deviation σ, and learnable affine parameters γ and β cached during the forward pass, $\hat{Y}^q$ is transformed back to the load forecasting sequence in real physical units:

$$\hat{Y}^q = \frac{\hat{Y}^q - \beta}{\gamma} \odot \sigma + \mu \tag{16}$$

The model is trained end to end by minimizing the joint pinball loss L:

$$L = \frac{1}{|Q| \cdot L} \sum_{q \in Q} \sum_{h=1}^{L} \max\left(q \cdot e_{t+h}^q, (q-1) \cdot e_{t+h}^q\right) \tag{17}$$

where $e_{t+h}^q = y_{t+h} - \hat{y}_{t+h}^q$ denotes the difference between the actual load and the quantile prediction at time step $t+h$. Guided by this asymmetric loss, the model can simultaneously achieve median point forecasting and interval boundary forecasting.

4. Experimental Results and Analysis

To evaluate the effectiveness of the proposed architecture across different geographical and meteorological contexts, two real-world load datasets from Quanzhou, Fujian, China, and Australia are used. The Quanzhou and Australian datasets are provided in Supplementary Materials S1. Both datasets contain regionally aggregated electricity load measurements. Their main characteristics are summarized in Table 1.

Table 1. Main characteristics of the two load datasets.

Characteristic	Quanzhou Dataset	Australian Dataset
Observation period	1 January 2016 to 7 November 2018	1 January 2006 to 1 January 2011
Sampling interval	15 min	30 min
Number of valid observations	100,000	87,648
Load range (MW)	1306–12,296	5498–14,274
Mean load (MW)	6924	8894
Input feature groups	Historical load, weather, and calendar features	Historical load, weather, electricity price, and calendar features
Input dimensionality	15	15

For Quanzhou, the six original input variables are historical load, maximum, minimum, and mean temperatures, humidity, and rainfall. For Australia, they are historical load, dry-bulb, dew-point, and wet-bulb temperatures, humidity, and electricity price. Both datasets additionally include a holiday indicator, a working-day indicator, and seven day-of-week indicators, resulting in 15 input dimensions. Electricity load is the prediction target for both datasets.

For both datasets, a sliding-window mechanism is used to construct sequential samples, with a historical observation window of 96 time steps and a prediction horizon of 4 time steps. Given the 15-min sampling interval of the Quanzhou dataset, these settings correspond to a 24-h historical window and forecasts for the next hour at 15-min intervals. For the Australian dataset, which has a 30-min sampling interval, the same settings correspond to a 48-h historical window and forecasts for the next 2 h at 30-min intervals. Thus, the two datasets share the same input and output sequence lengths but differ in their physical time spans.

During preprocessing, timestamps are first parsed to generate holiday indicators, working-day indicators, and one-hot-encoded day-of-week features. These calendar features are combined with historical load and meteorological variables, together with electricity price for the Australian dataset, to form a 15-dimensional input vector for each observation. Each time series is first divided chronologically, with the first 80% allocated to an initial training portion and the remaining 20% reserved for testing. The initial training portion is further divided chronologically at an 80%/20% ratio for model training and validation, resulting in an overall training–validation–testing split of 64%/16%/20%. The input variables and the load target are then standardized using parameters fitted exclusively on the resulting training set. The same parameters are applied to the validation and test sets. Sliding-window samples use the preceding 96 observations to predict the load at the next four time steps. Training and validation targets remain within the initial training portion, while testing windows are confined to the testing portion. The validation set is used for early stopping and checkpoint selection, and the test set is used exclusively for final performance evaluation. Windows containing missing or infinite values in their inputs are discarded, and no interpolation or missing-value imputation is performed.

The experiments are implemented using Python 3.9.21 and PyTorch 2.7.0 + cu128 (CUDA 12.8 build) on a single NVIDIA RTX 5090 GPU. The temporal branch contains three parallel scale modules, each comprising two Transformer encoder layers, giving six encoder layers in total across the three modules. Each encoder layer uses a feed-forward network dimension of 512, three attention heads, and a dropout rate of 0.1. The patch lengths are 4, 8, and 12, with corresponding strides of 4, 8, and 12. The learnable frequency-filter tensor has a shape of 49×30. The quantile levels are set to 0.1, 0.5, and 0.9.

The Adam optimizer is used with a batch size of 64 and no weight decay. The learning rate remains fixed at 1×10^{-4} throughout training, and no learning-rate scheduler is used.

Models are trained for a maximum of 300 epochs. Early stopping monitors the joint Pinball Loss on the validation set, with a patience of 15 epochs and a minimum improvement threshold of 10^{-5}. Training is terminated when the validation loss fails to decrease by more than this threshold for 15 consecutive epochs. The checkpoint with the lowest validation loss is retained for final testing. The fixed-seed experiment uses a random seed of 100. No validation-based hyperparameter search is conducted. The validation set is used for early stopping and checkpoint selection, while the test set is used exclusively for final performance evaluation.

4.1. Evaluation Metrics

To comprehensively and objectively evaluate the practical application value of the model in power grid dispatching, this paper constructs a dual evaluation system that includes both point forecasting and interval forecasting.

For point forecasting performance, the median prediction with $q = 0.5$ is taken as the baseline predicted value $\hat{y}_i$. The mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) are selected as evaluation metrics. Equations (18)–(22) give the metric definitions for a single forecasting step, where N denotes the number of test samples and $\overline{y}$ is the mean observed load at that step. For the four-step results reported in the tables, MSE, MAE, and R^2 are calculated separately at each forecasting step and then averaged with equal weights. Overall RMSE is obtained by taking the square root of the averaged MSE. Smaller values of MSE, RMSE, MAE, and MAPE, together with an R^2 value closer to 1, indicate smaller prediction errors and higher fitting accuracy:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{18}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{19}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{20}$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{21}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{22}$$

For interval forecasting performance, $q = 0.1$ is taken as the lower bound L_i, and $q = 0.9$ is taken as the upper bound U_i, thereby constructing a prediction interval with a nominal confidence level of 80%. The prediction interval coverage probability (PICP) and the prediction interval normalized average width (PINAW) are adopted for comprehensive evaluation:

$$\mathrm{PICP} = \frac{1}{N}\sum_{i=1}^{N} c_i, \quad c_i = \begin{cases} 1, & L_i \le y_i \le U_i \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

$$\mathrm{PINAW} = \frac{1}{N \cdot R}\sum_{i=1}^{N}(U_i - L_i) \tag{24}$$

where N is the total number of samples in the testing set, y_i is the observed actual load, and R is the range of the actual load in the testing set, namely the difference between the maximum and minimum values. A probabilistic forecasting model with high engineering practicality should reduce PINAW as much as possible while ensuring that PICP meets the required level.

Additionally, Winkler scores, Pinball Loss, quantile crossing rates, and reliability diagrams for the four forecasting steps are used to supplement the evaluation. For a nominal $1-\alpha$ prediction interval with lower bound L, upper bound U, and observation y, the Winkler score is defined as:

$$W_\alpha(L,U;y) = (U-L) + \frac{2}{\alpha}(L-y)\mathbb{I}(y<L) + \frac{2}{\alpha}(y-U)\mathbb{I}(y>U) \quad (25)$$

where $\alpha = 0.2$ for the nominal 80% interval. Lower scores indicate a more favorable combination of interval width and penalties for missed observations. Pinball Loss evaluates quantile prediction accuracy using the loss defined in Section 2.1, while the quantile crossing rate measures violations of the predicted quantile ordering. In the reliability diagrams, each nominal quantile is compared with the observed fraction of outcomes below its prediction; ideal empirical calibration corresponds to the diagonal line.

4.2. Forecasting Performance Evaluation and Comparative Analysis

To comprehensively evaluate the effectiveness and superiority of the proposed model, several representative models, including classical deep learning models and advanced load forecasting models, are selected for comparison on two datasets: (1) the Quanzhou main dataset and (2) the Australian dataset. For the probabilistic forecasting comparison, all baseline models are equipped with quantile prediction heads and retrained to jointly predict the quantiles at levels 0.1, 0.5, and 0.9. The same quantile levels and joint Pinball Loss used by the proposed model are applied to all baselines. The 0.5 quantile prediction is used to calculate point forecasting metrics, while the 0.1 and 0.9 quantile predictions directly form the lower and upper bounds of the nominal 80% prediction interval. No residual-based interval estimation is used. All models use the same input features, chronological training–validation–test split, 96-step input window, and four-step forecasting horizon. They are trained using the same optimizer, learning rate, batch size, maximum epoch limit, and validation-based early stopping and checkpoint-selection criteria described in Section 4. All methods are evaluated using the same metric definitions on the corresponding test sets.

As shown in Table 2, in terms of point forecasting accuracy (RMSE, MAE, and MAPE), the proposed model achieves superior performance on both datasets. On the Quanzhou dataset, the MAPE of the proposed model is only 0.49%. Compared with MCTF-Net [15] and ST-ScaleFusion [16], its RMSE is significantly reduced by 22.5% and 29.8%, respectively. On the Australian dataset, the proposed model still maintains the best performance, with the MAPE controlled at 0.77%. These results support its effectiveness on the two evaluated datasets.

Interval forecasting performance is assessed jointly through empirical coverage and normalized interval width. On the Quanzhou dataset, the proposed model achieves a PICP of 0.7918, which is 0.82 percentage points below the nominal 80% level, and a PINAW of 0.0119. Compared with MCTF-Net, which obtains a PICP of 0.7245 and a PINAW of 0.0126, the proposed model provides coverage closer to the nominal level together with narrower intervals.

On the Australian dataset, the proposed model achieves a PICP of 0.8065 and a PINAW of 0.0274. MCTF-Net produces narrower intervals, with a PINAW of 0.0228, but its PICP of 0.7105 falls 8.95 percentage points below the nominal level. The proposed model reduces the absolute coverage deviation to 0.65 percentage points, accompanied by an increase in interval width. These results indicate a favorable coverage–width trade-off on the evaluated datasets. No post-training calibration is applied; the interval bounds are obtained directly from the predicted 0.1 and 0.9 quantiles.

Table 2. Comparison of forecasting results with different models on two datasets.

Dataset	Model	MSE (MW^2)	RMSE (MW)	MAE (MW)	MAPE	R^2	PICP	PINAW
Quanzhou	FEDformer [11]	41,024	202	99	1.27%	0.9790	0.6052	0.0121
	LSTM [8]	23,722	154	103	1.22%	0.9895	0.5265	0.0167
	PatchTST [13]	16,990	130	80	0.98%	0.9957	0.6888	0.0194
	TimesNet [19]	13,107	114	68	0.87%	0.9967	0.5554	0.0126
	iTransformer [14]	10,483	102	60	0.75%	0.9973	0.6052	0.0121
	ST-ScaleFusion [16]	8930	94	55	0.67%	0.9978	0.6510	0.0135
	MCTF-Net [15]	7320	85	49	0.60%	0.9982	0.7245	0.0126
	Proposed model	4399	66	39	0.49%	0.9989	0.7918	0.0119
Australia	FEDformer [11]	38,332	195	117	1.34%	0.9707	0.7718	0.0458
	LSTM [8]	22,511	150	109	1.26%	0.9797	0.6219	0.0303
	PatchTST [13]	18,356	135	101	1.15%	0.9903	0.6151	0.0297
	TimesNet [19]	18,207	134	99	1.12%	0.9904	0.4219	0.0167
	iTransformer [14]	14,252	119	86	0.96%	0.9925	0.5362	0.0182
	ST-ScaleFusion [16]	12,650	112	80	0.89%	0.9932	0.6034	0.0196
	MCTF-Net [15]	11,320	106	76	0.84%	0.9939	0.7105	0.0228
	Proposed model	9375	96	68	0.77%	0.9943	0.8065	0.0274

To supplement the comparison in Table 2 and assess variability across random seeds, the proposed model is evaluated using five additional random seeds on each dataset.

SD denotes standard deviation. For each random-seed run, MSE, RMSE, MAE, and R^2 are calculated using the four-step aggregation procedure described in Section 4.1. Their means and standard deviations are then computed across the five runs. The standard deviations of MAPE and PICP are expressed in percentage points. As shown in Table 3, the mean RMSE and MAPE on the Quanzhou dataset are 68 MW and 0.48%, with standard deviations of 1.46 MW and 0.0038 percentage points, respectively. On the Australian dataset, the corresponding mean values are 101 MW and 0.79%, with standard deviations of 1.83 MW and 0.0125 percentage points. These results indicate relatively limited variation in point forecasting performance across the five runs. For interval forecasting, the mean PICP values are 77.70% and 78.06%, with corresponding mean PINAW values of 0.0109 and 0.0273. The mean coverage probabilities on both datasets are close to the nominal coverage level of 80%.

Table 3. Mean and standard deviation of forecasting metrics across five random seeds.

Metric	Quanzhou Mean	Quanzhou SD	Australian Mean	Australian SD
MSE (MW^2)	4679	200	10,251	373
RMSE (MW)	68	1.46	101	1.83
MAE (MW)	38	0.31	70	1.21
MAPE (%)	0.48	0.0038	0.79	0.0125
R^2	0.9988	0.0001	0.9945	0.0001
PICP (%)	77.70	0.79	78.06	0.49
PINAW	0.0109	0.0002	0.0273	0.0008

To more intuitively demonstrate the probabilistic forecasting performance of the proposed model, Figure 3a,b presents the continuous prediction curves of the model over a selected period on the two datasets. As shown in Figure 3a,b, the 80% prediction interval constructed by $q = 0.1$ and $q = 0.9$ tightly and smoothly envelops the fluctuation trajectory of the actual load. In particular, the enlarged local views clearly show that even during high-frequency nonlinear abrupt-change periods when the load rises sharply, the median

prediction curve ($q = 0.5$, green line) can still closely follow the evolution trend of the actual values. This further verifies the modeling capability of the dual-branch architecture in handling highly volatile power grid data.

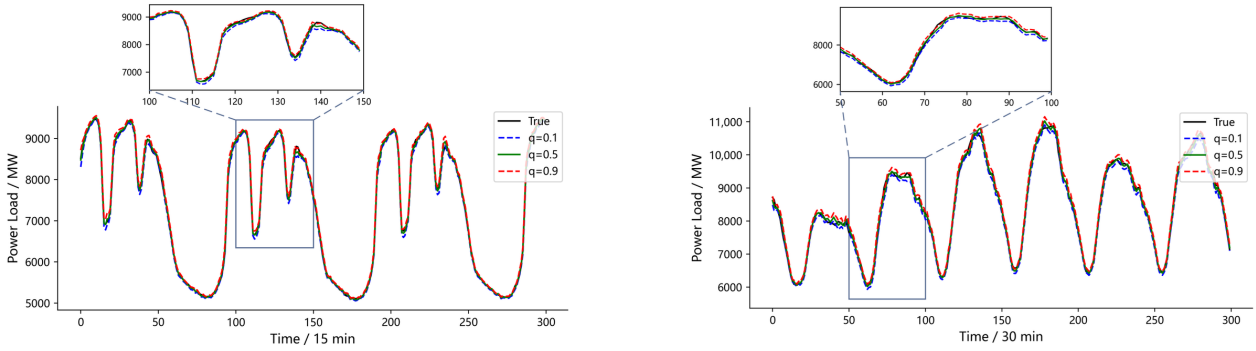


Figure 3. Interval forecasting results on different datasets: (**a**) Quanzhou dataset; (**b**) Australian dataset.

For supplementary evaluation, additional probabilistic forecasting diagnostics are reported. Over the four forecasting steps, the mean Winkler scores are 179 and 335 for the Quanzhou and Australian datasets, respectively. The corresponding Pinball Loss values are 12.4 and 22.5. No quantile crossings are observed on the Quanzhou dataset, while the quantile crossing rate on the Australian dataset is 0.008%.

Figure 4a,b examines empirical quantile calibration separately for each forecasting step. The observed fractions generally follow the nominal quantile levels, although departures from the diagonal remain. The largest absolute deviations among the displayed points occur at the fourth-step median forecasts: 0.473 for Quanzhou and 0.463 for Australia, compared with the nominal level of 0.5. These correspond to deviations of 2.7 and 3.7 percentage points, respectively.

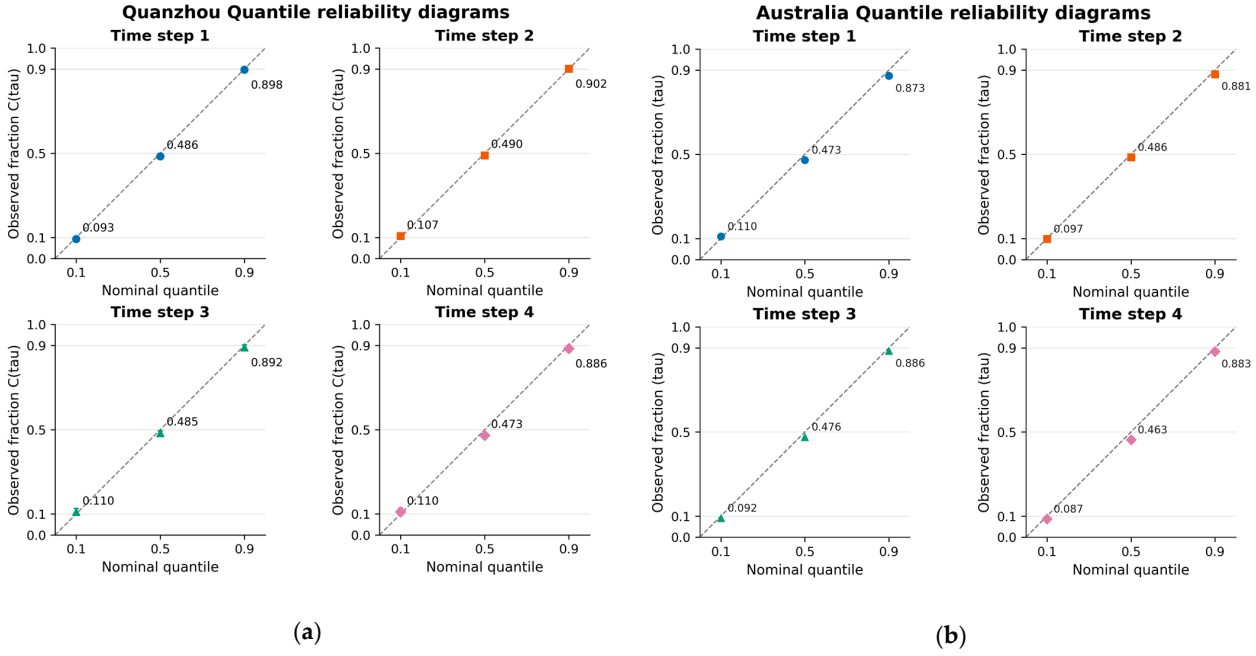


Figure 4. Quantile reliability diagrams for the four forecasting steps: (**a**) Quanzhou dataset; (**b**) Australian dataset. The dashed diagonal indicates ideal empirical calibration.

To further evaluate forecasting performance under different seasonal conditions, seasonal subsets are constructed according to the local seasonal definitions for Quanzhou and Australia. The seasonal subsets contain 25,231–26,199 samples for Quanzhou and 20,657–21,585 samples for Australia. Each seasonal subset is independently divided into training, validation, and testing sets at an 80%/10%/10% ratio, and forecasting metrics are calculated on the corresponding seasonal test set. These experiments assess performance under season-specific training and testing conditions. The comparative forecasting results for Quanzhou and Australia are presented in Tables 4 and 5, respectively.

Table 4. Seasonal forecasting performance on the Quanzhou dataset.

Season	PatchTST [13]		iTransformer [14]		ST-ScaleFusion [16]		MCTF-Net [15]		Proposed Model	
	RMSE (MW)	MAPE	RMSE (MW)	MAPE	RMSE (MW)	MAPE	RMSE (MW)	MAPE	RMSE (MW)	MAPE
Spring	120	1.27%	113	1.40%	96	1.12%	81	0.86%	67	0.62%
Summer	138	0.85%	118	0.99%	104	0.82%	90	0.69%	83	0.49%
Autumn	137	0.94%	119	1.29%	101	0.98%	87	0.87%	78	0.52%
Winter	113	1.52%	89	1.28%	81	1.04%	72	0.81%	60	0.62%

Table 5. Seasonal forecasting performance on the Australian dataset.

Season	PatchTST [13]		iTransformer [14]		ST-ScaleFusion [16]		MCTF-Net [15]		Proposed Model	
	RMSE (MW)	MAPE	RMSE (MW)	MAPE	RMSE (MW)	MAPE	RMSE (MW)	MAPE	RMSE (MW)	MAPE
Spring	131	1.23%	122	1.14%	113	1.06%	104	0.93%	95	0.82%
Summer	145	1.29%	141	1.23%	131	1.11%	123	1.03%	116	0.93%
Autumn	138	1.17%	116	1.05%	119	0.97%	108	0.89%	106	0.82%
Winter	157	1.26%	142	1.14%	129	0.99%	116	0.87%	105	0.77%

Since power load is strongly influenced by seasonal meteorological variations, it exhibits pronounced seasonal fluctuations, and the load components differ substantially across seasons. The seasonal comparisons show that the proposed model achieves good forecasting performance on the evaluated seasonal subsets of both the Quanzhou and Australian datasets.

4.3. Ablation Study

To assess the contribution of each core component across both datasets, five ablation variants are evaluated on the Quanzhou and Australian datasets: removing RevIN, removing the frequency-domain branch, removing the frequency-domain filter, removing the temporal branch, and replacing gated fusion with static equal-weight addition. The results for Quanzhou and Australia are presented in Table 6 and Table 7, respectively.

First, on the Quanzhou dataset, removing the temporal branch leads to the most significant increase in model error, with MAPE increasing by 95.9% compared with the original model. This result highlights the importance of the temporal branch in capturing load dynamics and improving point forecasting accuracy. In addition, removing the RevIN module and the frequency-domain branch increases MAPE by 53.1% and 42.9%, respectively, indicating that both components contribute to improving point forecasting accuracy.

Second, in terms of probabilistic forecasting reliability, removing the RevIN module causes the most severe degradation in coverage probability, reducing PICP by 15.9% compared with the original model. This result suggests that RevIN contributes to maintaining prediction interval coverage, potentially through its ability to mitigate the effects of distribution shifts. Meanwhile, removing the frequency-domain filtering mechanism reduces

PICP by 10.5%, demonstrating the contribution of the learnable filter to maintaining interval stability. Finally, replacing gated fusion with static equal-weight addition slightly increases forecasting errors, supporting the benefit of adaptively combining temporal and frequency-domain features.

Table 6. Ablation study results on the Quanzhou dataset.

Model Variant	MSE (MW^2)	RMSE (MW)	MAE (MW)	MAPE	R^2	PICP	PINAW
Original Model	4399	66	39	0.49%	0.9989	0.7918	0.0119
Without RevIN Module	11,242	106	59	0.75%	0.9971	0.6662	0.0139
Without Frequency-Domain Branch	10,395	101	56	0.70%	0.9973	0.7691	0.0168
Without Frequency-Domain Filtering	8419	91	51	0.63%	0.9979	0.7085	0.0127
Without Temporal Branch	21,252	145	73	0.96%	0.9946	0.7682	0.0203
Without Gated Fusion	6867	82	44	0.56%	0.9982	0.7871	0.0132

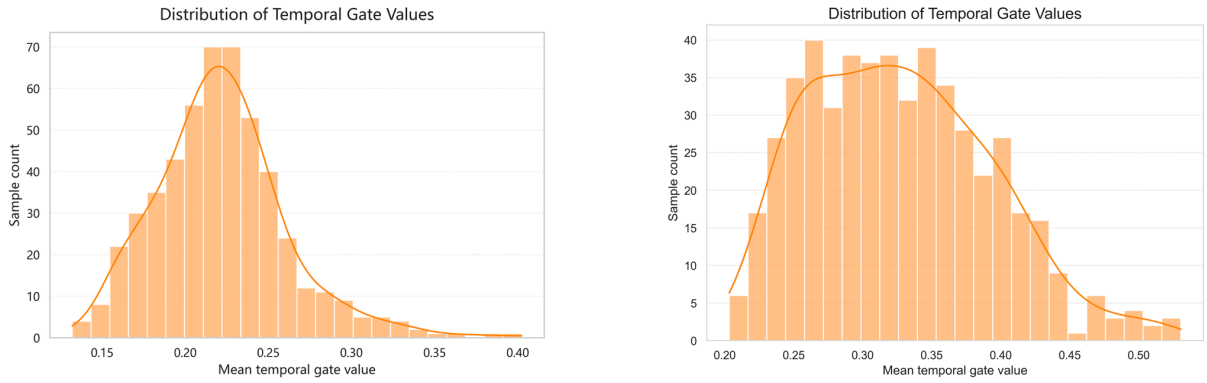
Table 7. Ablation study results on the Australian dataset.

Model Variant	MSE (MW^2)	RMSE (MW)	MAE (MW)	MAPE	R^2	PICP	PINAW
Original Model	9375	96	68	0.77%	0.9943	0.8065	0.0274
Without RevIN Module	16,703	129	83	0.94%	0.9912	0.8923	0.0332
Without Frequency-Domain Branch	15,654	125	80	0.91%	0.9917	0.8676	0.0324
Without Frequency-Domain Filtering	14,973	122	79	0.89%	0.9921	0.8492	0.0277
Without Temporal Branch	19,582	139	93	1.05%	0.9896	0.8408	0.0421
Without Gated Fusion	10,857	104	73	0.83%	0.9942	0.8359	0.0285

The Australian results show a consistent pattern in point forecasting performance. Removing the temporal branch causes the largest error increase, with MAPE rising from 0.77% to 1.05%, a relative increase of 36.4%. Removing RevIN, the frequency-domain branch, or the frequency-domain filter also increases forecasting errors. Replacing gated fusion with equal-weight addition increases MAPE to 0.83%, supporting the benefit of adaptive fusion. For interval forecasting, all five variants produce wider intervals and coverage probabilities above the nominal 80% level. The full model achieves coverage closest to the nominal level while maintaining the smallest PINAW among the evaluated variants. These results provide further evidence for the contributions of the model components on the Australian dataset.

Beyond the component ablation experiments, this study further investigates the distribution of learned gating values on two datasets to explore the internal behavior of the proposed dynamic fusion mechanism.

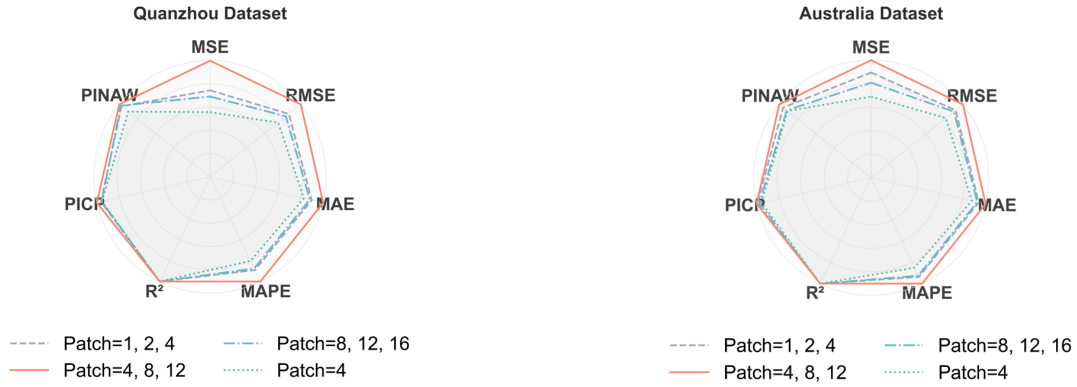
As shown in Figure 5a,b, the mean temporal gating values exhibit continuous and non-degenerate distributions on both datasets, rather than concentrating at a fixed value. This indicates that the proposed gating mechanism generates sample-dependent fusion weights and does not reduce to static equal-weight aggregation. Moreover, the gating values are mainly distributed below 0.5. According to Equation (14), this result indicates that the frequency-domain representation generally receives a relatively larger fusion coefficient, while the temporal branch continues to participate in the fused representation with sample-varying weights. The differences between the two distributions further suggest that the learned fusion patterns are affected by the characteristics of the input data. These observations provide additional evidence for the adaptive coordination of temporal and frequency-domain representations.



(**a**) (**b**)

Figure 5. Distributions of the learned temporal gating values on different datasets: (**a**) Quanzhou dataset; (**b**) Australian dataset.

To further investigate the impact of different scale combinations in the temporal branch, this paper includes an additional multi-scale parameter ablation experiment. Figure 6a,b presents the comprehensive performance radar chart of the model when using a single scale and three different multi-scale combinations.



(**a**) (**b**)

Figure 6. Performance comparison of different scale combinations on two datasets: (**a**) Quanzhou dataset; (**b**) Australian dataset.

As shown in Figure 6a,b, using only a single scale (Patch = 4) produces the smallest enclosed area in the radar charts. Among the multi-scale combinations, both an excessively small receptive field (Patch = {1,2,4}) and an excessively large receptive field (Patch = {8,12,16}) affect the forecasting performance of the model. Among the evaluated scale combinations, Patch = {4,8,12} achieves the best overall performance across the point and interval forecasting metrics on both datasets.

4.4. SHAP-Based Post Hoc Interpretability Analysis

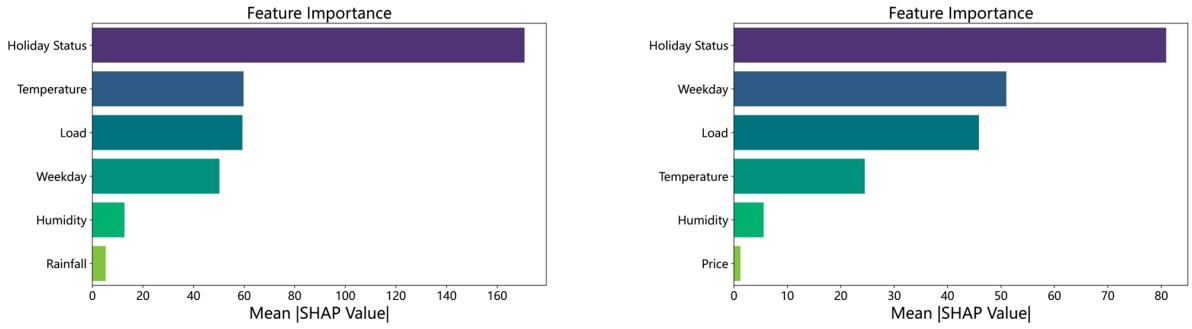
In practical power system dispatching, interpretability helps users understand the factors contributing to load forecasts. This study therefore employs SHAP-based post hoc attribution to quantify the contributions of input features to the model predictions.

Specifically, a gradient-based SHAP explainer using the expected gradients method is employed to approximate feature attribution values. After invalid windows are excluded, 150 complete sliding windows are randomly sampled without replacement from the training set to form the background set, and 50 windows are randomly sampled without replacement

from the test set for attribution analysis. Each window contains 96 historical time steps and 15 input variables. No clustering or stratified sampling is used to select the background samples. The scalar output being explained is the mean of the model's $q = 0.5$ predictions over the four future time steps. Thus, the resulting SHAP values represent input contributions to the average median forecast over the prediction horizon.

For visualization, SHAP results are presented for three feature groups and three individual features. The temperature group includes the three temperature variables, the weekday group includes the seven one-hot day-of-week indicators, and the holiday group combines the holiday and working-day indicators. Signed SHAP values are summed within these three groups. Historical load, humidity, and either rainfall for Quanzhou or electricity price for Australia are analyzed individually, without cross-variable aggregation. Further aggregation across time steps and samples is performed according to the purpose of each visualization.

Figure 7a,b presents the global feature importance obtained from SHAP attribution on the two datasets. "Holiday status" ranks first in both datasets, indicating that this feature group receives the largest mean absolute attribution among the reported feature groups for the explained samples. On the Quanzhou dataset, the "temperature group" has greater attribution importance than "historical load", while "historical load" and "weekday features" receive relatively high importance on the Australian dataset. These differences characterize the feature-attribution patterns of the model predictions on the two datasets.



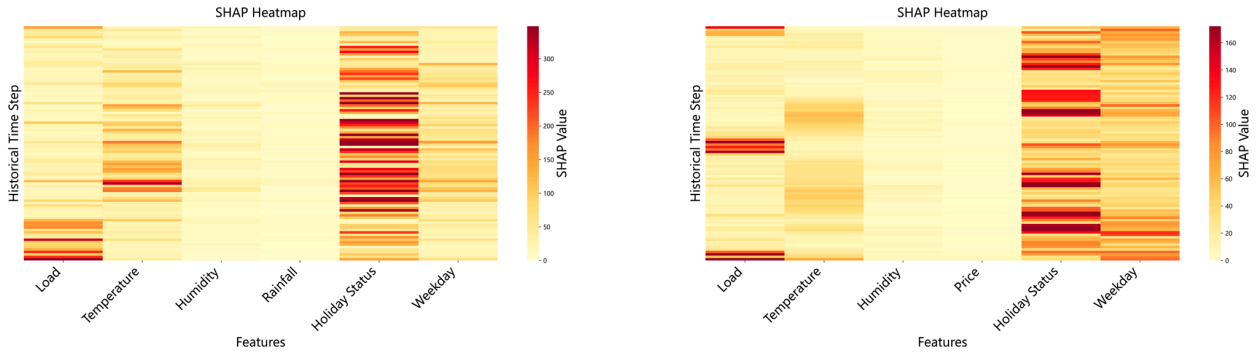
(**a**) (**b**)

Figure 7. Global feature importance on different datasets: (**a**) Quanzhou dataset; (**b**) Australian dataset.

As shown in Figure 8a,b, time-feature heatmaps are plotted in this paper to investigate the temporal attribution patterns of different variables. The heatmaps show that the attribution magnitudes of different features are not uniformly distributed but instead exhibit significant local concentration and periodic evolution patterns. In both datasets, "holiday status" presents continuous high-intensity color bands, indicating relatively large attribution magnitudes across multiple historical time steps. In the Quanzhou dataset, the "temperature group" and "historical load" show dark clustered regions with extremely high SHAP values at recent time steps. In contrast, in the Australian dataset, "historical load" and "weekday features" exhibit regular periodic stripe patterns, reflecting their strong cyclical characteristics.

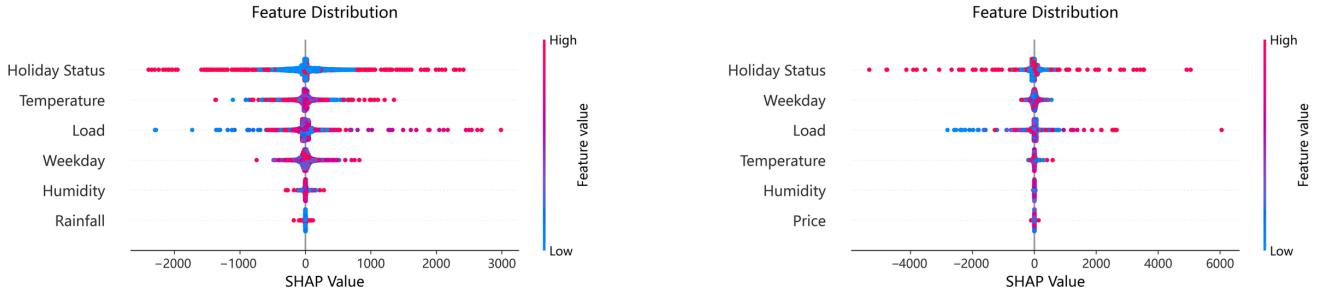
As shown in Figure 9a,b, SHAP aggregated feature distribution plots are introduced in this paper. Each scatter point in the plots represents a time-step sample in the test set. The color of the scatter points changes from blue to red, indicating that the corresponding feature value varies from low to high, while the horizontal axis quantifies the positive or negative contribution of that feature to the final predicted load. By observing the distribution plots of the Quanzhou and Australian datasets, it can be found that both datasets exhibit an obvious

positive contribution trend for the "historical load" feature, with red points concentrated on the positive side and blue points concentrated on the negative side. In the displayed samples from the Quanzhou dataset, higher values of the "temperature group" tend to correspond to positive SHAP values, indicating that these inputs contribute to increasing the model's predicted load.



(**a**) (**b**)

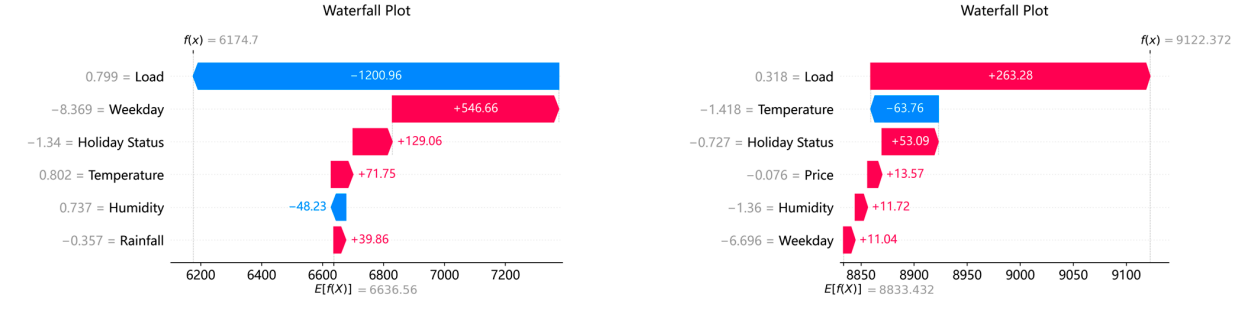
Figure 8. SHAP time-feature heatmaps on different datasets: (**a**) Quanzhou dataset; (**b**) Australian dataset.



(a) **(b)**

Figure 9. SHAP summary plots on different datasets: (**a**) Quanzhou dataset; (**b**) Australian dataset.

As shown in Figure 10a,b, the SHAP waterfall plots illustrate the feature contributions to individual predictions. Taking the Quanzhou dataset as an example, as the prediction changes from the baseline expectation (6636.56) to the final predicted value (6174.7), "historical load" imposes a strong negative correction (−1200.96), while "weekday features" and "holiday status" provide positive compensation, preventing the predicted value from being excessively pulled downward.



(**a**) (**b**)

Figure 10. SHAP waterfall plots on different datasets: (**a**) Quanzhou dataset; (**b**) Australian dataset.

4.5. Model Size

To provide a quantitative reference for the size of the proposed model, Table 8 reports the number of trainable parameters in each component. The model contains 202,161 trainable parameters in total. Temporal feature extraction and frequency-domain processing account for most of the parameters, while normalization, fusion projections, gated fusion, and quantile prediction heads require relatively few parameters.

Table 8. Trainable parameter counts of the proposed model.

Component	Number of Trainable Parameters
RevIN	30
Multi-scale temporal feature extraction	105,597
Multi-scale fusion projection	690
Frequency-domain branch	95,187
Gated fusion	465
Quantile prediction heads	192
Total	202,161

5. Conclusions

To address the non-stationarity and "black-box" challenges of load forecasting in new power systems, this paper proposes a time-frequency dual-branch architecture that combines complementary temporal and spectral representations. Learnable frequency-domain filtering and multi-scale temporal patching are employed to refine global periodic patterns and capture local load dynamics, respectively, while dynamic gating enables input-dependent feature fusion. On the Quanzhou and Australian datasets, the model achieves MAPE values of 0.49% and 0.77%, respectively. Compared with MCTF-Net, the strongest point forecasting benchmark in the experiments, RMSE is reduced by 22.5% on the Quanzhou dataset and approximately 9% on the Australian dataset. By combining reversible instance normalization with quantile regression, the model produces nominal 80% prediction intervals with empirical coverage probabilities of 79.18% and 80.65%, respectively. The corresponding PINAW values of 0.0119 and 0.0274 indicate a favorable balance between coverage and interval width on the evaluated datasets. In addition, SHAP-based post hoc attribution quantifies the contributions of input features, providing further insight into the model's predictions.

Despite the favorable results, the present evaluation is limited to two datasets and a fixed configuration of 96 historical time steps and four forecasting steps. Although the datasets represent different geographical and meteorological contexts, further evaluation is needed to assess the model's generalizability to other load systems. Moreover, the current experiments do not establish its performance across different historical window lengths or longer forecasting horizons. Future work will extend the evaluation to additional regions and forecasting configurations, with particular attention to longer horizons and robustness under changing load distributions. Further comparisons will examine point forecasting and multi-quantile training objectives within the same backbone, as well as alternative probabilistic forecasting approaches. Regarding interpretability, future studies will compare SHAP with alternative post hoc explanation methods and further analyze temporal- and frequency-domain contributions and dynamic gating behavior under different load conditions. The stability of feature attributions across forecasting horizons and operating conditions will also be examined.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/en19215160/s1, Supplementary Materials S1: Quanzhou and Australian datasets.

Author Contributions: Conceptualization, H.X. and F.L.; methodology, H.X.; software, H.X.; validation, F.L.; formal analysis, H.X.; investigation, H.X.; data curation, H.X.; writing—original draft preparation, H.X.; writing—review and editing, F.L.; supervision, F.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by grants from the Engineering Research Center of Offshore Wind Technology Ministry of Education (Shanghai University of Electric Power).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

ARIMA	Autoregressive integrated moving average
SVM	Support vector machine
LSTM	Long short-term memory
TCN	Temporal convolutional network
SHAP	SHapley Additive exPlanations
RevIN	Reversible instance normalization
rFFT	Real fast Fourier transform
iFFT	Inverse fast Fourier transform
MSE	Mean squared error
RMSE	Root mean square error
MAE	Mean absolute error
MAPE	Mean absolute percentage error
PICP	Prediction interval coverage probability
PINAW	Prediction interval normalized average width

References

1. Khalili, S.; Lopez, G.; Breyer, C. Role and trends of flexibility options in 100% renewable energy system analyses towards the Power-to-X Economy. *Renew. Sustain. Energy Rev.* **2025**, *212*, 115383. [CrossRef]
2. Dong, X.; Guo, W.; Zhou, C.; Luo, Y.; Tian, Z.; Zhang, L.; Wu, X.; Liu, B. Hybrid model for robust and accurate forecasting building electricity demand combining physical and data-driven methods. *Energy* **2024**, *311*, 133309. [CrossRef]
3. Wang, Q.; Liu, C.; Zhou, Z.; Han, Y.; Fang, L.; Zhao, M.; Cao, X. Grid-supplied load prediction under extreme weather conditions based on CNN-BiLSTM-attention model with transfer learning. *Energy Eng.* **2025**, *122*, 4715–4732. [CrossRef]
4. Liu, J.; Huang, Z.; Xiang, J.; Liu, L.; Hu, M. Seasonal short-term load forecasting for power systems based on modal decomposition and feature-fusion multi-algorithm hybrid neural network model. *Energy Eng.* **2024**, *121*, 3461–3486. [CrossRef]
5. Liu, W.; Mao, Z.; Jiang, L. A cloud-edge collaborative framework for power load forecasting using deep learning. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 1–12. [CrossRef]
6. Tarmanini, C.; Sarma, N.; Gezegin, C.; Ozgonenel, O. Short term load forecasting based on ARIMA and ANN approaches. *Energy Rep.* **2023**, *9*, 550–557. [CrossRef]
7. Wang, Z.; Ku, Y.; Liu, J. The power load forecasting model of combined SaDE-ELM and FA-CAWOA-SVM based on CSSA. *IEEE Access* **2024**, *12*, 41870–41882. [CrossRef]
8. Wang, M.; Xin, R.; Xia, M.; Zuo, Z.; Ge, Y.; Zhang, P.; Ye, H. A federated LSTM network for load forecasting using multi-source data with homomorphic encryption. *AIMS Energy* **2025**, *13*, 265–289. [CrossRef]
9. Dai, Y.; Yu, W. Short-term power load forecasting based on Seq2Seq model integrating Bayesian optimization, temporal convolutional network and attention. *Appl. Soft Comput.* **2024**, *166*, 112248. [CrossRef]
10. Ahmad, A.; Xiao, X.; Mo, H.; Dong, D. TFTformer: A novel transformer based model for short-term load forecasting. *Int. J. Electr. Power Energy Syst.* **2025**, *166*, 110549. [CrossRef]
11. Hou, K.; Zhang, X.; Yang, J.; Hu, J.; Yao, G.; Zhang, J. Short-term load forecasting based on multi-frequency sequence feature analysis and multi-point modified FEDformer. *Front. Energy Res.* **2025**, *12*, 1524319. [CrossRef]

12. Zeng, L.; Zheng, K.; Lin, G.; Yang, J.; Wu, M.; Chen, G.; Jiang, H. Enhanced short-term load forecasting based on adaptive residual fusion of Autoformer and Transformer. *Energies* **2025**, *18*, 6496. [CrossRef]
13. Qu, Z.; Meng, Y.; Hou, X.; Chi, R.; Ai, Y.; Wu, Z. Integrated energy short-term multivariate load forecasting based on PatchTST secondary decoupling reconstruction for progressive layered extraction multi-task learning network. *Expert Syst. Appl.* **2025**, *269*, 126446. [CrossRef]
14. Meng, X.; Wang, J.; Li, D.; Zhang, H.; Sun, Q.; Wang, H. ICEEMDAN- and LSTM-enhanced hybrid cosine-attention iTransformer for ultra-short-term load forecasting. *Electronics* **2025**, *14*, 4857. [CrossRef]
15. Gao, W.; Tian, Z. End-to-end mid-term load forecasting framework based on multi-channel technique and time-frequency domain fusion. *Appl. Soft Comput.* **2026**, *186*, 114029. [CrossRef]
16. Jiang, E.; Wang, Z.; Jiang, S. A multi-scale time–frequency complementary load forecasting method for integrated energy systems. *Energies* **2025**, *18*, 3103. [CrossRef]
17. Ye, H.; Chen, J.; Jiang, F.; He, X.; Zhang, T.; Chen, J.; Gao, X. Fremer: Lightweight and effective frequency transformer for workload forecasting in cloud services. *Proc. VLDB Endow.* **2025**, *18*, 3812–3825. [CrossRef]
18. Makatjane, K.; Sigauke, C.; Shoko, C.; Moroke, N. Real-time electricity load forecasting in South Africa using SOM-enriched deep learning ensembles. *AIMS Energy* **2026**, *14*, 310–334. [CrossRef]
19. Lin, R.; Chen, S.; Wu, B.; Zhao, X.; Li, Q. WaveTimes: Short-term load forecast based on wavelet decomposition and improved TimesNet. *Chin. J. Electron.* **2025**, *34*, 1911–1923. [CrossRef]

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