processes

MDPI

Article

Short-Term Load Forecasting Under Extreme Temperatures Based on Transformer–TCN and BiGRU Residual Correction

Shiyu Liu [1], Yating Wang [1,*], Yijun Wang [1], Qiang Liu [1], Hanyan Huang [1], Yanpei Zhao [2] and Xiaomin Xu [2]

[1] China Electric Power Planning and Engineering Institute Co., Ltd., Beijing 100120, China
[2] School of Economics and Management, North China Electric Power University, Beijing 102206, China
* Correspondence: 18192568929@163.com

Abstract

To address the intensified power load fluctuations and the degraded accuracy of conventional short-term load forecasting (STLF) models under extreme heat and extreme cold conditions, this paper proposes a two-stage STLF method that integrates XGBoost–RF feature selection, Transformer–TCN base forecasting, and BiGRU residual correction. First, a dual-model consensus importance evaluation mechanism is constructed using XGBoost and random forest (RF) to screen key input features from meteorological and calendar factors. Second, a Transformer is employed to model the global temporal dependencies of the load series, while a gated temporal convolutional network (TCN) extracts local variations and multi-scale periodic features. On this basis, a bidirectional gated recurrent unit (BiGRU) is introduced to model the historical forecasting residuals of the base model in a rolling manner, and the final forecasts are obtained through direct error correction. An empirical analysis is conducted using hourly load and meteorological data from 2019 to 2023 for a region in eastern China. The results show that, on the independent 2023 test set, the mean absolute error (MAE) of the proposed model is reduced by approximately 21.7% compared with the Transformer–TCN base model, and by approximately 62.6% and 42.8% under extreme heat and extreme cold scenarios, respectively, indicating that the residual correction mechanism can effectively improve forecasting performance under extreme temperatures. Ablation experiments and statistical tests further verify the effectiveness and stability of the model improvements. These results indicate that the proposed two-stage forecasting framework can simultaneously achieve overall forecasting accuracy and forecasting capability under extreme temperature scenarios, and can provide a reference for short-term load forecasting and power grid dispatch under extreme weather conditions.

Keywords: short-term load forecasting; extreme temperature; Transformer; TCN; BiGRU; SHAP

Check for updates

Academic Editor: Mattia Bartoli

Received: 24 October 2026
Revised: 19 November 2026
Accepted: 23 November 2026
Published: 26 November 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

Short-term load forecasting is a fundamental component of power grid dispatch planning, unit commitment, and electricity market trading, and its accuracy directly affects the economics and security of power system operation [1,2]. As the penetration of renewable generation continues to rise and end-use electrification keeps advancing, the volatility and uncertainty on the load side have increased considerably, and short-term load forecasting (STLF) has become a key technology for maintaining real-time power balance in the grid. Meanwhile, against the backdrop of global climate change, the frequency and intensity of extreme heat and cold wave events have both increased markedly, and the sensitivity of

electricity demand to temperature has strengthened accordingly [3]. Existing studies have shown that climate change not only raises average electricity demand but also markedly increases the frequency and magnitude of peak loads, thereby amplifying the investment and operational pressure on power systems [4,5]; a pronounced nonlinear relationship generally exists between temperature and electricity consumption, and meteorological factors such as humidity cannot be ignored in summer demand modeling [6]. In recent years, power grids in many countries have experienced load spikes far beyond historical levels during extreme temperature events, and large-scale blackouts have even been triggered: during the February 2021 cold wave in Texas, USA, more than 4.5 million users lost power during peak hours, with some outages lasting several days, exposing the vulnerability of power systems under extreme low temperatures [7,8]. Empirical research based on the Texas region further indicates that the sensitivity of peak load to temperature will continue to increase with climate warming [9]. Therefore, how to maintain high STLF accuracy under both extreme heat and extreme cold has become an urgent practical problem in the construction of new-type power systems.

From the perspective of methodological evolution, STLF was initially dominated by statistical methods such as multiple linear regression and time series analysis. Subsequently, machine learning methods such as support vector machines and random forests were introduced into load forecasting: Lahouar and Slama implemented day-ahead load forecasting using random forests combined with expert input selection, and verified the contribution of meteorological variables through feature importance analysis [10]; the XGBoost method proposed by Chen and Guestrin improves model generalization through second-order Taylor expansion and a regularized objective function, while providing a stable measure of feature importance, and has become a common tool for factor screening [11]. In recent years, deep learning methods have further improved forecasting accuracy: Kong et al. constructed a residential load forecasting framework based on long short-term memory (LSTM) networks, verifying the capability of recurrent networks to model temporal dependencies [12]; Sajjad et al. proposed a CNN-GRU hybrid model that accounts for both local feature extraction and long-range dependency modeling [13]; Lim et al. proposed the temporal fusion Transformer, which achieves interpretable multi-horizon forecasting with multi-source variables [14]; Zhou et al. and Wu et al. proposed Informer and Autoformer, respectively, extending Transformers to long-sequence forecasting scenarios [15,16]; and Song et al. extracted multi-scale temporal features of load series through gated convolution and captured the correlations among loads of different energy types using a spatiotemporal attention mechanism, effectively improving forecasting accuracy [17].

However, a single model can hardly characterize both the long-term trend and the local abrupt changes in a load series at the same time, so hybrid modeling and error correction have attracted extensive attention. Wang et al. (2021) introduced a decomposition-based error correction method into multi-step electricity load forecasting to improve the accuracy and stability of multi-step forecasts [18]. Fan et al. (2024) introduced a gated recurrent unit (GRU) into the error correction stage of industrial load forecasting to dynamically correct the forecasting bias of the base model, thereby improving the accuracy and stability of industrial load forecasting [19]. Laouafi et al. incorporated an error correction strategy into an adaptive ensemble model to improve short-term forecasting [20]. Subsequently, Xie et al. further proposed bias calibration and error propagation adjustment methods for regional power grid loads [21]. These studies show that base models usually achieve relatively high accuracy under normal meteorological conditions, but tend to produce systematic bias during periods of persistently high or low load; if the temporal structure of the residual is not explicitly modeled, this bias will be further amplified under extreme scenarios, forming a vicious circle.

For extreme temperature scenarios, existing studies mostly assess the long-term response relationship between temperature and load from climatological or demand-side perspectives [3,4,9]. Kim et al. evaluated changes in heat wave characteristics in South Korea and their impacts on electricity demand [22]; Yang et al. combined multi-source factors with a bidirectional gated recurrent network on wildfire-season distribution network data to study forecasting stability under extreme weather [23]. To address the shortage of heat wave samples, Xiao et al. adopted data augmentation and a dual-attention mechanism for interpretable short-term load forecasting [24]; Zhang et al. handled dynamic changes under extreme weather through contrastive online learning and concept drift detection [25]. At the methodological level, Liu et al. constructed EWELD, a large-scale industrial and commercial load dataset for extreme weather events, providing a data foundation for extreme scenario modeling [26]; Lu et al. proposed a data-centric load forecasting method under extreme heat and cold waves, and pointed out that the forecasting errors of conventional models during extreme temperatures are significantly higher than under normal weather [27].

From the reported quantitative results, the improvement margins and evaluation protocols differ considerably across studies. At the method level, Informer by Zhou et al. and Autoformer by Wu et al. reduce the mean squared error of long-sequence forecasting by approximately 41.5–60.7% and 38%, respectively, the latter decreasing from 0.280 to 0.231 (about 18%) on the Electricity dataset; the hierarchical multi-task spatiotemporal attention model of Song et al. reduces MAPE by up to 25.09%, 25.96%, and 17.46% for cooling, heating, and electrical loads. At the error-correction level, the hybrid ensemble error-correction model of Fan et al. achieves an NMAE of 3.684% and an NRMSE of 7.266% on Korean industrial load and outperforms eight comparative models; the adaptive ensemble method of Laouafi et al. attains a MAPE of 1.03–4.26% on data from New South Wales, Australia, outperforming 24 state-of-the-art methods; and the bias calibration method of Xie et al. nearly eliminates systematic bias in the PJM regional grid. At the extreme-weather level, the wildfire-season model of Yang et al. reduces MAPE by 14.37% and 20.86% for the Bi-GRU and ViT models, respectively, with an average MAPE of about 3%; the generative augmentation and dual-attention model of Xiao et al., developed for scarce heat-wave samples, reduces MAPE by 18.03% and 33.62% on residential and public-service data; and the extreme-sample synthesis–disentanglement method of Lu et al. over 80 datasets from six regions reduces nMAE by 9.8–33.5% and nRMSE by 7.5–34.5%, while noting that heat- and cold-wave samples account for only 6.4% of all data. Overall, the improvements reported under extreme weather are mostly in the 10–35% range and depend strongly on the data partitioning, baseline selection, and proportion of extreme samples; moreover, most studies do not report the extreme-sample proportion, event counts, and threshold definitions together, so results across studies are not directly comparable. This paper therefore reports errors stratified by temperature state, compares the proposed model with 11 baselines under the same data partitioning, and reports the sample size, extreme-event counts, and threshold sensitivity of each stratum.

It should be noted that the above studies share two common limitations in their evaluation protocols. First, most works report improvements as the average error over the whole year or all test samples [15,16,19], whereas extreme-temperature samples usually account for only a small proportion of the test set, so average metrics cannot reflect the true performance in the tail of the temperature distribution. This paper reports errors stratified by temperature state and simultaneously gives the sample size of each stratum, the number of extreme-weather events, and statistical test results, so that the improvement can be interpreted together with the sample structure. Second, error correction is mostly implemented by joint training or decomposition–reconstruction, in which the correction term and the base forecast are parametrically coupled [18,19], making it difficult to deter-

mine whether the performance gain comes from residual modeling or from refitting of the base model. This paper completely freezes the base-model parameters in the residual stage, and the correction module receives only the observed residual series, so the attribution of the improvement is clearer. In addition, some studies do not specify how extreme-scenario thresholds are defined or how sensitive the results are to them [22,23], which limits the comparability of results across studies; this is also why threshold sensitivity results are further provided here.

Overall, three shortcomings remain in existing research. First, extreme temperature samples account for a small proportion of training data, making it difficult for models to learn their specific load response patterns, which leads to relatively large forecasting errors during high- and low-temperature periods. Second, most hybrid models take overall average accuracy as the optimization objective, so error correction and base forecasting are separated from each other and the autocorrelation structure of the residual is not exploited. Third, model interpretability is insufficient, making it difficult to answer the question most concerning to dispatch operators: which factors dominate load changes under extreme heat or extreme cold. To address these shortcomings, this paper proposes a two-stage STLF framework that integrates XGBoost–RF dual-model factor screening, Transformer–TCN base forecasting, and BiGRU residual correction, and validates it using measured load and meteorological data from 2019 to 2023 for a region in eastern China. The main contributions of this paper are as follows:

(1) A Transformer–TCN base forecasting model is constructed, in which the Transformer models global temporal dependencies and a gated TCN supplements local load variation information, enabling rolling load forecasting.
(2) A BiGRU residual prediction and direct error correction module is designed, which corrects the base forecast using only the residual series observed before the forecasting instant, avoiding future information leakage, and substantially reduces forecasting errors under extreme heat and extreme cold scenarios.
(3) A scenario-based evaluation system is established for short-term load forecasting under extreme temperatures. Model performance is validated at various levels, including overall forecasting, extreme temperature conditions, and consecutive extreme temperature events. Feature importance analysis is used to examine changes in load drivers under different temperature conditions.

2. Forecasting Method and Framework

2.1. Feature Selection Based on XGBoost–RF

Feature selection aims to compress the input dimension while retaining key information, so that redundant variables do not increase model complexity and weaken generalization capability. The candidate factors in this paper consist of three categories—load, meteorological, and calendar variables. Load variables refer to the historical load series; meteorological variables include temperature, dew point, humidity, wind speed, solar radiation and precipitation; and calendar variables include hour, day of week, month, seasonal position, and holiday and working-day indicators. The raw data are first subjected to time alignment, outlier identification and replacement, missing-value imputation, and normalization [11], so as to eliminate dimensional differences and ensure strict temporal correspondence among variables.

To prevent information leakage from the validation and test sets, feature selection for XGBoost–RF was performed using only the 2019–2020 training data. The feature importance scores obtained from the two models were normalized and then averaged with equal weights. The features were then ranked in descending order of their combined importance, and the smallest subset of features whose cumulative contribution reached 95% was selected. Given that the feature importance of a single tree model is subject to randomness, this paper

adopts a dual-model consensus screening method composed of XGBoost and random forest. XGBoost additively integrates K regression trees, and its objective function is:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

$$\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \|w_k\|^2 \quad (2)$$

where $l(\ )$ is a differentiable convex loss function, $\Omega(\ )$ is the regularization term, T_k is the number of leaf nodes in the k-th tree, w_k is the leaf weight, and γ and λ are the regularization coefficients for the number of leaves and the leaf weights, respectively. XGBoost measures feature importance by the cumulative contribution of split gains in each iteration [12], whereas random forest measures feature importance by the decrease in mean squared error (reduction in impurity) of in-bag and out-of-bag samples. After normalizing the importance of the two models separately and averaging them, the consensus importance of feature k is obtained:

$$I_k^{con} = \frac{1}{2}(I_k^{XGB} + I_k^{RF}) \quad (3)$$

The consensus importances are sorted in descending order and their cumulative importance is calculated; the smallest feature subset whose cumulative importance is no less than 95% is then selected:

$$R_m = \frac{\sum_{k=1}^{m} I_{(k)}^{con}}{\sum_{k=1}^{K} I_{(k)}^{con}}, m^* = min\{m | R_m \geq 0.95\} \quad (4)$$

Based on these criteria, seven key features were selected from the candidate factors: historical load, seasonal position, holiday, temperature, time of day, dew point, and solar radiation. Sine-cosine periodic encoding is applied to seasonal position and time of day; after encoding, the seven feature categories collectively form nine input channels.

2.2. Base Forecasting Model Based on Transformer–TCN

The basic forecasting model takes filtered and feature-engineered multi-source time-series data as input and uses a 168-h sliding time window. Construct an input matrix consisting of historical load, meteorological, and periodic coding features from the week preceding the t-th forecast time step (dimension 168×9). All meteorological variables in the model inputs are based on historical meteorological observations obtained prior to the start of the forecast; actual meteorological observations from the target forecast time and beyond are not used, nor are future weather forecast data incorporated. The input sequence is first mapped to a d-dimensional embedding space through a linear projection, and positional encoding is added to preserve temporal ordering:

$$h_t^0 = W_e x_t + b_e + PE_t \quad (5)$$

where PE_t is the positional encoding. Subsequently, L layers of Transformer encoders extract global temporal dependencies layer by layer. Each layer contains two sublayers—multi-head self-attention and a feedforward network—and both sublayers adopt residual connections and layer normalization. The scaled dot-product attention is computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

Multi-head attention projects the queries, keys, and values into h subspaces for parallel computation, and then concatenates and linearly maps the results:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \quad (7)$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (8)$$

The self-attention mechanism enables the representation at each instant to access information from any instant within the window directly, which is conducive to characterizing the long-range periodicity of load and the persistent influence of meteorological processes [15]. However, a pure attention structure is not sufficiently sensitive to local abrupt changes (such as intraday ramping and the rapid load drop when a cold wave arrives), so a gated TCN is introduced in parallel as a local correction branch. TCN is based on causal dilated convolution, and the output at position i of layer l is:

$$F^{(l)}(i) = \sum_{j=0}^{k-1} f^{(l)}(j) \cdot x^{(l)}(i - d_l \cdot j) \quad (9)$$

where k is the kernel size and d_l is the dilation factor of layer l, which grows as 2^l; with depth, thereby enlarging the receptive field without increasing the number of parameters; causal padding ensures that the output at position i depends only on preceding inputs, avoiding information leakage. To enhance nonlinear representation capability, the convolution results are modulated by a gating unit:

$$z^{(l)} = \tanh\left(W_f * x^{(l)} + b_f\right) \odot \sigma\left(W_g * x^{(l)} + b_g\right) \quad (10)$$

where $\sigma(\ )$ is the sigmoid function and $\odot$ denotes element-wise multiplication. Considering that the load series exhibits the strongest autocorrelation at the three lag scales of 1 h, 24 h, and 168 h, the TCN branch sets lag anchors at these three positions; its output is concatenated with the attention output and then passed through a fully connected layer to obtain the base forecast:

$$\hat{y}_{base,t} = W_o\left[z_t^T; z_t^C\right] + b_o \quad (11)$$

where z_t^T and z_t^C are the representations of the Transformer branch and the TCN branch at instant t, respectively. The base model was trained on the 2019–2020 data, and the 2021 data were used as the validation set for model selection and early stopping, based on which hyperparameters such as the number of network layers, the number of attention heads, the dilation factor, and the kernel size were determined [16,18].

2.3. BiGRU Residual Prediction and Direct Error Correction Model

The base model can fit load variations well under normal meteorological conditions, but during periods of sustained high or low temperature it tends to produce systematic bias that is consistent in direction and cumulative in magnitude. Therefore, this paper constructs a BiGRU residual prediction model after fixing the parameters of the base model. Define the baseline forecast residual at t as:

$$\varepsilon_t = y_t - \hat{y}_{base,t} \quad (12)$$

A fixed Transformer–TCN model performs hourly 1-h-ahead forecasts for 2022 data, and the out-of-sample residuals are calculated using Equation (12). Using the observed baseline forecast residuals from the most recent $q = 10\ h$ as input, the model predicts the residuals that the baseline model might generate for the next hour:

$$\hat{\varepsilon}_t = f_\theta\left(\varepsilon_{t-q}, \varepsilon_{t-q+1}, \ldots, \varepsilon_{t-1}\right), q = 10 \quad (13)$$

In this equation, ε_t represents the actual prediction error of the baseline model, and $\hat{\varepsilon}_t$ represents the error predicted by the BiGRU. BiGRU performs bidirectional encoding only within a $10 - h$ window of historical residuals observed prior to time step t; therefore, its backward computation does not include ε_t or future residual information after t [19]. The

GRU cell uses reset and update gates to control the retention and updating of historical information. The calculation process is as follows:

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{14}$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{15}$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \tag{16}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{17}$$

BiGRU concatenates the forward and backward hidden states and maps them to the residual forecast; the final load forecast is obtained by directly adding the residual forecast to the base forecast:

$$h_t^{bi} = [\overrightarrow{h_t}; \overleftarrow{h_t}], \hat{e}_t = W_e h_t^{bi} + b_e \tag{18}$$

$$\hat{y}_{final,t} = \hat{y}_{base,t} + \hat{e}_t \tag{19}$$

The residual model was split chronologically based on the 2022 data, with the first 85% used for BiGRU training and the remaining 15% for parameter selection and validation; the 2023 data was used solely as an independent test set. Once model selection is complete, the internal validation subset is no longer incorporated into the training set for refitting. Ultimately, the BiGRU model parameters are learned from approximately 85% of the training subset, while the remaining 15% of the validation subset is used solely for model selection and early stopping. For any prediction time step t, the model first generates $\hat{e}_t$ using the historical residuals observed at the previous time step and then derives the final prediction. The window for the first test time step in 2023 is initialized using the residuals observed during the last 10 h of 2022.

2.4. SHAP-Based Interpretability Method

To quantify the contribution of each input feature to the forecast, this paper adopts the Shapley-value-based SHAP method for post-hoc explanation of the model [28]. For a given forecast sample, the Shapley value of feature i is defined as the weighted average of its marginal contributions over all feature subsets:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{20}$$

where N is the set of all input features and $f(S)$ is the model output when only the feature subset S is used. The model prediction can be decomposed into the sum of a base value and the contributions of individual features:

$$f(x) = \phi_0 + \sum_{i=1}^{|N|} \phi_i \tag{21}$$

$\varphi_i > 0$ indicates that the feature increases the forecast, whereas $\varphi_i < 0$ indicates that it decreases the forecast, and the absolute value reflects the strength of the influence. Because the Transformer–TCN models used here are non-tree-structured models, the gradient-based GradientSHAP estimator is employed to compute Shapley values, approximating marginal contributions by the expectation of the model output gradient with respect to input features over random baseline samples [29,30]. At the global level, the mean absolute Shapley value of each feature is used as the basis for importance ranking, and attribution results are computed separately for normal, extreme heat, and extreme cold samples to compare the dominant factors under different temperature states.

2.5. Model Evaluation Metrics

The root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) were selected as evaluation metrics for the forecasting models, defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{22}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{23}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100\% \tag{24}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \tag{25}$$

where n is the number of test samples, y_t and $\hat{y}_t$ are the actual and forecast load at instant t, respectively, and $\bar{y}$ is the mean actual load. RMSE and MAE reflect the absolute magnitude of forecasting errors (in MW), MAPE reflects the relative error level, and R^2 reflects the capability of the model to explain load fluctuations. Statistical significance tests were performed using the moving-block bootstrap and the Newey–West HAC methods, with 1000 bootstrap repetitions, a block length of 168 h, and an HAC lag order of 24 h. Paired sequences were constructed based on the difference in hourly absolute errors between the final model and the baseline model, and 95% confidence intervals and HAC test results were reported.

2.6. Overall Forecasting Framework

This paper constructs a short-term load forecasting framework consisting of factor screening and feature construction, Transformer–TCN base forecasting, BiGRU residual correction, and multilevel performance validation, as shown in Figure 1. First, the XGBoost–RF dual model is used for factor screening. Second, a base forecasting model is established through Transformer–TCN, and BiGRU is used to correct historical residuals in a rolling manner. Finally, the forecasting results are comprehensively evaluated in terms of overall performance, extreme temperature scenarios, statistical robustness, and model interpretability. For details on the key model parameters and experimental settings for each model in this paper, refer to Appendix A.

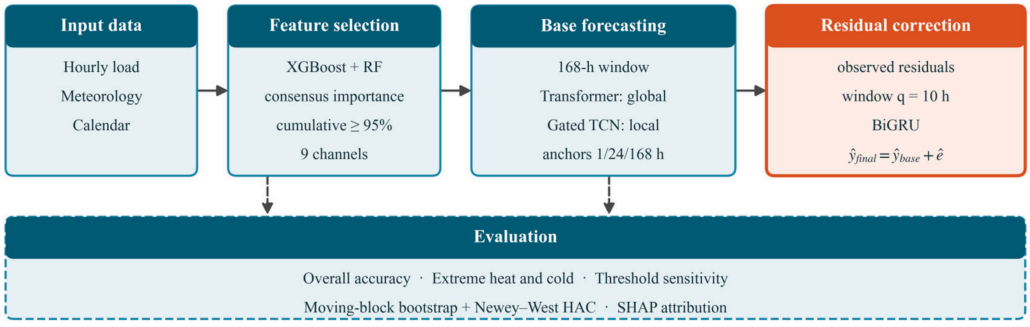


Figure 1. Forecasting framework. Solid arrows indicate the main forecasting process, while dashed arrows indicate connections to the evaluation stage.

3. Empirical Analysis

3.1. Data Source and Preprocessing

To validate the effectiveness of the proposed model, hourly electric load and concurrent meteorological data from 2019 to 2023 for a region in eastern China were analysed. The

sampling interval is 1 h, and the dataset contains 43,824 valid records in total, mainly covering load, temperature, dew point, relative humidity, wind speed, precipitation, solar radiation, and calendar information. During the sample period, the hourly load of this region averaged 58,348.94 MW and ranged from 29,059.75 to 97,718.01 MW, while the concurrent temperature ranged from −18.36 to 38.71 °C; the region is therefore treated as a regional case for the empirical analysis. Descriptive statistics of the main load and meteorological variables are listed in Table 1.

Table 1. Descriptive statistics of the main load and meteorological variables.

Variable	Mean	Std. Dev.	Min.	Max.
Load/MW	58,348.94	9243.72	29,059.75	97,718.01
Temperature/°C	14.74	10.92	−18.36	38.71
Dew point/°C	6.34	11.97	−34.74	27.62
Relative humidity/%	60.93	21.24	8.40	99.64
Wind speed/($m \cdot s^{-1}$)	2.74	1.25	0.38	8.89
Precipitation/mm	0.08	0.37	0.00	7.68
Solar radiation/($W \cdot m^{-2}$)	182.59	255.66	0.00	972.03

In terms of seasonal distribution, both the regional load and temperature show clear seasonal variation. The average load is highest in summer (62,928.83 MW), higher than that in spring (55,198.43 MW); the average summer temperature is 26.78 °C and the average winter temperature is 1.03 °C. The study sample therefore covers load operating states with considerable differences across seasons and temperature conditions, providing a data basis for the subsequent analysis of load forecasting under extreme heat and extreme cold scenarios.

The data were divided strictly in chronological order. The 2019–2020 data were used for input feature selection and the initial training of the Transformer–TCN base model, and the 2021 data were used for base-model validation and determination of the optimal number of training epochs. After model selection, the base model was re-initialized and retrained on the 2019–2021 data to obtain the final Transformer–TCN model. The base-model parameters were then frozen, and the model was used to perform hourly forecasting on the 2022 data; the out-of-sample residual series was constructed from the difference between the actual load and the base forecast for the development of the BiGRU residual model. The 2023 data did not participate in model parameter training or selection and were used only for the final evaluation of rolling 1 h ahead forecasting performance.

Data integrity checks showed that no missing records existed in the load and meteorological data required for modelling, so no missing-value imputation was performed. The Hampel method was used to identify 597 statistical outlier candidates, which were further verified against physical value ranges and temporal continuity; five anomalous observations were finally confirmed. Among the identified candidates, genuine high-temperature, low-temperature, and load-peak processes were retained, and only isolated observations confirmed as measurement or data-acquisition anomalies were corrected. Anomalous observations requiring correction during model development were linearly interpolated, given by:

$$y = y_0 + \frac{y_1 - y_0}{x_1 - x_0} \cdot (x - x_0) \tag{26}$$

where x_0 is the time of the instant preceding the missing point, x_1 is the time of the instant following the missing point, y_0 and y_1 are the known values at the corresponding instants, x is the time point of the missing value, and y is the imputed value. To avoid calling observations after the target instant when linear interpolation is used in the validation, residual-development, and test stages, forward filling was adopted whenever interpolation would require future observations; that is, the previous nearest valid observation was used instead.

Extreme temperature thresholds were determined from the daily maximum and minimum temperature distributions during the 2019–2021 development period. The 90th percentile of the daily maximum temperature was taken as the extreme-heat threshold and the 10th percentile of the daily minimum temperature as the extreme-cold threshold, which are 32.02 °C and −3.49 °C, respectively. After classifying the independent 2023 test set with these thresholds, the Extreme Heat, Extreme Cold, and Normal scenarios contain 345, 504, and 7911 hourly samples, respectively. This definition uses the historical climate distribution of the study region as a reference and is intended to evaluate model performance in the tail states of the temperature distribution.

3.2. Results of XGBoost–RF Input Feature Selection

The importance of input features was evaluated using the XGBoost–RF method, and the screening results are shown in Figure 2. To reduce the contingency of feature evaluation by a single model, the feature importances of the XGBoost model and the random forest (RF) model were calculated separately, normalized, and then evaluated comprehensively. According to the ranking of comprehensive importance and the cumulative contribution rate, seven features were finally selected as the main inputs of the forecasting model: historical load (0.6409), seasonal position (0.1185), holiday (0.0489), temperature (0.0449), time of day (0.0442), dew point (0.0412), and solar radiation (0.0267). Since historical load mainly reflects the temporal autocorrelation of load itself rather than an external influencing factor, its importance is not displayed in Figure 2, and only the remaining external factors are compared. Among the external factors, seasonal position has the highest comprehensive importance, followed by holiday, temperature, time of day, dew point, and solar radiation, indicating that load variations are simultaneously affected by seasonal cycles, daily electricity-use patterns, and meteorological conditions.

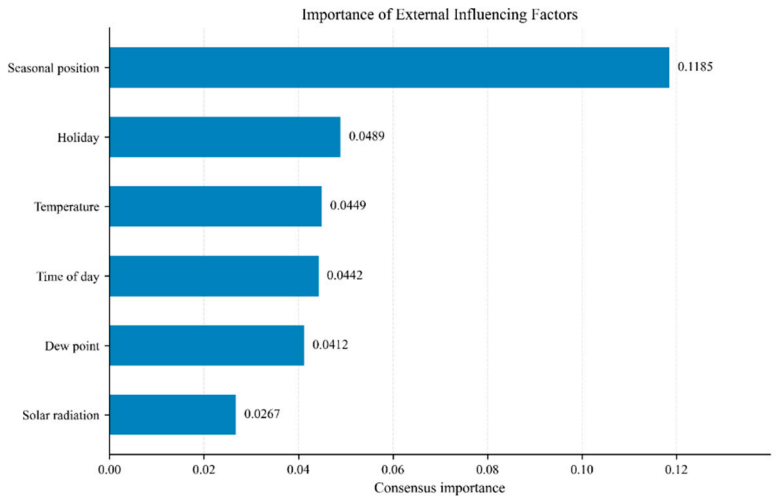


Figure 2. Ranking of XGBoost–RF input feature importance.

To examine the robustness of the cumulative importance threshold, 90%, 95%, and 99% thresholds were further compared. The validation performance obtained with the 90% and 95% thresholds was essentially identical, whereas raising the threshold to 99% introduced additional low-importance features without improving performance. Therefore, the 95% cumulative importance threshold was adopted in this study.

3.3. Forecasting Results of Different Models

To verify the forecasting performance of the proposed model, comparative experiments were conducted with the persistence model, the daily naive model (same hour on the

previous day), the weekly naive model (same hour in the previous week), the temporal convolutional network (TCN) model, the gated recurrent unit (GRU) model, the long short-term memory (LSTM) model, the bidirectional LSTM (BiLSTM) model, the BiLSTM with attention mechanism (BiLSTM-Attention) model, the Transformer model, the PatchTST Model and the iTransformer Model. To further analyse the improvement brought by the residual correction stage, the Transformer–TCN base forecasting model constructed in the first stage of this paper (Transformer–TCN Base) was used as the direct comparison model, and the complete model with BiGRU residual correction was denoted as the Proposed Model. All models adopted the same data partition and were compared under identical evaluation metrics. The overall forecasting results on the independent 2023 test set are listed in Table 2, and the forecasting errors of different models are compared in Figure 3.

Table 2. Comparison of overall forecasting results of different models.

Model	Model Number	MAE/MW	RMSE/MW	MAPE/%	R^2
Persistence	No. 1	1875.97	2444.08	2.857	0.9360
Daily-Naive	No. 2	1623.60	2344.79	2.424	0.9411
Weekly-Naive	No. 3	3496.30	4841.48	5.333	0.7487
TCN	No. 4	1191.38	1969.51	1.683	0.9584
GRU	No. 5	1064.34	1431.01	1.576	0.9780
LSTM	No. 6	1157.50	1535.87	1.724	0.9746
BiLSTM	No. 7	1251.43	1688.12	1.848	0.9693
BiLSTM-Attention	No. 8	1205.36	1652.82	1.780	0.9707
PatchTST	No. 9	1236.18	2079.84	1.733	0.9519
iTransformer	No. 10	1166.93	1877.24	1.621	0.9608
Transformer	No. 11	1015.71	1524.39	1.455	0.9751
Transformer–TCN Base	No. 12	911.79	1337.66	1.334	0.9808
Proposed Model	No. 13	714.31	939.67	1.090	0.9905

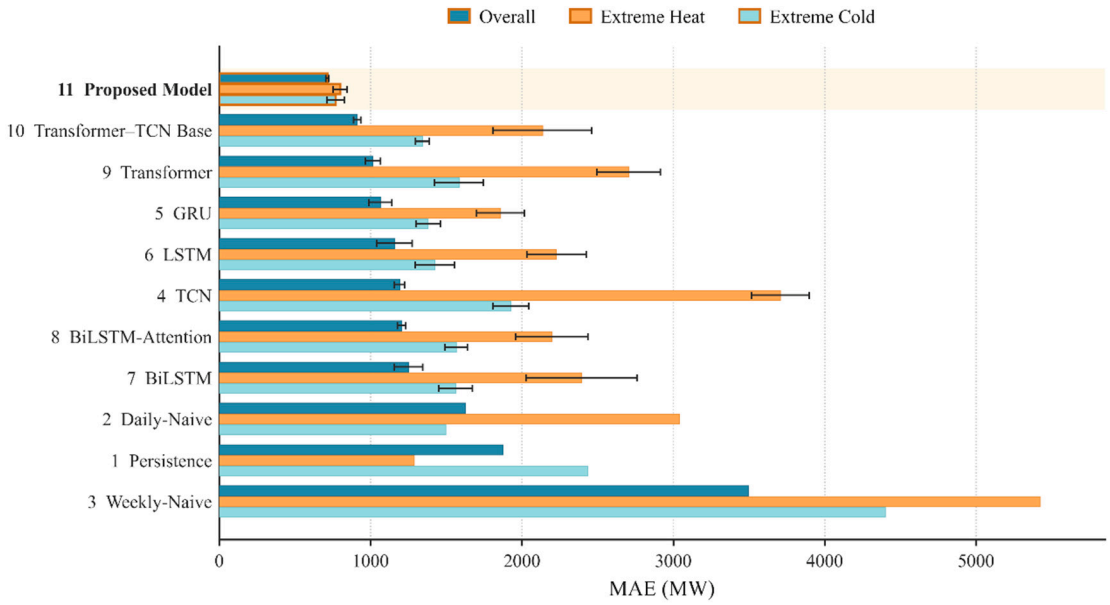


Figure 3. Comparison of forecasting performance of different models.

As shown in Table 2, model No. 13 (the proposed model) achieves the best performance, with MAE, RMSE, and MAPE of 714.31 MW, 939.67 MW, and 1.090%, respectively, and R^2 of 0.9905. The MAE of model No. 11 is 1015.71 MW; after adding the gated TCN local correction, the MAE of model No. 12 decreases to 911.79 MW, and after further adding BiGRU residual correction, the MAE decreases to 714.31 MW. Compared with model No. 12, the MAE and RMSE of model No. 13 are reduced by approximately 21.7% and 29.8%, respectively.

Figure 4 presents the forecast curves for a representative week in 2023. The proposed model tracks the multi-day peak–valley variations well, and around rapid load rises and peak periods the final forecast curve is closer to the actual load overall.

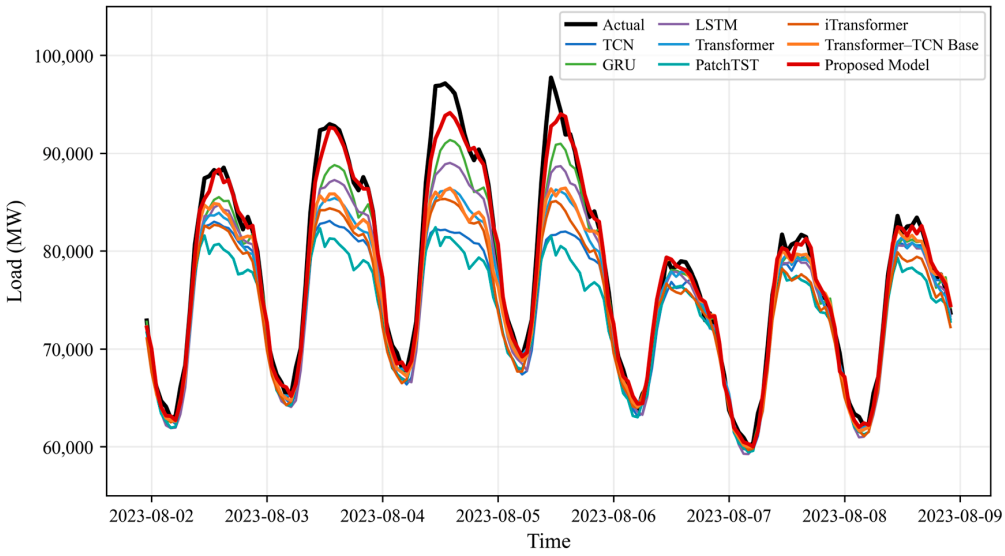


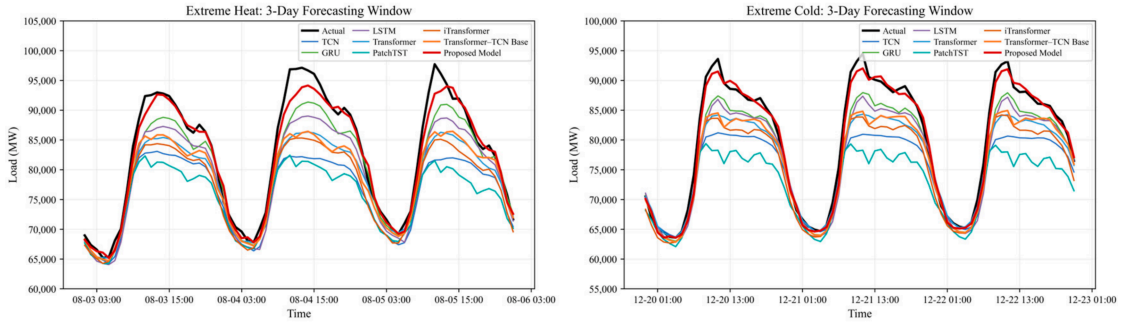
Figure 4. Weekly load forecasting results.

3.4. Forecasting Results Under Extreme Temperature Scenarios

Based on the temperature quantile thresholds determined in the training stage, the 2023 test samples were divided into three states—Normal, Extreme Heat, and Extreme Cold—and the forecasting results of Transformer–TCN Base and the Proposed Model were compared across temperature states. The error metrics under different temperature scenarios are listed in Table 3, and the forecast curves for representative periods under Extreme Heat and Extreme Cold are shown in Figure 5.

Table 3. Forecasting results under different temperature states (average of three independent runs).

Scenario	Model	MAE/MW	RMSE/MW	MAPE/%	R^2
Normal	Transformer–TCN Base	830.98	1147.69	1.253	0.9832
Normal	Proposed Model	707.17	922.08	1.093	0.9892
Extreme Heat	Transformer–TCN Base	2135.63	2947.89	2.467	0.8179
Extreme Heat	Proposed Model	798.74	1093.08	0.957	0.9753
Extreme Cold	Transformer–TCN Base	1342.50	2101.07	1.833	0.9744
Extreme Cold	Proposed Model	768.51	1081.44	1.139	0.9932



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

Figure 5. Forecasting results for representative periods. (**a**) Forecast results under Extreme Heat; (**b**) Forecast results under Extreme Cold.

Under the Normal, Extreme Heat, and Extreme Cold states, the MAE of the proposed model is 707.17, 798.74, and 768.51 MW, respectively, and the corresponding R^2 values are 0.9892, 0.9753, and 0.9932. Together with the overall results, the coefficients of determination of the proposed model reach 99.05%, 97.53%, and 99.32% for the Overall, Extreme Heat, and Extreme Cold scenarios, respectively.

Compared with Transformer–TCN Base, the MAE under Extreme Heat decreases from 2135.63 MW to 798.74 MW (a reduction of approximately 62.6%), and under Extreme Cold from 1342.50 MW to 768.51 MW (approximately 42.8%); under the Normal state it decreases from 830.98 MW to 707.17 MW. Residual correction not only improves the overall average result, but reduces errors more markedly under high- and low-temperature states.

3.5. Ablation Results

Six groups of ablation experiments (A1–A6) were designed to examine successively the effects of TCN local correction, the gating structure, explicit lag anchors, and GRU/BiGRU residual correction on forecasting results. All ablation models adopted the same data partition and evaluation metrics. The error results under the Overall, Extreme Heat, and Extreme Cold scenarios are listed in Table 4, the comparison of overall MAE is shown in Figure 6, and the comparison of ablation forecasts is shown in Figure 7.

Table 4. Error comparison under different scenarios.

Ablation Model	Model Number	Scenario	MAE/MW	RMSE/MW	MAPE/%	R^2
Transformer	A1	Overall	1015.71	1524.39	1.455	0.9751
		Extreme Heat	2703.26	3483.93	3.145	0.7492
		Extreme Cold	1584.53	2492.33	2.121	0.9639
+TCN (no gate)	A2	Overall	944.09	1411.56	1.378	0.9784
		Extreme Heat	2403.50	3324.89	2.776	0.7690
		Extreme Cold	1366.07	2193.06	1.869	0.9716
+Gate (no lag anchors)	A3	Overall	1151.86	1786.60	1.655	0.9657
		Extreme Heat	3252.11	4457.45	3.742	0.5895
		Extreme Cold	1734.25	2840.25	2.296	0.9532
Full Transformer–TCN Base	A4	Overall	911.79	1337.66	1.334	0.9808
		Extreme Heat	2135.63	2947.89	2.467	0.8179
		Extreme Cold	1342.50	2101.07	1.833	0.9744
+GRU residual + DEC	A5	Overall	728.23	966.61	1.106	0.9900
		Extreme Heat	916.97	1297.23	1.087	0.9644
		Extreme Cold	813.02	1163.00	1.190	0.9921
Proposed	A6	Overall	714.31	939.67	1.090	0.9905
		Extreme Heat	798.74	1093.08	0.957	0.9753
		Extreme Cold	768.51	1081.44	1.139	0.9932

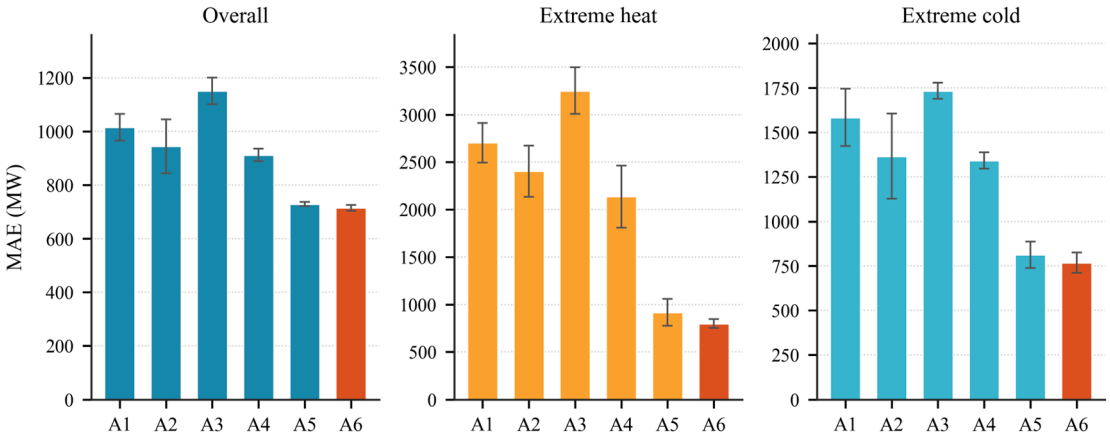


Figure 6. Overall MAE of the A1–A6 ablation models.

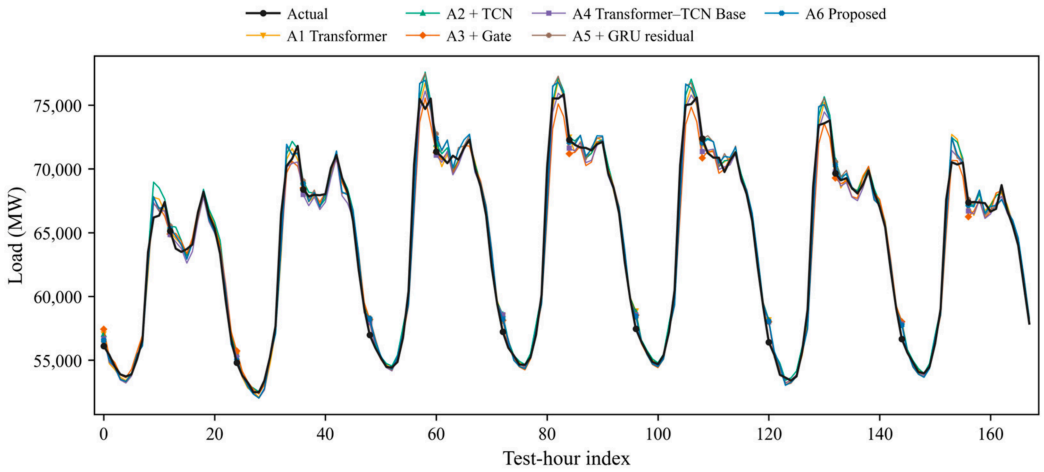


Figure 7. Comparison of ablation forecasts.

A1 performs load forecasting with a Transformer and serves as the basic control model for evaluating the capability of the Transformer to model global temporal dependencies of load, providing a benchmark for subsequent model improvements. A2 introduces a non-gated TCN branch on the basis of A1 to examine the influence of local convolutional feature extraction on forecasting performance. The overall forecasting performance of A3, which uses a gated TCN but does not set explicit lag anchors, is lower than that of A2 and A4; after adding 1-h, 24-h, and 168-h explicit lag anchors, the performance of A4 improves significantly. These results indicate that, under the model configuration used in this study, the introduction of explicit periodic information provides supplementary information to the gated TCN, thereby improving the model's predictive performance. A4 is the complete Transformer–TCN base forecasting model, in which explicit 1 h, 24 h, and 168 h lag anchors are added to the gated TCN to verify the base forecasting performance after fusing global temporal dependencies, local variation features, and multi-scale periodic information; it also serves as the unified benchmark for subsequent residual correction experiments.

In the residual correction stage, A5 introduces a unidirectional GRU on the basis of A4 to model historical forecasting residuals and adopts direct error correction (DEC) to obtain the final forecast, which verifies the corrective effect of explicit residual learning on the systematic forecasting errors of the base model. A6 replaces the unidirectional GRU with a BiGRU to model historical residuals while keeping the base model and the direct error correction scheme unchanged; this is the complete model proposed in this paper, and it is used to further compare bidirectional residual feature extraction with unidirectional residual modeling, thereby verifying the gain of BiGRU residual correction.

The ablation results show that the overall MAE, MAPE, and R^2 of A1 are 1015.71 MW, 1.455%, and 0.9751, respectively. After adding the non-gated TCN, the MAE of A2 decreases to 944.09 MW, the MAPE decreases to 1.378%, and R^2 increases to 0.9784. For the complete Transformer–TCN Base (A4), the MAE and MAPE further decrease to 911.79 MW and 1.334%, and R^2 increases to 0.9808, indicating that the combination of Transformer and TCN can further improve base forecasting performance. In contrast, the overall MAE and MAPE of A3 are 1151.86 MW and 1.655%, respectively, and R^2 decreases to 0.9657, indicating that gated local correction must be analyzed together with explicit periodic anchors, and the performance change of A3 alone cannot be attributed to the gating mechanism itself.

Under the same base model and residual training procedure, the introduction of residual correction further improves model performance. The overall MAE, RMSE, and MAPE of A5 are 728.23 MW, 966.61 MW, and 1.106%, respectively, and R^2 reaches 0.9900. With BiGRU residual correction, the corresponding metrics of A6 further improve to 714.31 MW, 939.67 MW,

1.090%, and 0.9905. Under extreme temperature scenarios, the MAE from A5 to A6 decreases from 916.97 MW to 798.74 MW under Extreme Heat and from 813.02 MW to 768.51 MW under Extreme Cold, while the corresponding R^2 values increase from 0.9644 to 0.9753 and from 0.9921 to 0.9932, respectively. Further examination of the incremental differences between A5 and A6 was conducted using the moving-block bootstrap and Newey–West HAC methods, consistent with Section 3.6. Defining ΔMAE as A6 − A5, the overall ΔMAE values under random seeds 42, 123, and 2024 are −14.21, −11.72, and −15.84 MW, respectively, and the 95% confidence intervals are [−19.60, −8.77] and [−19.42, −5.03], and [−32.65, −1.20] MW, corresponding to HAC *p*-values of 4.718×10^{-15}, 4.809×10^{-7}, and 8.403×10^{-4}, respectively; none of the three confidence intervals crosses zero. These results indicate that residual learning is important for reducing forecasting errors, and that under the same residual correction framework, BiGRU can further improve forecasting performance in overall and extreme temperature scenarios compared with a unidirectional GRU.

3.6. Statistical Significance and Threshold Robustness

To examine whether the error reduction of the proposed model relative to Transformer–TCN Base is stable, moving-block bootstrap and the Newey–West heteroskedasticity- and autocorrelation-consistent (HAC) method were further applied to perform statistical tests on hourly paired errors, and the sensitivity of the results to scenario partitioning was examined by varying the extreme temperature quantile thresholds. The paired error test results are listed in Table 5, and the forecasting errors under different extreme temperature thresholds are listed in Table 6.

Table 5. Paired error tests between the base model and the final model under different random seeds.

Seed	Scenario	ΔMAE/MW	95% Block-Bootstrap CI	HAC *p*-Value
42	Overall	−222.09	[−328.59, −125.67]	9.93×10^{-18}
42	Normal	−129.01	[−190.28, −78.21]	9.99×10^{-15}
42	Extreme Heat	−1677.43	[−2668.17, −897.57]	1.88×10^{-5}
42	Extreme Cold	−687.01	[−1380.18, −212.19]	1.65×10^{-4}
123	Overall	−169.28	[−233.97, −119.17]	1.34×10^{-27}
123	Normal	−111.40	[−145.03, −83.97]	9.93×10^{-30}
123	Extreme Heat	−1003.49	[−1617.05, −513.81]	3.66×10^{-5}
123	Extreme Cold	−506.77	[−995.48, −161.60]	7.93×10^{-5}
2024	Overall	−201.07	[−279.23, −135.08]	5.06×10^{-25}
2024	Normal	−131.01	[−169.53, −94.41]	7.62×10^{-26}
2024	Extreme Heat	−1329.75	[−2167.86, −618.51]	2.26×10^{-5}
2024	Extreme Cold	−528.20	[−1066.70, −92.48]	1.72×10^{-3}

Table 6. Sensitivity to extreme temperature thresholds.

Threshold	Scenario	Sample Size/h	Base MAE/MW	Final MAE/MW
P90/P10	Extreme Heat	345	2135.63	798.74
P90/P10	Extreme Cold	504	1342.50	768.51
P95/P5	Extreme Heat	174	2436.41	818.77
P95/P5	Extreme Cold	324	1432.82	768.54

For all three random seeds, the ΔMAE values under the Overall, Normal, Extreme Heat, and Extreme Cold scenarios are all less than 0, the 95% confidence intervals of the moving-block bootstrap do not cross 0, and the *p*-values of the HAC test all reach the significance level. The two types of tests yield consistent results, indicating that the error reduction of the final model relative to the base model is not caused by a small number of isolated instants and remains stable after accounting for hourly error correlation.

When the extreme threshold is tightened from P90/P10 to P95/P5, the MAE of the final model on the high- and low-temperature subsets remains lower than that of the base

model, indicating that the direction of improvement under extreme temperature scenarios does not depend on a single threshold setting.

3.7. SHAP-Based Interpretability Analysis

This paper uses SHAP to conduct an interpretability analysis of the Transformer–TCN baseline prediction model. The attribution of each factor to the model output is shown in Figure 8, where Figure 8a presents the mean absolute SHAP value and relative contribution of different external factors, and Figure 8b further shows the magnitude of feature values, the sign of SHAP values, and their distribution characteristics for each factor. The feature importance percentages reported here are used to compare the relative contributions among meteorological and other exogenous factors; historical load is excluded from the normalization of these percentages, while the feature-selection results show that historical load remains one of the most important inputs for load forecasting.

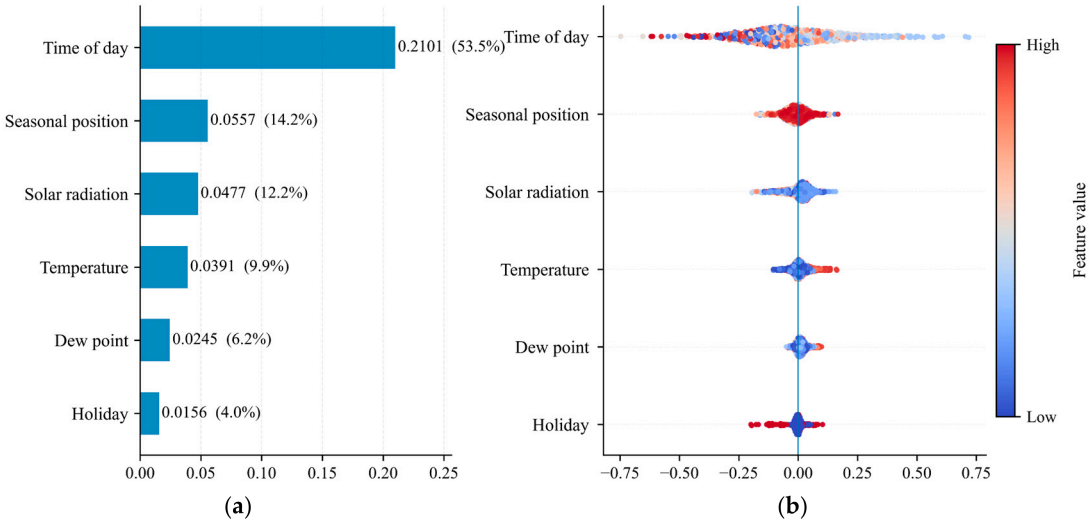


Figure 8. Interpretability analysis results. (**a**) Feature contribution; (**b**) Global interpretability.

Figure 8a shows that among the analysed external factors, time of day has the highest contribution (53.5%), followed by seasonal position, solar radiation, and temperature, with contributions of 14.2%, 12.2%, and 9.9%, respectively; the contributions of dew point and holiday are 6.2% and 4.0%, respectively. The cumulative contribution of the two types of temporal information, time of day and seasonal position, reaches 67.7%, indicating that the model makes considerable use of intraday load variation and seasonal position information in short-term load forecasting. Among the meteorological factors, solar radiation, temperature, and dew point all show certain model attributions, and the mean absolute SHAP value of temperature is 0.0391, which provides a basis for the subsequent analysis of changes in the attribution of temperature information under extreme temperature states. Figure 8b presents the global SHAP distributions of different external factors. The SHAP values of time of day have the widest distribution range, indicating that the model attributions at different times of day differ noticeably, which is consistent with its relatively high overall contribution in Figure 8a. The SHAP values of seasonal position are mainly distributed around 0, but still show certain positive and negative differences across seasonal positions, indicating that the model exploits load variation information at different stages of the year.

The distribution of temperature is relatively clear: higher-temperature samples are mainly distributed in the positive SHAP region, whereas lower-temperature samples are mostly located in the negative region, indicating that in the current model and test samples, high-temperature information generally corresponds to attributions that raise the model

output, while low-temperature information corresponds more to attributions that lower the model output. The attribution strength of temperature was further analysed by temperature state. The mean absolute SHAP values of temperature under the Normal, Extreme Heat, and Extreme Cold states are 0.0243, 0.0540, and 0.0389, respectively. The value under Extreme Heat is approximately 2.22 times that under Normal, and the value under Extreme Cold is also higher than that under Normal. This result indicates that once temperature enters an extreme range, the model relies more strongly on temperature information.

4. Conclusions

To address the large short-term load forecasting errors under extreme temperature conditions, this paper constructed a two-stage forecasting model combining XGBoost–RF factor screening, Transformer–TCN base forecasting, and BiGRU residual correction, and conducted an empirical analysis based on hourly load and meteorological data from 2019 to 2023. The main conclusions are as follows.

(1) Based on the XGBoost–RF model, the factors affecting load forecasting were screened, and seven features were finally retained—historical load, seasonal position, holiday, temperature, time of day, dew point, and solar radiation—with a cumulative contribution above 95%. The screening results show that historical load information is of high importance in short-term load forecasting, while temporal rhythm and meteorological information also provide effective information for forecasting, yielding a more suitable input feature set for subsequent model construction.

(2) A Transformer–TCN base forecasting model was constructed, in which the Transformer extracts global temporal dependency features of the load series and the TCN further extracts local variation information. On the independent 2023 test set, the overall MAE and RMSE of the base model are 911.79 MW and 1337.66 MW, respectively, and the MAE under the Extreme Heat and Extreme Cold scenarios is 2135.63 MW and 1342.50 MW, respectively. Ablation experiments further show that the combination of Transformer and TCN improves base forecasting performance and provides a relatively stable base forecast for subsequent residual correction.

(3) A BiGRU residual correction mechanism was further introduced on the basis of the base forecasting model. On the overall test set, the MAE, RMSE, and MAPE of the final model reach 714.31 MW, 939.67 MW, and 1.09%, respectively, and R^2 reaches 99.05%. Compared with the Transformer–TCN base model, the overall MAE is reduced by approximately 21.7%, and the MAE under the Extreme Heat and Extreme Cold scenarios is reduced by approximately 62.6% and 42.8%, respectively, with corresponding R^2 values of 97.53% and 99.32%. Ablation experiments and statistical tests further indicate that BiGRU residual correction can effectively reduce the forecasting errors of the base model and shows more pronounced error improvement under extreme temperature scenarios.

(4) SHAP interpretability analysis shows that, among external factors excluding the autoregressive information of historical load, time of day has the highest relative contribution (53.5%), while seasonal position, solar radiation, and temperature contribute 14.2%, 12.2%, and 9.9%, respectively. High-temperature samples mainly exhibit positive temperature attributions, whereas lower-temperature samples exhibit more negative attributions. Further scenario-specific analysis shows that the mean absolute SHAP values of temperature under the Normal, Extreme Heat, and Extreme Cold scenarios are 0.0243, 0.0540, and 0.0389, respectively, with the value under Extreme Heat being approximately 2.22 times that under Normal, indicating that the model exhibits a stronger attribution response to temperature information under extreme temperature states.

The method described in this paper is designed for hourly rolling load forecasting. It improves the stability and accuracy of forecast results under conditions of rapid load changes caused by extreme high and low temperatures, and can provide forecasting references for short-term power grid dispatch, reserve capacity allocation, and early warnings regarding supply–demand balance during periods of extreme temperatures. The empirical analysis in this paper is currently based primarily on data from a regional power grid in eastern China; the model's applicability under different climatic conditions, load structures, and grid scales still requires further validation, and the geographical generalization ability of the proposed model has not yet been independently verified. At the same time, the existing interpretability analysis focuses primarily on the contributions of features during the baseline forecasting stage. Future research could further validate the model's cross-regional generalization capability and, by combining it with probabilistic forecasting methods, characterize the uncertainty in load forecasting under extreme weather conditions, with the aim of providing a more comprehensive reference for load forecasting and power grid operation during extreme weather events.

Author Contributions: Conceptualization, S.L. and X.X.; methodology, S.L. and Y.W. (Yating Wang); validation, S.L., Y.W. (Yating Wang) and Y.W. (Yijun Wang); formal analysis, Q.L.; investigation, Y.W. (Yijun Wang), Q.L. and H.H.; resources, Y.W. (Yijun Wang); data curation, S.L. and Y.W. (Yijun Wang); writing—original draft, S.L.; writing—review & editing, Y.W. (Yating Wang), Q.L., H.H., Y.Z. and X.X.; supervision, Y.Z. and X.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this study are not publicly available due to confidentiality restrictions imposed by the data provider.

Conflicts of Interest: Authors Shiyu Liu, Yating Wang, Yijun Wang, Qiang Liu, and Hanyan Huang are employed by the company China Electric Power Planning and Engineering Institute Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Model Parameters and Experimental Setup

Table A1. Model parameters and experimental settings.

Module	Hyperparameter	Value
XGBoost	Number of estimators	800
	Max depth	6
	Learning rate	0.035
	Subsample	0.85
	Column subsample	0.85
Random Forest	Number of estimators	800
	Max features	Sqrt
	Min samples per leaf	2
Transformer	Attention heads	4
	Embedding dimension	64
	Dropout	0.15
	Encoder layers	2

Table A1. *Cont.*

Module	Hyperparameter	Value
TCN	Channels	64
	Kernel size	3
	Dilation factors	1, 2, 4, 8, 16
	Number of blocks	5
	Dropout	0.15
BiGRU	Hidden units	32
	Layers	1
	Learning rate	1×10^{-3}
	Loss function	SmoothL1Loss ($\beta = 0.5$)
	Maximum epochs	100
	Early-stopping patience	12
Training	Optimizer	AdamW
	Learning rate	1×10^{-3}
	Batch size	256
Repeated runs	Random seeds	42, 123, 2024

References

1. Chen, S.; Fu, T.; Lan, H.; Hao, L.; Yang, Y.; Weng, Z. A data-driven framework for unit commitment considering ramping and forecasting information. *Front. Energy Res.* **2025**, *13*, 1693639. [CrossRef]
2. Shiblee, M.F.H.; Koukaras, P. Short-term load forecasting in the Greek power distribution system: A comparative study of gradient boosting and deep learning models. *Energies* **2025**, *18*, 5060. [CrossRef]
3. Perera, A.T.D.; Nik, V.M.; Chen, D.; Scartezzini, J.-L.; Hong, T. Quantifying the impacts of climate change and extreme climate events on energy systems. *Nat. Energy* **2020**, *5*, 150–159. [CrossRef]
4. Auffhammer, M.; Baylis, P.; Hausman, C.H. Climate change is projected to have severe impacts on the frequency and intensity of peak electricity demand across the United States. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 1886–1891. [CrossRef] [PubMed]
5. Van Ruijven, B.J.; De Cian, E.; Sue Wing, I. Amplification of future energy demand growth due to climate change. *Nat. Commun.* **2019**, *10*, 2762. [CrossRef] [PubMed]
6. Maia-silva, D.; Kumar, R.; Nateghi, R. The critical role of humidity in modeling summer electricity demand across the United States. *Nat. Commun.* **2020**, *11*, 1686. [CrossRef] [PubMed]
7. Doss-gollin, J.; Farnham, D.J.; Lall, U.; Modi, V. How unprecedented was the February 2021 Texas cold snap? *Environ. Res. Lett.* **2021**, *16*, 064056. [CrossRef]
8. Busby, J.W.; Baker, K.; Bazilian, M.D.; Gilbert, A.Q.; Grubert, E.; Rai, V.; Rhodes, J.D.; Shidore, S.; Smith, C.A.; Webber, M.E. Cascading risks: Understanding the 2021 winter blackout in Texas. *Energy Res. Soc. Sci.* **2021**, *77*, 102106. [CrossRef]
9. Alipour, P.; Mukherjee, S.; Nateghi, R. Assessing climate sensitivity of peak electricity load for resilient power systems planning and operation: A study applied to the Texas region. *Energy* **2019**, *185*, 1143–1153. [CrossRef]
10. Lahouar, A.; Ben Hadj Slama, J. Day-ahead load forecast using random forest and expert input selection. *Energy Convers. Manag.* **2015**, *103*, 1040–1051. [CrossRef]
11. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 785–794.
12. Kong, W.; Dong, Z.Y.; Jia, Y.; Hill, D.J.; Xu, Y.; Zhang, Y. Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Trans. Smart Grid* **2019**, *10*, 841–851. [CrossRef]
13. Sajjad, M.; Khan, Z.A.; Ullah, A.; Hussain, T.; Ullah, W.; Lee, M.Y.; Baik, S.W. A novel CNN-GRU-based hybrid approach for short-term residential load forecasting. *IEEE Access* **2020**, *8*, 143759–143768. [CrossRef]
14. Lim, B.; Arik, S.Ö.; Loeff, N.; Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [CrossRef]
15. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115. [CrossRef]
16. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 22419–22430.
17. Song, C.; Yang, H.; Cai, J.; Yang, P.; Bao, H.; Xu, K.; Meng, X.-B. Multi-energy load forecasting via hierarchical multi-task learning and spatiotemporal attention. *Appl. Energy* **2024**, *373*, 123788. [CrossRef]

18. Wang, D.; Yue, C.; Elamraoui, A. Multi-step-ahead electricity load forecasting using a novel hybrid architecture with decomposition-based error correction strategy. *Chaos Solitons Fractals* **2021**, *152*, 111453. [CrossRef]
19. Fan, C.; Nie, S.; Xiao, L.; Yi, L.; Li, G. Short-term industrial load forecasting based on error correction and hybrid ensemble learning. *Energy Build.* **2024**, *313*, 114261. [CrossRef]
20. Laouafi, A.; Laouafi, F.; Boukelia, T.E. An adaptive hybrid ensemble with pattern similarity analysis and error correction for short-term load forecasting. *Appl. Energy* **2022**, *322*, 119525. [CrossRef]
21. Xie, K.; Giacomoni, A.; Deng, X.; Wu, Y. Bias calibration and error propagation adjustment for ML-based time series forecasting: A systematic study for PJM's electricity load forecast amid Virginia's data center surge. *Energy* **2025**, *336*, 138411. [CrossRef]
22. Kim, Y.; Choi, Y.; Min, S.K. Future changes in heat wave characteristics and their impacts on the electricity demand in South Korea. *Weather Clim. Extrem.* **2022**, *37*, 100485. [CrossRef]
23. Yang, W.; Sparrow, S.N.; Wallom, D.C.H. A comparative climate-resilient energy design: Wildfire Resilient Load Forecasting Model using multi-factor deep learning methods. *Appl. Energy* **2024**, *368*, 123365. [CrossRef]
24. Xiao, Y.; Hu, X.; Lin, Y.; Lu, Y.; Jing, R.; Zhao, Y. Interpretable short-term electricity load forecasting considering small sample heatwaves. *Appl. Energy* **2025**, *398*, 126417. [CrossRef]
25. Zhang, X.; Wang, J.; Wang, Y.; Wang, S.; Yang, S.; Wang, Y.; Gao, K. ACOLM: Adaptive contrastive online learning model for urban extreme weather load forecasting. *Energy* **2025**, *340*, 139255. [CrossRef]
26. Liu, G.; Liu, J.; Bai, Y.; Wang, C.; Wang, H.; Zhao, H.; Liang, G.; Zhao, J.; Qiu, J. EWELD: A large-scale industrial and commercial load dataset in extreme weather events. *Sci. Data* **2023**, *10*, 615. [CrossRef] [PubMed]
27. Lu, N.; Qin, D.; Zhou, Y.; Bai, L.; Wen, Q.; Kang, C.; Wang, Y. Electricity load forecasting under extreme heat and cold waves. *Nat. Commun.* **2026**, *17*, 10041. [CrossRef] [PubMed]
28. Dong, J.; Jiang, Y.; Chen, P.; Li, J.; Wang, Z.; Han, S. Short-term power load forecasting using bidirectional gated recurrent units-based adaptive stacked autoencoder. *Int. J. Electr. Power Energy Syst.* **2025**, *165*, 110459. [CrossRef]
29. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
30. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*; Curran Associates: Long Beach, CA, USA, 2017; pp. 4765–4774.

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.