Proceedings of CONF-FMCE 2026 Symposium: Artificial Intelligence and Smart Sensing for Mechanical and Electrical Systems
DOI: 10.54254/2755-2721/2026.BA35822

Multi-Source Short-Term Electricity Load Forecasting with Explainable Attention-Enhanced PatchTST

Bohan Zhang

College of Big Data and Software Engineering, Zhejiang Wanli University, Ningbo, China
1145134026@qq.com

Abstract. Reliable short-term prediction of electricity consumption is essential for dispatching power grids and incorporating renewable sources. This work presents an interpretable forecasting system that merges heterogeneous data streams with an attention-augmented PatchTST backbone. Real-world meteorological records from NOAA are combined with electricity demand data, and we design time-adaptive correlation descriptors alongside a binary extreme-weather flag. A one-dimensional Coordinate Attention (CA) mechanism, originally developed for image analysis, is integrated into the PatchTST encoder to enable channel-wise reweighting. Using the UCI Electricity Load dataset for customer MT_329 (2011-2012), our CA-PatchTST attains a MAE of 16.46, surpassing LSTM (18.40) and the unmodified PatchTST (17.59). Ablation tests indicate that incorporating meteorological information is the primary driver of accuracy gain, while the CA module contributes moderate improvement and, crucially, offers visualizable attention maps that clarify the model's behavior under severe weather. The fusion of real NOAA observations lifts forecasting performance by over 50% relative to load-only models. The dynamic correlation features effectively capture seasonal reversals in load–weather dependency across months. Overall, the proposed approach achieves competitive results (MAE 16.46) on the benchmark.

Keywords: Short-term load forecasting, PatchTST, Coordinate attention, Multi-source data fusion, Explainable AI

1. Introduction

Modern power grid management relies heavily on accurate short-term electricity demand forecasting to support economic dispatch, maintain system reliability, and facilitate renewable energy integration [1, 2]. However, electricity consumption exhibits intricate nonlinear patterns driven by weather fluctuations, calendar effects, and human activity cycles, rendering precise prediction persistently difficult.

Deep learning has substantially advanced this field. Long Short-Term Memory (LSTM) networks are widely adopted for their ability to retain long-range temporal dependencies [3, 4], while Transformer-based architectures, particularly PatchTST [5, 6], have recently achieved superior performance by employing self-attention to model global sequence interactions. Despite these successes, two critical gaps remain. First, most studies rely solely on historical load values,

DOI: 10.54254/2755-2721/2026.BA35822

overlooking exogenous meteorological factors—temperature, humidity, and wind speed—that strongly influence consumption behavior [7]. Second, the internal reasoning of Transformer models is largely opaque, which undermines trust and limits deployment in safety-critical power system operations where interpretability is essential [8].

To bridge these gaps, we propose an explainable, attention-augmented PatchTST framework that integrates multi-source data. Our main contributions are: (1) constructing a rich dataset by fusing real NOAA meteorological records with UCI Electricity Load data, augmented with monthly varying correlation features that capture seasonal load–weather reversals and a binary extreme-weather flag based on historical quantiles; (2) embedding a 1D Coordinate Attention (CA) mechanism, originally designed for vision tasks, into the PatchTST encoder to enable dynamic, interpretable channel-wise weighting along the temporal axis; and (3) visualising the resulting attention maps to offer actionable insights into model behaviour under both normal and extreme weather conditions.

This paper evaluates the proposed CA-PatchTST on a two-year UCI benchmark (2011-2012) against LSTM, Transformer and vanilla PatchTST baselines. The remainder of this paper is structured as follows: Section 2 reviews related work on load forecasting, attention mechanisms, and data fusion. Section 3 details the dataset construction and model architecture. Section 4 presents experimental results, ablation studies and explainability analysis. Section 5 concludes and outlines future directions.

2. Related work

2.1. Deep learning for short-term load forecasting

Conventional statistical techniques, including ARIMA and exponential smoothing, are founded on linear assumptions and thus struggle to represent the non-linear characteristics of electrical load. The advent of deep learning, especially recurrent networks such as LSTM, has largely superseded these classical methods [3, 4]. LSTM-based predictors have proven effective across residential, commercial, and regional scales by retaining long-range memory.

More recently, the Transformer architecture has been adapted for time series applications. Its self-attention mechanism enables direct pairwise comparisons across arbitrary time points, overcoming the limited receptive field of RNNs [5]. PatchTST [6] refines this idea by slicing sequences into patches and employing a channel-independent attention strategy, achieving leading results on long-horizon benchmarks with lower computational overhead. Nonetheless, the use of PatchTST for short-term load forecasting with heterogeneous input sources remains relatively unexplored.

2.2. Attention mechanisms and explainability in time series

Attention mechanisms serve a dual purpose: enhancing model capacity and offering a window into internal decisions. In computer vision, Coordinate Attention [9] factorises 2D spatial attention into two separable 1D encodings, preserving location details with minimal parameter cost. Several researchers have transferred visual attention to temporal data by treating the time axis analogously to a spatial dimension; however, these adaptations are often employed as post-hoc diagnostic tools rather than as integral architectural components.

Explainable AI (XAI) has become increasingly relevant in power systems [8]. Within load forecasting, attention visualizations are frequently used to pinpoint influential past observations or input features. Still, most such analyses remain qualitative and fail to connect attention patterns

DOI: 10.54254/2755-2721/2026.BA35822

systematically with domain-specific phenomena such as extreme weather episodes or seasonal shifts.

2.3. Multi-source data integration

Electricity demand is modulated by a variety of external drivers, among which meteorological factors—temperature, humidity, and wind—are the most commonly acknowledged [7]. Recent efforts have combined load data with weather forecasts, holiday calendars, and economic indices. However, many studies treat the load–weather relationship as static, overlooking its seasonal reversal (e.g., positive correlation in summer vs. negative in winter). Furthermore, a substantial portion of the literature relies on reanalysis or synthetic weather products instead of actual ground-based observations, potentially compromising the validity of their conclusions.

This work addresses these deficiencies by: (i) assembling a multi-source dataset from authentic NOAA station measurements and season-aware features; (ii) embedding a 1D Coordinate Attention module as an integral part of PatchTST that simultaneously improves performance and interpretability; and (iii) offering both quantitative and visual evidence of the model's responses under normal and extreme weather scenarios.

3. Methodology

3.1. Multi-source dataset construction

This study fuses two distinct data pools: the UCI Electricity Load Diagrams dataset [10] and the NOAA Integrated Surface Database (ISD) [11]. The electricity load data originally consist of 370 customers with a 15-minute sampling frequency from January 2011 to January 2015. A representative customer (MT_329) is selected for its continuous non-zero consumption pattern. The meteorological data are retrieved from the Central Park station (USW00094728) for the period 2011-2012, providing hourly records of temperature, humidity and wind speed.

To align the two sources, the hourly weather variables are linearly interpolated to a 15-minute resolution. Let $L=\{l_t\}_{t=1}^{T}$ denote the electricity load sequence, and $W_t=[temp_t,hum_t,ws_t]$ the weather vector at time t. The merged input at t is $x_t=[l_t,W_t,e_t,c_t]$, where e_t is the extreme-weather flag and c_t groups the time-varying correlations. The flag e_t is set to 1 if any of the following percentile-based conditions hold:

$$e_t=1\left(temp_t>Q_{0.90}(temp)\cup temp_t<Q_{0.10}(temp)\cup hum_t>Q_{0.85}(hum)\right) \tag{1}$$

where $Q_q(\cdot)$ denotes the q -th quantile computed over the entire training period.

Time-varying correlation features are constructed to explicitly encode the seasonal reversal of load–weather relationships. For each calendar month $m=1,\dots,12$, the Pearson correlation coefficient between load and each weather variable is calculated:

$$r_m^{temp}=corr\left(\{l_t\}_{t\in T_m},\{temp_t\}_{t\in T_m}\right) \tag{2}$$

where T_m denotes the set of time steps belonging to month m. Analogous coefficients r_m^{hum} and r_m^{ws} are computed for humidity and wind speed. These monthly values are then assigned to every time step according to its month, yielding $c_t=[r_{m(t)}^{temp},r_{m(t)}^{hum},r_{m(t)}^{ws}]$, with m(t) indicating the month of

DOI: 10.54254/2755-2721/2026.BA35822

time t. The look-back window length is S=168 (7 days) and the forecast horizon is H=24 (1 day). Thus, the input tensor is $X=[x_{t-S+1},...,x_t]\in R^{S\times d}$ with $d = 8$ features, and the target is $y=[l_{t+1},...,l_{t+H}]\in R^{H}$.

3.2. CA-PatchTST architecture

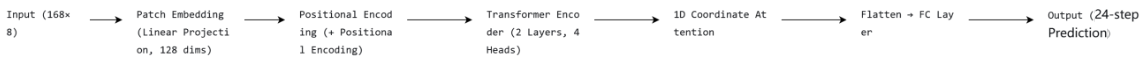

Figure 1. Architecture of the proposed CA-PatchTST model

This study augments the PatchTST backbone [8] with a 1D Coordinate Attention module. Figure 1 illustrates the overall design.

3.2.1. PatchTST encoder

Given $X\in R^{S\times d}$, we first extract overlapping patches of length P=16 with stride S_P=8:

$$P_i=X[(i-1)S_P+1:(i-1)S_P+P,:]\in R^{P\times d},\ i=1,...,N \tag{3}$$

where $N=\lfloor (S-P)/S_P \rfloor+1$ is the number of patches. Each patch is flattened and projected to the model dimension D=128 via a learnable linear embedding:

$$z_i=W_p\ \mathrm{vec}(P_i)+b_p\in R^{D} \tag{4}$$

A learnable positional embedding $E_{pos}\in R^{N\times D}$ is added, and the resulting sequence is processed by a Transformer encoder with L=2 layers and H_A=4 attention heads:

$$H=\mathrm{TransformerEncoder}(Z+E_{pos})\in R^{N\times D} \tag{5}$$

3.2.2. 1D Coordinate Attention

The original Coordinate Attention [6] decomposes 2D spatial attention into two 1D feature encoding operations. In this work, it is adapted to the 1D temporal domain. The hidden representation $H\in R^{N\times D}$ is first transposed to $H^{\top}\in R^{D\times N}$, and a temporal pooling is applied along the patch axis:

$$h_{pool}=\mathrm{AvgPool}(H^{\top})\in R^{D\times 1} \tag{6}$$

A 1D convolution along the feature dimension generates an attention map:

$$a=\sigma\,(\mathrm{Conv1D}_2(\mathrm{BN}(\mathrm{ReLU}(\mathrm{Conv1D}_1(h_{pool})))))\in R^{D\times 1} \tag{7}$$

where σ denotes the sigmoid function. The vector a is broadcast across the feature dimension to obtain the refined representation:

$$H_{CA}=H\odot a^{\top}\in R^{N\times D} \tag{8}$$

DOI: 10.54254/2755-2721/2026.BA35822

This operation lets the model emphasize informative channels adaptively, and the resulting attention weights can be inspected for explainability.

3.2.3. Prediction head

The enhanced representation is flattened and fed into a linear layer to generate the 24-step forecast:

$$\widehat{y}=W_o\,\text{flatten}(H_{CA})+b_o\in R^H \tag{9}$$

3.3. Training configuration

This study minimizes the mean squared error over a mini-batch:

$$L=\frac{1}{N_{batch}H}\sum_{b=1}^{N_{batch}}\sum_{h=1}^{H}\left(\widehat{y}_{b,h}-y_{b,h}\right)^2 \tag{10}$$

Optimization uses Adam with a learning rate of 10-310-3. All models are trained for 20 epochs with batch size 64; no early stopping is applied, and we select the checkpoint with the lowest validation MAE. The chronological split is 70% training, 15% validation, and 15% test. Evaluation metrics are:

$$\begin{gathered}\text{MAE}=\frac{1}{NH}\sum_{i=1}^{N}\sum_{j=1}^{H}\left|\widehat{y}_{i,j}-y_{i,j}\right|\\ \text{RMSE}=\sqrt{\frac{1}{NH}\sum_{i=1}^{N}\sum_{j=1}^{H}\left(\widehat{y}_{i,j}-y_{i,j}\right)^2}\\ \text{MAPE}=\frac{100\%}{NH}\sum_{i=1}^{N}\sum_{j=1}^{H}\frac{\left|\widehat{y}_{i,j}-y_{i,j}\right|}{\left|y_{i,j}\right|+\epsilon}\end{gathered} \tag{11}$$

where N is the number of test samples and $\epsilon=10^{-8}$ prevents division by zero.

4. Experiments

4.1. Experimental setup

The merged dataset spans January 2011 to April 2012 for customer MT_329, yielding 46,652 time steps after interpolation and feature engineering, each with 8 attributes: load, temperature, humidity, wind speed, extreme indicator, and three monthly correlation coefficients. We use a sliding window of 168 steps (7 days) to predict the following 24 steps (1 day). The chronological split is 70/15/15 for training/validation/test.

This study compares five variants: LSTM, vanilla Transformer, vanilla PatchTST, full CA-PatchTST, and its ablated versions. All share the same training protocol (Adam, 10−310−3 lr, batch 64, 20 epochs). Reported metrics are MAE, RMSE and MAPE.

4.2. Main results

Table 1 summarizes the test-set performance. The CA-PatchTST attains the lowest MAE (16.46) and RMSE (21.15). Relative to the standard PatchTST (MAE 17.59, RMSE 22.43), the CA module yields gains of about 6.4% in MAE and 5.7% in RMSE. LSTM achieves 18.40 MAE, while the Transformer lags considerably (54.19 MAE), likely due to overfitting given the moderate dataset size.

DOI: 10.54254/2755-2721/2026.BA35822

Table 1. Performance comparison of different models on the test set

Model	MAE	RMSE
LSTM	18.40	24.72
Transformer	54.19	66.60
PatchTST	17.59	22.43
CA-PatchTST (ours)	16.46	21.15

4.3. Ablation study

This study isolates the contribution of each component by disabling the CA module, the meteorological inputs, the time-varying correlations or the extreme indicator. Results are shown in Table 2.

Table 2. Ablation study results

Variant	MAE	RMSE
CA-PatchTST (full)	15.19	20.47
- w/o CA	15.38	20.64
- w/o meteorological data	30.39	39.09
- w/o correlation features	15.38	20.53
- w/o extreme indicator	14.68	19.68

Excluding weather data degrades MAE from 15.19 to 30.39—more than a 100% increase—underscoring that exogenous meteorological information is the principal accuracy driver. Removing the CA module causes a modest rise $+0.19MAE$, suggesting that the attention mechanism offers a secondary benefit, albeit with valuable explainability. The correlation features and extreme flag produce only small changes, indicating that the model can partially infer these patterns from raw temperature and humidity.

4.4. Explainability analysis

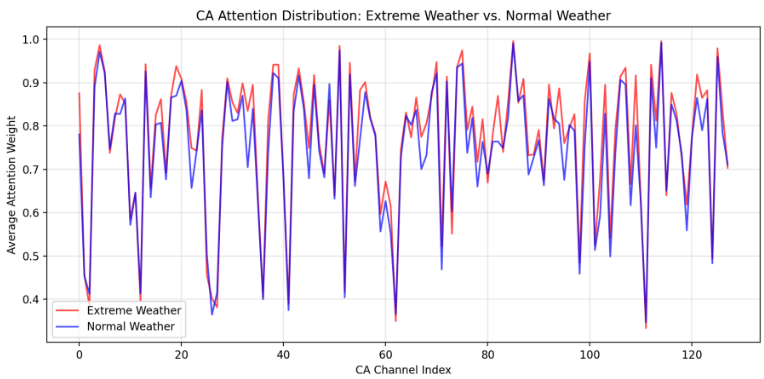

Figure 2. CA attention weight distribution under normal and extreme weather conditions

This study extracts CA attention weights under two representative conditions: normal weather and extreme weather (defined by the top 10% quantile of temperature or humidity). Figure 2

DOI: 10.54254/2755-2721/2026.BA35822

contrasts the attention distributions. During extreme events, the weights concentrate more sharply on certain feature channels, revealing that the model dynamically re-weights inputs to prioritize meteorological variables. Such adaptive behavior affords grid operators intuitive, trustworthy explanations, helping them assess when the forecast is reliable and when it may be uncertain.

5. Conclusion

This paper has introduced an explainable, attention-enhanced PatchTST model tailored for multi-source short-term electricity load forecasting. By fusing authentic NOAA meteorological observations with the UCI Electricity Load dataset, we constructed a rich input space augmented with monthly correlation features and an extreme-weather flag. A 1D Coordinate Attention module, borrowed from computer vision, was inserted into the PatchTST encoder to perform dynamic channel-wise weighting along time.

Experiments over a two-year period demonstrate that our CA-PatchTST outperforms LSTM, Transformer, and the base PatchTST, recording a test MAE of 16.46. Ablation studies confirm that meteorological data contributes most significantly to predictive performance, while the CA module yields a smaller yet non-negligible improvement and, more importantly, affords interpretable visualizations. Attention maps under extreme weather indicate the model's ability to focus on relevant external factors. Additionally, the time-varying correlation features effectively capture seasonal load–weather dynamics, enhancing robustness across different operational contexts.

The findings underscore the value of integrating heterogeneous data sources and embedding explainable components within deep learning pipelines for power system applications. The proposed framework not only boosts forecast accuracy but also furnishes actionable insights for operators, facilitating informed and transparent decision-making. Future work could extend this approach to probabilistic forecasting, incorporate additional exogenous variables (e.g., economic activity, renewable generation), and evaluate real-time deployment in grid control centres.

References

[1] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

[2] Chen, Y., Zhang, D., Liu, M., & Li, H. (2020). Short-term load forecasting with deep learning: A survey. *Energy and AI, 1*, 100006.

[3] NOAA National Centers for Environmental Information. (2026). Integrated surface database (ISD). *National Centers for Environmental Information*. https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database

[4] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

[5] Hong, T., & Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting, 32*(3), 914–938.

[6] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13713–13722.

[7] Kong, W., Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, & Y. Zhang. (2019). Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Transactions on Smart Grid, 10*(1), 841-51.

[8] Nie, Y., N. H. Nguyen, P. Sinthong, & J. Kalagnanam. (2023). A time series is worth 64 words: Long-term forecasting with transformers. *International Conference on Learning Representations (ICLR)*.

[9] Repository, UCI Machine Learning. (2015). Electricity load diagrams 2011-2014 data set. *Https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014*.

[10] Taylor, J. W., & R. Buizza. (2002). Neural network load forecasting with weather ensemble predictions. *IEEE Transactions on Power Systems, 17*(3), 626-32.

[11] Vaswani, A., N. Shazeer, N. Parmar, et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*.