# WEATHER-CONDITIONED ATTENTION CNN-LSTM FOR DAY-AHEAD LOAD FORECASTING: A LEAKAGE-FREE COMPARISON WITH FEED-FORWARD BENCHMARKS

**Abdulla Mohannad Habbib[1*], Omar Gheyath Ibrahim[2], Mohammed Tariq Majeed[3]**

*[1)], [2)] Department of Medical Instrumentation Techniques Engineering, Polytechnic College Mosul, Northern Technical University, Mosul, Iraq, [3)] Department of Medical Instrumentation Techniques Engineering, Polytechnic College Mosul, Northern Technical University, Mosul, Iraq.*

*Email: [1*)]abdulla1985@ntu.edu.iq, [2)] mti.lec17.omar@ntu.edu.iq, [3)] aldabgh68@ntu.edu.iq*

***Article Info***

*Article history:*
*Received Oct 05, 2026*
*Revised Oct 15, 2026*
*Accepted Oct 28, 2026*

***ABSTRACT***

*Attention-based deep networks are increasingly proposed for day-ahead load forecasting in smart grids. This study develops a CNN-LSTM forecaster whose temporal attention query is conditioned on the calendar type and hourly temperature of the target day, and it evaluates the model under a protocol designed to avoid three common biases: target leakage across data splits, unequal inputs for benchmarks, and comparison of seed ensembles with single runs. Hourly PJM East load and temperature data for 2013–2018 were split chronologically with purging, and the model was compared with persistence, ridge regression, a multilayer perceptron (MLP), an LSTM and a CNN-LSTM, each stochastic model being trained with three random seeds. With temperature inputs, the proposed model reached a test MAPE of 2.21%, was the most accurate recurrent model, and its attention layer reduced the error of the CNN-LSTM by 20.8% ($p < 0.001$). However, an MLP given the same information was equally accurate in single runs, significantly more accurate as a seed ensemble (2.13%, $p = 0.018$) and more robust to temperature-forecast errors, and an exploratory combination of both models reached 2.02%. Deep hybrid forecasters should therefore be judged against well-configured feed-forward benchmarks under symmetric, leakage-free protocols.*

***Keywords:*** *Short-term load forecasting; Attention mechanism; LSTM; Temperature: Data leakage; Benchmarking.*

This is an open access article under the CC BY-SA License

***ABSTRAK***

*Jaringan deep learning berbasis mekanisme attention semakin banyak dikembangkan untuk memprediksi beban listrik pada hari berikutnya pada jaringan listrik cerdas (smart grid). Penelitian ini mengembangkan model CNN-LSTM dengan temporal attention yang menggunakan jenis hari berdasarkan kalender dan suhu per jam pada hari yang diramalkan sebagai pengondisi kueri attention. Evaluasi dirancang untuk menghindari tiga sumber bias, yaitu kebocoran informasi target antarpemisahan data (target leakage), ketidaksetaraan masukan model pembanding, serta perbandingan tidak setara antara model tunggal dan ansambel dari beberapa random seed. Data beban listrik dan suhu per jam wilayah PJM East periode 2013–2018 dibagi secara kronologis dengan purging. Model dibandingkan dengan persistensi, regresi ridge, multilayer perceptron (MLP), LSTM, dan CNN-LSTM. Setiap model stokastik dilatih menggunakan tiga random seed. Dengan masukan suhu, model usulan mencapai MAPE 2,21% pada pengujian dan menjadi model rekuren terbaik. Mekanisme attention menurunkan galat CNN-LSTM sebesar 20,8% ($p < 0,001$). Namun, MLP menunjukkan akurasi sebanding pada pelatihan tunggal, kinerja ansambel lebih baik dengan MAPE 2,13% ($p = 0,018$), serta ketahanan lebih tinggi terhadap kesalahan prakiraan suhu. Kombinasi eksploratif kedua model menghasilkan MAPE 2,02%. Temuan ini menegaskan perlunya mengevaluasi model hibrida deep learning terhadap model feed-forward yang dikonfigurasi optimal melalui protokol evaluasi setara dan bebas kebocoran informasi.*

***Kata kunci:*** *peramalan beban listrik jangka pendek; mekanisme attention; LSTM; suhu; kebocoran data; evaluasi komparatif model.*

***Corresponding Author:***
*Abdulla Mohannad Habbib,*
*Department of Medical Instrumentation Techniques Engineering,*
*Polytechnic College Mosul, Northern Technical University,*
*Mosul, Iraq.*
*Email: abdulla1985@ntu.edu.iq*

## 1. INTRODUCTION

Electricity demand must be balanced by generation at every instant, so system operators commit units, procure reserves and trade in day-ahead markets on the basis of a load forecast. Short-term load forecasting (STLF), with lead times from one hour to about one week, is therefore a core function of power-system operation, and its importance has grown with the deployment of smart grids, in which distributed generation, electric vehicles and demand response make demand more variable while smart meters and supervisory systems provide long, high-resolution histories [1]–[3]. Load forecasting has consequently become a meeting point of electrical engineering, which supplies the physical understanding of demand, and computer engineering, which supplies the learning algorithms and the computing platforms on which forecasts are produced [4].

Weather, and temperature in particular, is the dominant exogenous driver of demand. Heating and air-conditioning make load a V-shaped function of temperature, and the selection and representation of temperature variables strongly affect STLF accuracy [5]. Models that use future temperatures consistently outperform models without them, but the gain shrinks considerably when observed temperatures are replaced by actual weather forecasts [6]. Deep recurrent networks have become a standard tool for capturing these non-linear, multi-seasonal dependencies: long short-term memory (LSTM) networks have been deployed for grid operators [7] and compared with support vector machines across time scales [8], hybrid convolutional–recurrent (CNN-LSTM) models extract local load shapes before modelling their evolution [9], and attention mechanisms have been added to weight the most relevant past time steps [10], [11]. Transformer-based load forecasters extend this line of work [12], [13], and attention weights are frequently presented as a source of interpretability, although their behaviour can depend strongly on the encoding length [14].

At the same time, the time-series forecasting community has questioned whether such complexity pays off. Simple linear models outperformed elaborate Transformers on standard long-horizon benchmarks [15], and multilayer-perceptron (MLP) architectures have matched or exceeded attention-based models at a fraction of the computational cost [16]–[18]. Forecasting reviews stress that strong, simple benchmarks and sound evaluation are indispensable [19]. Hewamalage et al. catalogue evaluation pitfalls in forecasting, including leakage between training and test data and inappropriate benchmarks [20], and Kapoor and Narayanan show that data leakage has affected hundreds of machine-learning-based studies [21]. Ensembles of networks trained with different random initialisations are known to improve accuracy [22], [23], which implies that comparing an ensemble of one model with a single run of another is not a like-for-like comparison.

Most published attention-based STLF models are evaluated against other deep networks and rarely against a well-configured MLP under identical conditions; when they are, benchmarks often receive less input information, splits defined only by the forecast origin let multi-hour target windows cross the boundary between training, validation and test periods, and stochastic models are reported as single runs or with asymmetric ensembling. It also remains unclear under which inputs an attention layer actually helps. The objective of this study is therefore to design a CNN-LSTM forecaster whose temporal attention query is conditioned on the calendar type and hourly temperature of the target day, to quantify by ablation when this attention is useful, and to compare the model with persistence, ridge regression, an MLP, an LSTM and a CNN-LSTM under a protocol that removes these three biases: a purged chronological split, identical input information for all models, and identical early stopping and seed treatment with both single-run and ensemble results. The novelty lies in this combination of a weather-conditioned attention query with a leakage-free, seed-controlled benchmark, which reveals that the attention layer is what makes recurrent models competitive when temperature is available, but that a feed-forward network remains more accurate.

## 2. RESEARCH METHODS

### 2.1. Data and Purged Chronological Split

The study uses the hourly metered load of the PJM East region of PJM Interconnection, a regional transmission organisation in the eastern United States [24], [25]. The period from 1 January 2013 to 2 August 2018 (48,960 hours) was retained; duplicate time stamps at daylight-saving transitions were averaged and 7 missing hours were interpolated. Hourly 2-m air temperatures were obtained from the Open-Meteo historical archive [26] for Philadelphia, Newark, Baltimore and Washington, D.C., and their mean was used as a zonal index; the archive was sampled every three hours and interpolated to hourly resolution with a

cubic spline, which is accurate for the smooth diurnal course of air temperature. The correlation between daily load and temperature was 0.92 on days above 18 °C and -0.89 below, so temperature was represented by cooling- and heating-degree terms with an 18 °C base:

$$\mathrm{CD}_t = \max(T_t - 18,\ 0), \qquad \mathrm{HD}_t = \max(18 - T_t,\ 0) \tag{1}$$

The data were split chronologically into training (2013–2016), validation (2017) and test (1 January–2 August 2018) periods (Figure 1). Because every sample has a 24-hour target window, a split by forecast origin alone lets the last 23 windows of one period reach into the next; in a preliminary version of this study, 23 test-period hours were thereby used as validation targets. The split was therefore purged: a sample with origin $o$ belongs to a subset only if its whole target window $[o,\ o + 23]$ lies inside that subset. This yields the sample counts in Table 1, and a script verified that the target hours of the three subsets are disjoint. The 168-hour input windows of the first validation and test samples still contain earlier observations, which is legitimate because these values are known at the forecast origin. The load was standardised with the mean and standard deviation of the training period only.

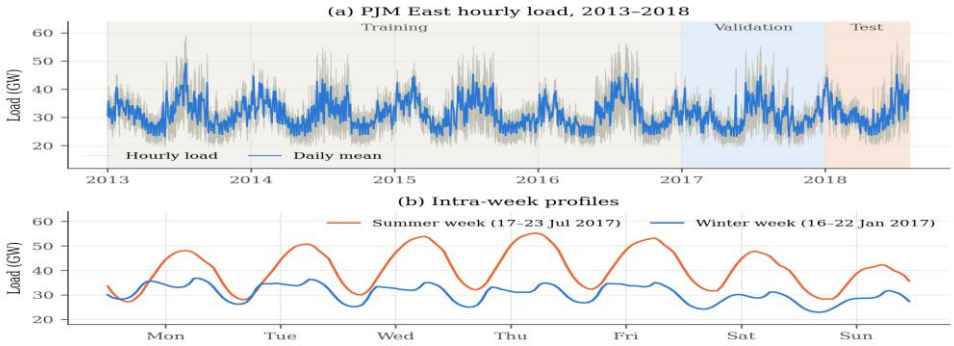



**Figure 1** Hourly PJM East load with the training, validation and test periods (a), and representative summer and winter weeks (b)

**Table 1** Data subsets: number of forecast samples after purging, and mean and maximum hourly load of each period

| Subset | Period | Samples | Mean load (MW) | Max. load (MW) |
|---|---|---|---|---|
| Training | Jan 2013 – Dec 2016 | 35,041 | 31,562 | 59,122 |
| Validation | Jan – Dec 2017 | 8,737 | 30,652 | 55,218 |
| Test | 1 Jan – 2 Aug 2018 | 5,113 | 31,781 | 56,609 |

## 2.2. Problem Formulation and Input Settings

Each past hour is described by a vector $\mathbf{x}_t$ containing the standardised load, sine–cosine encodings of the hour of day, day of week and day of year, weekend and holiday indicators and, with temperature, the standardised temperature and the scaled degree terms. A target-day vector $\mathbf{u}$ contains the calendar type of the day to be forecast and, with temperature, the degree terms of its 24 hours. Observed temperatures are used as a proxy for a perfect forecast; the effect of forecast errors is quantified separately. At every hourly origin $o$, the model maps the previous week to the next 24 loads:

$$\hat{\mathbf{y}}_o = [\hat{y}_o, \ldots, \hat{y}_{o+23}] = f_\theta(\mathbf{x}_{o-168}, \ldots, \mathbf{x}_{o-1}, \mathbf{u}_o) \tag{2}$$

Two input settings were evaluated with otherwise identical models: Setting A uses load and calendar features only (9 features per hour, $\mathbf{u}$ of dimension 4), and Setting B adds temperature (12 features per hour, $\mathbf{u}$ of dimension 52).

## 2.3. Proposed CNN-LSTM with Conditioned Attention

The architecture (Figure 2) has four stages. Two one-dimensional convolutional layers (32 filters, kernel 3, ReLU) learn local shape detectors such as morning ramps and evening peaks, and max-pooling halves the sequence to 84 two-hour steps. An LSTM with 64 units then processes the pooled sequence $\mathbf{z}_1, \ldots, \mathbf{z}_{84}$; with input, forget and output gates $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{o}_t$ and candidate $\mathbf{g}_t$, its cell and hidden states evolve as

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t, \qquad \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \tag{3}$$

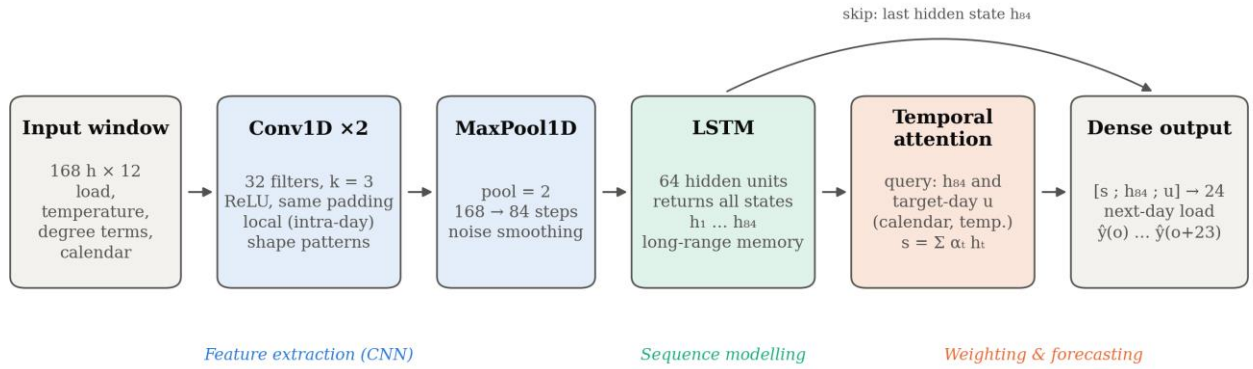



**Figure 2** Architecture of the proposed CNN-LSTM with weather- and calendar-conditioned temporal attention (Setting B dimensions)

Instead of using only the last hidden state, the model scores every hidden state with an additive attention whose query combines the final state $\mathbf{h}_{84}$ with the target-day vector $\mathbf{u}$:

$$e_t = \mathbf{v}_a^{\top}\tanh(\mathbf{W}_a\mathbf{h}_t + \mathbf{U}_a\mathbf{h}_{84} + \mathbf{V}_a\mathbf{u} + \mathbf{b}_a) \quad (4)$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{\tau=1}^{84}\exp(e_\tau)}, \qquad \mathbf{s} = \sum_{t=1}^{84}\alpha_t\,\mathbf{h}_t \quad (5)$$

The conditioning mirrors the similar-day reasoning of dispatchers, who look for recent days of the same type and temperature, and, because $\mathbf{u}$ enters through the tanh non-linearity, it lets the network combine target-day weather non-linearly with the history. After dropout (rate 0.2), the context vector, the final state and $\mathbf{u}$ are mapped linearly to the 24 outputs:

$$\hat{\mathbf{y}} = \mathbf{W}_y[\mathbf{s}\,;\,\mathbf{h}_{84}\,;\,\mathbf{u}] + \mathbf{b}_y \quad (6)$$

The network minimises the mean squared error over all horizons and has 39,288 parameters in Setting B. All layers, including back-propagation through time and attention, were implemented from first principles in NumPy, and every analytical gradient was verified against finite differences (maximum relative discrepancy below $10^{-4}$). Table 2 lists the training configuration, which was identical for all recurrent models.

**Table 2** Training configuration of the recurrent models

| Item | Setting |
|---|---|
| Window / horizon | 168 h input, 24 h multi-output |
| Convolution | 2 × Conv1D (32 filters, kernel 3, ReLU), max-pool 2 |
| LSTM / attention | 64 hidden units / additive, dimension 32 |
| Regularisation | dropout 0.2; decoupled weight decay $10^{-4}$; clipping 1.0 |
| Optimiser | Adam, learning rate $10^{-3}$, batch 64, halved after 3 stagnant epochs |
| Stopping | patience 8, max. 40 epochs, best weights restored |
| Seeds | 42, 7, 123 (recurrent models and MLP) |

### 2.4. Benchmarks and Fair Comparison Protocol

Six benchmarks were used: day-ahead and week-ahead persistence; ridge regression on the 168 lagged loads, the calendar features and $\mathbf{u}$; an MLP with two hidden layers (128 and 64 ReLU units) trained on the same inputs with Adam (learning rate $10^{-3}$, mini-batches of 64) and the same chronological early-stopping rule as the recurrent models, i.e., the loss on the 2017 validation set with a patience of 10 epochs and restoration of the best weights; an LSTM reading the raw 168-step sequence; and a CNN-LSTM without attention. The last two form an ablation sequence that isolates the convolutional stage and the attention layer. In Setting B, the flat models were given the temperature terms either for all 168 input hours, i.e., exactly the information of the recurrent models, or for the last 24 hours only; for each model, the variant with the lower validation loss was retained (168 h for ridge regression with $\alpha = 0.01$, 24 h for the MLP with 48,344 parameters), and both variants are reported. The MLP and the three recurrent models were each trained with three random seeds (42, 7 and 123). Every stochastic model is reported

both as the mean ± standard deviation of single runs and as a seed ensemble, i.e., the average of the three forecasts, formed identically for all models.

### 2.5. Evaluation Metrics and Significance Test

Accuracy over all test origins and horizons was measured by the mean absolute error (MAE), the root mean squared error (RMSE), the coefficient of determination $R^2$ and the mean absolute percentage error:

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

Differences in accuracy were tested with the Diebold–Mariano (DM) test [19], [20] on the 214 day-ahead forecasts issued at 00:00, using the daily mean squared-error differential $d_d$ between a benchmark and the proposed model and a Newey–West variance with $K = 7$ days:

$$\text{DM} = \frac{\bar{d}}{\sqrt{\left[\hat{\gamma}(0) + 2\sum_{k=1}^{K}\left(1 - \frac{k}{K+1}\right)\hat{\gamma}(k)\right]/D}} \tag{8}$$

Here $\hat{\gamma}(k)$ is the autocovariance of $d_d$ at lag $k$ and $D$ the number of test days; positive values favour the proposed model. Because six benchmarks are compared with the proposed model, all conclusions were also checked with the Holm correction for multiple comparisons, which changed none of them.

## 3. RESULTS AND DISCUSSION

### 3.1. Research Results

**Overall accuracy.** Table 3 reports the test accuracy in Setting B. The MLP was the most accurate model on every metric (MAPE 2.13%, RMSE 935 MW). The proposed model was second (2.21%, 956 MW) and the most accurate recurrent model by a wide margin: the LSTM and CNN-LSTM reached only 2.82% and 2.79%. Relative to day-ahead persistence and ridge regression, the proposed model reduced the MAPE by 72% and 25%, and the DM tests show that it is significantly more accurate than every benchmark except the MLP, which is in turn significantly more accurate than the proposed model (DM = −2.37, p = 0.018). Without attention, the recurrent models were even less accurate than the MLP because the target-day temperature enters them only through the linear output layer, which cannot represent its interaction with the hour of day and the day type.

**Table 3** Test accuracy in Setting B (calendar and temperature inputs; 5,113 hourly origins × 24 horizons). † Seed ensemble of three runs. Best values in bold. DM statistics from day-ahead forecasts (D = 214); positive values favour the proposed model

| Model | MAE (MW) | RMSE (MW) | MAPE (%) | $R^2$ | Parameters | DM vs. proposed | p-value |
|---|---|---|---|---|---|---|---|
| Naive (previous day) | 2,532 | 3,401 | 7.92 | 0.705 | 0 | 8.01 | < 0.001 |
| Naive (previous week) | 3,609 | 4,930 | 11.09 | 0.380 | 0 | 5.11 | < 0.001 |
| Ridge regression | 935 | 1,254 | 2.95 | 0.960 | 17,592 | 7.31 | < 0.001 |
| MLP † | **689** | **935** | **2.13** | **0.978** | 48,344 | −2.37 | 0.018 |
| LSTM † | 899 | 1,203 | 2.82 | 0.963 | 22,520 | 7.75 | < 0.001 |
| CNN-LSTM † | 893 | 1,197 | 2.79 | 0.963 | 31,928 | 7.58 | < 0.001 |
| CNN-LSTM-Attention (proposed) † | 710 | 956 | 2.21 | 0.977 | 39,288 | – | – |

**Single runs, ensembles and input variants.** Table 4 shows that single runs of the MLP and the proposed model were equally accurate (2.36 ± 0.01% against 2.35 ± 0.02%), so the advantage of the MLP arises from ensembling: averaging the three runs lowered its MAPE by 0.23 points, but that of the recurrent models by only 0.06–0.14 points, because the individual MLP runs differ more from each other. This asymmetry matters: in the preliminary version of this study, in which only the recurrent models were ensembled and the MLP was stopped on a random part of the training data, the proposed model appeared more accurate than the MLP, a conclusion that does not survive the symmetric protocol. With the full 168-hour temperature history, i.e., exactly the information of the recurrent models, the MLP reached 2.39 ± 0.07% for single runs and 2.12% as an ensemble, so

the choice of variant does not affect the conclusion. On the validation set, the proposed model even had a slightly lower loss than the selected MLP (0.0194 versus 0.0204), so the test result is not an artefact of model selection. Removing the leakage itself changed the test MAPE of the proposed model only from 2.22% to 2.21%.

**Table 4** Single-run and ensemble accuracy in Setting B. * Variant selected on the validation set. Single runs: mean ± standard deviation over three seeds

| Model | Temp. history | Validation MSE | Single run MAPE (%) | Ensemble MAPE (%) |
|---|---|---|---|---|
| Ridge regression * | 168 h | 0.0297 | – | 2.95 |
| Ridge regression | 24 h | 0.0308 | – | 3.09 |
| MLP | 168 h | 0.0213 | 2.39 ± 0.07 | 2.12 |
| MLP * | 24 h | 0.0204 | 2.36 ± 0.01 | 2.13 |
| LSTM | 168 h | – | 2.88 ± 0.01 | 2.82 |
| CNN-LSTM | 168 h | – | 2.88 ± 0.05 | 2.79 |
| Proposed | 168 h | 0.0194 | 2.35 ± 0.02 | 2.21 |

**Accuracy by horizon and visual assessment.** Ridge regression is the most accurate model for the first two hours (MAPE 0.56% one hour ahead), the MLP for hours 3–11 and the proposed model for hours 12–23 (Figure 3). The gap in favour of the MLP is largest at short horizons, where the recurrent model must compress the latest observations into its state, whereas for the afternoon and evening hours of the next day, which matter most for day-ahead scheduling, the proposed model is slightly more accurate (by 0.01–0.09 points in hours 17–22). Figure 4 shows that both models follow the daily double peak of a winter week with a cold spell and the heat-driven pattern of a summer week far better than persistence; both over-forecast the morning of New Year's Day and under-forecast the heat peak of 5 July, and their errors are strongly correlated (0.80). The MLP was the most accurate model in five of the seven test months and for the daily peak (peak MAPE 2.23% versus 2.51%; timing error 1.08 h versus 1.73 h).

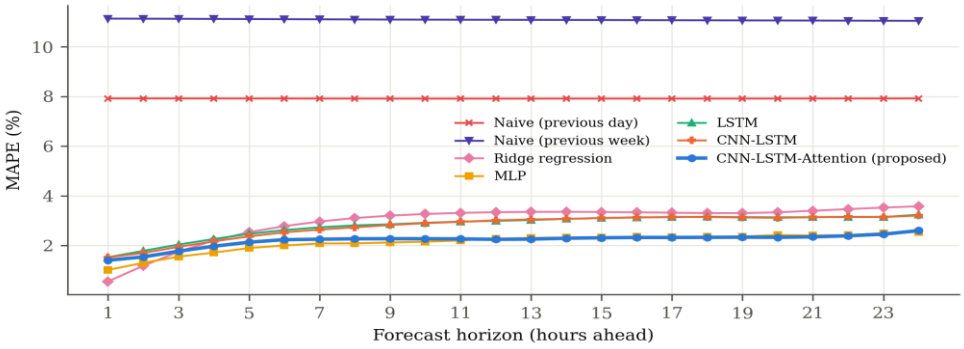



**Figure 3** Test MAPE as a function of the forecast horizon in Setting B (seed ensembles)

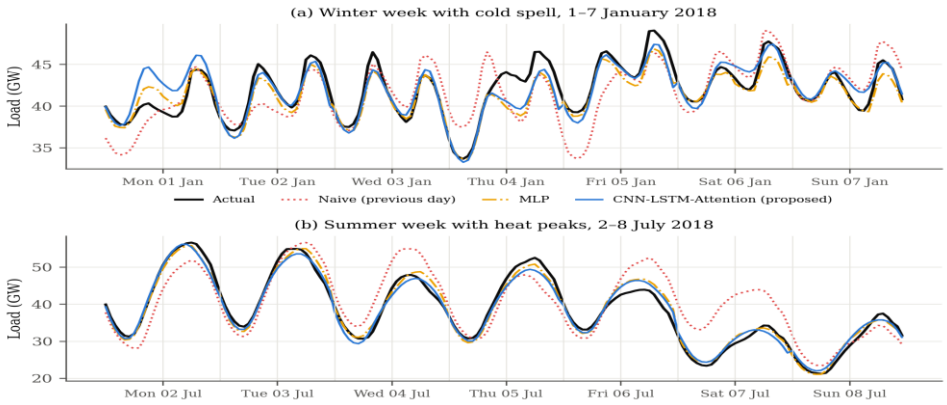



**Figure 4** Day-ahead forecasts issued at 00:00 for (a) a winter week and (b) a summer week of the test period (Setting B)

**Effect of temperature and ablation of attention.** Adding temperature reduced the MAPE of every learned model by 40–55% (Figure 5, Table 5). Without temperature, the three recurrent architectures performed almost identically (4.94%, 4.91% and 4.89%), and attention brought no significant gain over the CNN-LSTM ($p = 0.270$). The MLP had the lowest MAPE in this setting (4.76%) but not the lowest squared error (RMSE 2,194 versus 2,180 MW for the proposed model), and the DM test found no significant difference ($p = 0.267$). With temperature, the convolutional stage again brought little (2.82% versus 2.79%), but the attention layer reduced the MAPE of the CNN-LSTM from 2.79% to 2.21%, a highly significant reduction of 20.8% (18.6% for single runs). The MLP and the proposed model, which both let target-day temperature interact non-linearly with the history, gained about 55% from temperature, whereas the attention-free recurrent models gained only about 43%.

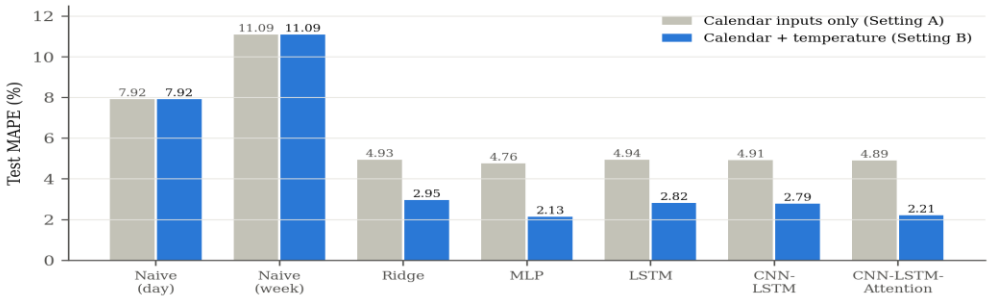



**Figure 5** Test MAPE of all models without (Setting A) and with (Setting B) temperature inputs

**Table 5** Test MAPE in both input settings with DM p-values against the proposed model. ‡ Difference favours the MLP

| Model | MAPE A (%) | p (A) | MAPE B (%) | p (B) |
|---|---|---|---|---|
| Ridge regression | 4.93 | < 0.001 | 2.95 | < 0.001 |
| MLP | 4.76 | 0.267 | 2.13 | 0.018 ‡ |
| LSTM | 4.94 | 0.028 | 2.82 | < 0.001 |
| CNN-LSTM | 4.91 | 0.270 | 2.79 | < 0.001 |
| Proposed | 4.89 | – | 2.21 | – |

**Sensitivity to temperature-forecast errors.** The target-day temperatures of the test set were perturbed by a day-level bias with standard deviation $s$ and an hourly component with standard deviation $s/3$, without retraining (Table 6). All temperature-aware models deteriorate, and the lead of the MLP over the proposed model widens from 0.08 points at $s = 0$ °C to 0.47 points at $s = 3$ °C. The proposed model remains second up to $s = 2$ °C, the range typical of day-ahead temperature forecasts for large cities, but at 3 °C it falls behind both attention-free recurrent models (5.33% versus 5.27% and 5.32%). This reflects its strong reliance on the target-day temperature through the attention query and agrees with the loss of benefit observed with real weather forecasts in [6].

**Table 6** Test MAPE (%) of Setting B models with simulated temperature-forecast errors of day-level standard deviation s

| Model | s = 0 °C | s = 1 °C | s = 2 °C | s = 3 °C |
|---|---|---|---|---|
| Ridge regression | 2.95 | 3.54 | 4.72 | 6.14 |
| MLP | **2.13** | **2.65** | **3.68** | **4.86** |
| LSTM | 2.82 | 3.28 | 4.21 | 5.32 |
| CNN-LSTM | 2.79 | 3.24 | 4.17 | 5.27 |
| Proposed | 2.21 | 2.82 | 3.99 | 5.33 |

**Attention weights and mode l combination.** For the seed-42 model, the mean attention profile has a diurnal rhythm spread over all seven past days rather than concentrated on the most recent one. The attention mass given to each past day correlated only weakly with its temperature similarity to the target day (mean Spearman $\rho = 0.15$, positive on 60% of days; the most-attended day was the most similar one on 18.2% of days against 14.3% by chance), so the layer acts mainly as a non-linear mixer of target-day information with the past rather than as a selector of analogue days, consistent with

the cautions of [14]. Because the errors of the MLP and the proposed model are not perfectly correlated, their ensembles were also averaged with equal weights as an exploratory check that was neither pre-specified nor tuned. The combination reached a MAPE of 2.02% and an RMSE of 895 MW, significantly better than the MLP ($p = 0.005$) and the proposed model ($p < 0.001$).

### 3.2. Discussion

The results answer the research question in two parts. First, conditioned attention helps exactly when its query carries informative, forward-looking information about the target day. With calendar inputs only, a 64-unit LSTM can retain everything relevant in its final state and the attention path is redundant; with temperature, the attention query is the only place in the recurrent architecture where target-day weather interacts non-linearly with the history, which explains the one-fifth error reduction. This complements earlier reports of gains from attention in CNN-LSTM and LSTM forecasters [10], [11], which did not separate the effect of the inputs from that of the architecture.

Second, a feed-forward MLP with the same information was as accurate as the proposed model in single runs and more accurate as a seed ensemble, in line with recent evidence that simple linear and MLP-based models rival far more complex attention-based forecasters [15]–[17]. Three properties of the problem favour the MLP: it sees every lagged load and the target-day temperatures simultaneously in its first hidden layer, whereas the recurrent models must first compress the past week into a 64-dimensional state; its direct access to the latest observations gives it a clear advantage at short horizons; and about 35,000 strongly autocorrelated training windows do not require the extra flexibility of recurrent architectures. The proposed model, in turn, was slightly more accurate for the later hours of the next day. The study also illustrates the evaluation pitfalls described in [20], [21]: target leakage had only a small effect here, but asymmetric seed ensembling and different early-stopping rules reversed the ranking between the two best models. Since seed ensembling benefits diverse models most and the two model families make partly different errors, combinations of different model families [22], [23] are a promising operational choice.

For utilities, temperature information is indispensable, a well-configured MLP and a regularised linear model should always be included as benchmarks, and among recurrent architectures a CNN-LSTM without a conditioned attention query should not be deployed when temperature forecasts are available. The NumPy implementation trains without a GPU in 5–15 minutes per recurrent model on a single CPU core, against 1–9 minutes for the MLP, which is relevant for lightweight deployment in substations and distribution control centres [27]. The study has limitations: observed temperatures were used as a proxy for forecasts, a single region and a seven-month test period were analysed, only point forecasts were produced although probabilistic forecasts are increasingly required [28], the hyperparameters of all neural models were fixed a priori, three seeds give only a rough estimate of run-to-run variability, and the combination result is exploratory.

## 4. CONCLUSIONS

This study developed a CNN-LSTM forecaster with a weather- and calendar-conditioned temporal attention mechanism and evaluated it for day-ahead load forecasting on PJM East data under a purged chronological split, identical inputs, identical early stopping and three training seeds for every stochastic model, and formal significance tests. With temperature inputs, the model reached a test MAPE of 2.21%, was significantly more accurate than persistence, ridge regression, an LSTM and a CNN-LSTM, and was the most accurate model for hours 12–23 ahead; its attention layer reduced the error of the CNN-LSTM by about one fifth, but only when the query carried target-day temperature information. A multilayer perceptron with the same information was equally accurate in single runs but significantly more accurate as a seed ensemble (2.13%) and more robust to temperature-forecast errors, and the advantage previously attributed to the attention model arose from asymmetric evaluation. Attention-based hybrids should therefore be judged against well-configured feed-forward benchmarks under symmetric, leakage-free protocols, and their most promising role may be as members of forecast combinations. Future work will train the models on archived numerical weather forecasts, tune all neural models with a common search budget, develop a validated combination scheme and extend the approach to probabilistic forecasting and other regions.

## REFERENCES

[1] Y. Eren and İ. Küçükdemiral, "A comprehensive review on deep learning approaches for short-term load forecasting," *Renewable and Sustainable Energy Reviews*, vol. 189, Art. no. 114031, 2024, doi: https://doi.org/10.1016/j.rser.2023.114031.

[2] H. Hou, C. Liu, Q. Wang, X. Wu, J. Tang, Y. Shi, and C. Xie, "Review of load forecasting based on artificial intelligence methodologies, models, and challenges," *Electric Power Systems Research*, vol. 210, Art. no. 108067, 2022, doi: https://doi.org/10.1016/j.epsr.2022.108067.

[3] F. Dewangan, A. Y. Abdelaziz, and M. Biswal, "Load forecasting models in smart grid using smart meter information: A review," *Energies*, vol. 16, no. 3, Art. no. 1404, 2023, doi: https://doi.org/10.3390/en16031404.

[4] R. Wazirali, E. Yaghoubi, M. S. S. Abujazar, R. Ahmad, and A. H. Vakili, "State-of-the-art review on energy and load forecasting in microgrids using artificial neural networks, machine learning, and deep learning techniques," *Electric Power Systems Research*, vol. 225, Art. no. 109792, 2023, doi: https://doi.org/10.1016/j.epsr.2023.109792.

[5] A. Candela Esclapez, M. López García, S. Valero Verdú, and C. Senabre Blanes, "Automatic selection of temperature variables for short-term load forecasting," *Sustainability*, vol. 14, no. 20, Art. no. 13339, 2022, doi: https://doi.org/10.3390/su142013339.

[6] N. Bashiri Behmiri, C. Fezzi, and F. Ravazzolo, "Incorporating air temperature into mid-term electricity load forecasting models using time-series regressions and neural networks," *Energy*, vol. 278, Art. no. 127831, 2023, doi: https://doi.org/10.1016/j.energy.2023. 127831.

[7] J. S. Caicedo-Vivas and W. Alfonso-Morales, "Short-term load forecasting using an LSTM neural network for a grid operator," *Energies*, vol. 16, no. 23, Art. no. 7878, 2023, doi: https://doi.org/10.3390/en16237878.

[8] S. Zhang, J. Liu, and J. Wang, "High-resolution load forecasting on multiple time scales using long short-term memory and support vector machine," *Energies*, vol. 16, no. 4, Art. no. 1806, 2023, doi: https://doi.org/10.3390/en16041806.

[9] F. Saeed, A. Paul, and H. Seo, "A hybrid channel-communication-enabled CNN-LSTM model for electricity load forecasting," *Energies*, vol. 15, no. 6, Art. no. 2263, 2022, doi: https://doi.org/10.3390/en15062263.

[10] A. Wan, Q. Chang, K. AL-Bukhaiti, and J. He, "Short-term power load forecasting for combined heat and power using CNN-LSTM enhanced by attention mechanism," *Energy*, vol. 282, Art. no. 128274, 2023, doi: https://doi.org/10.1016/j.energy.2023.128274.

[11] J. Lin, J. Ma, J. Zhu, and Y. Cui, "Short-term load forecasting based on LSTM networks considering attention mechanism," *International Journal of Electrical Power & Energy Systems*, vol. 137, Art. no. 107818, 2022, doi: https://doi.org/10.1016/j.ijepes.2021.107818.

[12] A. L'Heureux, K. Grolinger, and M. A. M. Capretz, "Transformer-based model for electrical load forecasting," *Energies*, vol. 15, no. 14, Art. no. 4993, 2022, doi: https://doi.org/10.3390/en15144993.

[13] C. Wang, Y. Wang, Z. Ding, T. Zheng, J. Hu, and K. Zhang, "A transformer-based method of multienergy load forecasting in integrated energy system," *IEEE Transactions on Smart Grid*, vol. 13, no. 4, pp. 2703–2714, 2022, doi: https://doi.org/10.1109/TSG.2022.3166600.

[14] W. Choi and S. Lee, "Interpretable deep learning model for load and temperature forecasting: Depending on encoding length, models may be cheating on wrong answers," *Energy and* Buildings, vol. 297, Art. no. 113410, 2023, doi: https://doi.org/10.1016/j.enbuild.2023.113410.

[15] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?," in *Proceedings of the AAAI Conference on* Artificial *Intelligence*, vol. 37, no. 9, pp. 11121–11128, 2023, doi: https://doi.org/10.1609/aaai.v37i9.26317.

[16] C. Challu, K. G. Olivares, B. N. Oreshkin, F. Garza Ramirez, M. Mergenthaler Canseco, and A. Dubrawski, "NHITS: Neural hierarchical interpolation for time series forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, pp. 6989–6997, 2023, doi: https://doi.org/10.1609/aaai.v37i6.25854.

[17] V. Ekambaram, A. Jati, N. Nguyen, P. Sinthong, and J. Kalagnanam, "TSMixer: Lightweight MLP-Mixer model for multivariate time series forecasting," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 459–469, 2023, doi: https://doi.org/10.1145/3580305.3599533.

[18] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. Available: https://openreview.net/forum?id=Jbdc0vTOcol.

[19] F. Petropoulos *et al.*, "Forecasting: theory and practice," *International Journal of*

*Forecasting*, vol. 38, no. 3, pp. 705–871, 2022, doi: https://doi.org/10.1016/j.ijforecast.2021.11.001.

[20] H. Hewamalage, K. Ackermann, and C. Bergmeir, "Forecast evaluation for data scientists: *Common* pitfalls and best practices," *Data Mining and Knowledge Discovery*, vol. 37, no. 2, pp. 788–832, 2023, doi: https://doi.org/10.1007/s10618-022-00894-5.

[21] S. Kapoor and A. Narayanan, "Leakage and the *reproducibility* crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Art. no. 100804, 2023, doi: https://doi.org/10.1016/j.patter.2023.100804.

[22] R. Gao, L. Du, P. N. Suganthan, Q. Zhou, and K. F. Yuen, "*Random* vector functional link neural network based ensemble deep learning for short-term load forecasting," *Expert Systems with Applications*, vol. 206, Art. no. 117784, 2022, doi: https://doi.org/10.1016/j.eswa.2022.117784.

[23] X. Wang, R. J. Hyndman, F. Li, and Y. Kang, "Forecast combinations: An over 50-year review," *International Journal of Forecasting*, vol. 39, no. 4, pp. 1518–1547, 2023, doi: https://doi.org/10.1016/j.ijforecast.2022.11.005.

[24] R. Mulla, "Hourly energy consumption (PJM)," Kaggle dataset, 2018. Accessed: Sep. 29, 2026. Available: https://www.kaggle.com/datasets/robikscube/hourly-energy-consumption.

[25] K. Xie, A. Giacomoni, X. Deng, and Y. Wu, "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*, vol. 336, Art. no. 138411, 2025, doi: https://doi.org/10.1016/j.energy.2025.138411.

[26] P. Zippenfenig, "Open-Meteo.com Weather API," Zenodo, 2024, doi: https://doi.org/10.5281/zenodo.7970649.

[27] S. Bani Melhem, M. Golec, S. Alrabaee, M. Alshaikh, and M. Uyar, "EdgeAI-powered hybrid ESN-GRU model for high-accuracy and efficient short-term load forecasting in smart grids," *IEEE Access*, vol. 13, pp. 198241–198251, 2025, doi: https://doi.org/10.1109/ACCESS.2025.3631672.

[28] Y. Yang, Q. Xing, K. Wang, C. Li, J. Wang, and X. Huang, "A novel combined probabilistic load forecasting system integrating hybrid quantile regression and knee improved multi-objective optimization strategy," *Applied Energy*, vol. 356, Art. no. 122341, 2024, doi: https://doi.org/10.1016/j.apenergy.2023.122341.