Proceedings of CONF-MLA 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA36234

One-Hour-Ahead Short-Term Electricity Load Forecasting Using Long Short-Term Memory Networks

Jingkai Gao

SWJTU-Leeds Joint School, Southwest Jiaotong University, Chengdu, China
gjk200561@my.swjtu.edu.cn

Abstract. Accurate one-hour-ahead electricity load forecasting sustains dispatch, reserve planning, and dependable power-system operation, but additional inputs do not always improve predictions. This study assesses how different parameter settings contribute to power demand forecasting. Four long short-term models with a unified architecture were trained on 48048 hourly observations and a 24-hour window. Input configurations include historical demand alone, demand with twelve weather variables, demand with two calendar indicators, or all inputs. Each configuration was trained five times and evaluated on a chronological test set. The weather-based model achieved the lowest mean errors: a mean absolute percentage error of 1.599% and a root mean square error of 26.333. The history-only model remained competitive, while calendar indicators and the all-input model offered no improvement. Recent demand therefore provides most of the useful information at this horizon, with lagged weather adding a modest signal. Careful feature selection can reduce complexity and support interpretable, dependable operational forecasts.

Keywords: Short-term electricity load forecasting, long short-term memory, feature ablation

1. Introduction

Short-term load forecasting, which helps maintain the balance between electricity generation and consumption, is crucial for power dispatch, reserve capacity planning, demand response, and system reliability. In particular, one-hour-ahead predictions are useful as operators can respond to near-term changes while recent observations remain relevant.

Recent studies have approached this problem through optimization, decomposition and hybrid networks. Hong and Chan proposed a method called enhanced elite-based particle swarm optimization, which was adopted to improve a convolutional network that utilized external variables including temperature, humidity, weekday, and holiday. It was found that in half of the tests, mean absolute error (MAE) declined by about 14%, root mean square error (RMSE) decreased by about 15%, and coefficient of determination (R-squared) rose by 0.0347 to 0.1018 in all the tests [1]. Fan et al. improved a random forest model that only used historical load as input by combining it with empirical wavelet transform, decreasing mean absolute percentage error (MAPE) by 31.34%, RMSE by 24.33%, and mean square error (MSE) by 16.14% compared with the normal wavelet transform random forest model [2]. In summary, it is indicated that the prediction accuracy can be enhanced through model optimization, signal decomposition, and hybrid architectures.

DOI: 10.54254/2755-2721/2026.BA36234

For LSTM-based research, Wang et al. used apparent temperature and weekday as external variables, observing that the average MAPE was reduced from 53.77% to 48.46% [3]. For commercial buildings, Chitalia et al. compared various feature sets and found that external variables including outside temperature, hour of day, and weekday were most beneficial to the model performance [4]. Eskandari et al. conducted an experiment on feature ablation, which showed that with the integration of external variables including temperature, season, weekday, and holiday, MAPE was reduced by 8.2% [5]. Mansoor selected month and historical load as model inputs. By introducing a graph convolutional network, the MAE of the LSTM model was decreased by 0.18% [6]. Consequently, leveraging auxiliary variables to improve model performance has become a common approach in this field.

For long-term forecasting, macroscopic variables such as weather and population provide more predictive information than recent load as they enhance the model performance greatly [7]. As the forecasting horizon narrows down, the smaller the time distance between historical load and predicted target is, the more information the historical load will have [8]. Meteorological and calendar variables probably provide only limited incremental information. The purpose of this study is to determine the marginal predictive value of meteorological and calendar information for one-hour-ahead load forecasting, helping the readers to select appropriate inputs when predicting short-term load.

2. Method

2.1. Data source and description

The data used in this study were downloaded from a Kaggle mirror of the Short-term electricity load forecasting (Panama case study) dataset, which was published by Aguilar Madrid [9]. The dataset covers 48048 hourly observations from 3 January 2015 at 1:00 to 27 June 2020 at 0:00 with no duplicate timestamps or missing values. The time index is continuous at one-hour intervals. The dataset also contains meteorological measurements from Tocumen, Santiago, and David, together with calendar information. The dataset is suitable for this study because it not only provides a continuous hourly time series of national electricity demand (NED), but also includes various inputs, which enable a controlled comparison of historical load, weather, and calendar inputs.

2.2. Data preprocessing and feature selection

During a data-quality check, it was found that there were 13 abnormally low NED records as they were below the empirical second percentile of the NED series. At the same time, it was lower than 60% of the median load levels for the same weekdays and the same times in the previous and following four weeks. Considering that these unusual values only lasted for one to five hours, they were replaced using linear interpolation between adjacent valid hourly observations, as it can maintain the feature of the current load as well as the short-term trend of changes. NED was used as the forecasting target. The meteorological features consist of temperature (T2M), specific humidity (QV2M), liquid precipitation (TQL), and wind speed (W2N) for each of the three cities. Calendar information includes Holiday ID, holiday, and school. Holiday ID was excluded as it represents nominal holiday categories rather than ordered numerical quantities. Using this input may cause an artificial ordinal relationship among holiday types. The binary holiday and school inputs were reserved as they represent road schedule effects.

DOI: 10.54254/2755-2721/2026.BA36234

2.3. LSTM forecasting method

Long Short-Term Memory (LSTM) is a gated recurrent network designed to model temporal dependencies while reducing the vanishing-gradient problem of conventional RNNs. As shown in Figure 1, its cell state carries information across time. The forget gate controls retained memory, the input gate controls new candidate information, and the output gate produces the hidden state. Sigmoid and hyperbolic tangent functions bound these signals, enabling LSTM to learn short- and long-term load patterns [3]. This gated design makes LSTM particularly suitable for hourly electricity-load forecasting, where recent demand, daily cycles, and longer calendar-related patterns may jointly affect future load. Compared with a conventional RNN, LSTM can selectively retain relevant historical information and discard less useful information, enabling it to model nonlinear relationships between past load, weather, calendar conditions, and future demand.

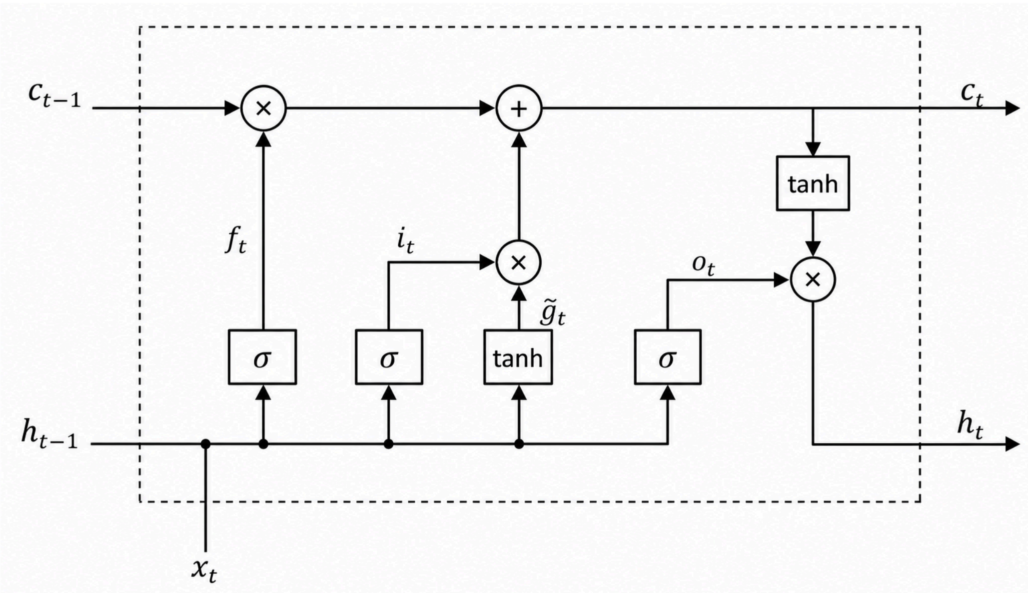

Figure 1. Structure and information flow of an LSTM memory unit [3]

The calculation proceeds in four stages. First, the forget gate f_t assigns values between 0 and 1 to the previous cell state and therefore determines how much earlier information is retained. Second, the input gate i_t controls the contribution of the candidate memory $\widetilde{C}_t$, which is formed from the current input and the previous hidden state. Third, C_t combines the retained memory with the selected new information. Finally, the output gate o_t filters the transformed cell state to produce h_t, which is passed to the next time step and, in the stacked network, to the next LSTM layer.

Figure 2 summarizes the four feature groups. LSTM-History uses lagged NED only (d = 1), which serves as a baseline. LSTM-Weather adds 12 meteorological variables from Tocumen, Santiago, and David (d = 13), enabling the model to examine whether meteorological information can supplement more changes. LSTM-Calendar adds the binary holiday and school indicators (d = 3), while LSTM-Select-All combines demand, all weather variables, and both calendar indicators (d = 15). Datetime orders the observations rather than serving as a numerical input. Each model uses only the preceding 24 hours to forecast demand at an hour.

DOI: 10.54254/2755-2721/2026.BA36234

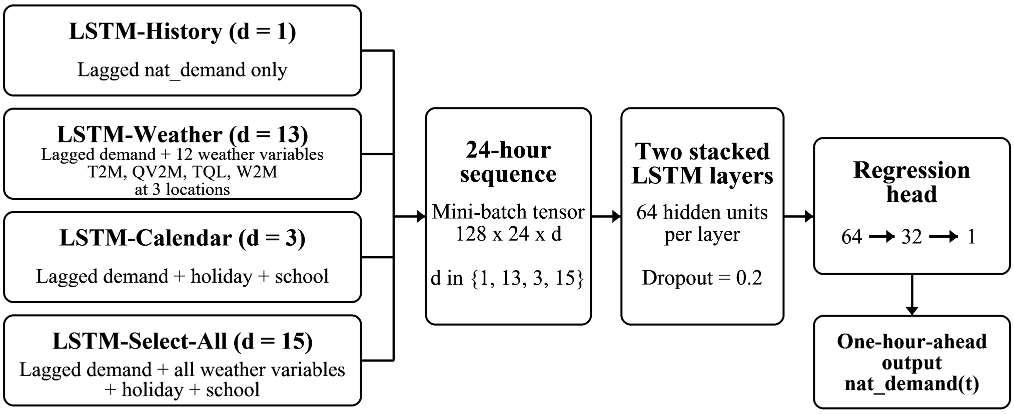

Figure 2. Four-group LSTM feature-ablation experiment (original)

The 24-hour lookback produces 48,024 sequences, divided chronologically into 33,616 training (70%), 7,204 validation (15%), and 7,204 test samples (15%). This split reflects forecasting later demand from earlier observations and prevents future information from entering training. Input scalers are fitted separately for each feature group using training data, while one common target scaler is fitted to training targets. The fitted scalers are then applied unchanged to validation and test data.

All configurations use the same architecture and training settings; only d changes. Each 24 x d sequence enters two 64-unit LSTM layers with dropout 0.2 and a 64-32-1 ReLU regression head. Mini-batches contain up to 128 sequences. Training uses MSE loss, Adam (learning rate 0.001), and gradient clipping at 1.0. Parameter count ranges from 52,545 for LSTM-History to 56,129 for LSTM-Select-All.

Each configuration is trained for 100 epochs with seeds 42-46. The checkpoint with the lowest validation loss is retained, without using the test set for selection. After inverse scaling, MAE, MAPE, RMSE, and R-squared are calculated for each run. Table 3 reports mean MAPE and RMSE across the five runs because these metrics express relative and scale-dependent error, respectively. Standard deviations and additional metrics are retained as diagnostics.

3. Results

3.1. Comparative forecasting performance

Table 1 reports mean test-set MAPE and RMSE across five 100-epoch runs. The four models use History (d = 1, demand only), Weather (d = 13, demand plus 12 weather variables), Calendar (d = 3, demand plus holiday and school), and Select-All (d = 15, all inputs). The lowest-validation-loss checkpoint is evaluated for each seed, and lower metric values indicate better performance.

Table 1. Mean test-set performance of the four LSTM feature groups across five runs

Model	Input dimension (d)	MAE	MAPE (%)	RMSE	R-squared
LSTM-History	1	19.351	1.625	27.244	0.9782
LSTM-Weather	13	19.054	1.599	26.333	0.9796
LSTM-Calendar	3	19.767	1.672	28.076	0.9768

DOI: 10.54254/2755-2721/2026.BA36234

Table 1. (continued)

LSTM-Select-All	15	19.553	1.647	27.278	0.9781

LSTM-Weather performed best, with a MAPE of 1.599% and an RMSE of 26.333; its RMSE was 3.34% below LSTM-History. LSTM-History ranked second (1.625%, 27.244). LSTM-Select-All was close to the history baseline (1.647%, 27.278), whereas LSTM-Calendar performed worst (1.672%, 28.076). Thus, weather improved average accuracy, but adding calendar indicators did not preserve that gain.

3.2. Prediction curve comparison

Predictions cover all 7,204 test hours from 31 August 2019 to 27 June 2020. Figure 3 enlarges the 17-hour interval with the greatest normalized mean pairwise separation among model predictions. The excerpt, from 22:00 on 18 January to 14:00 on 19 January 2020, was selected only after predictions had been generated for the complete test set. It is therefore a visual aid rather than a separate test. Each curve uses the run with the lowest validation loss; solid black lines and distinct markers distinguish the series.

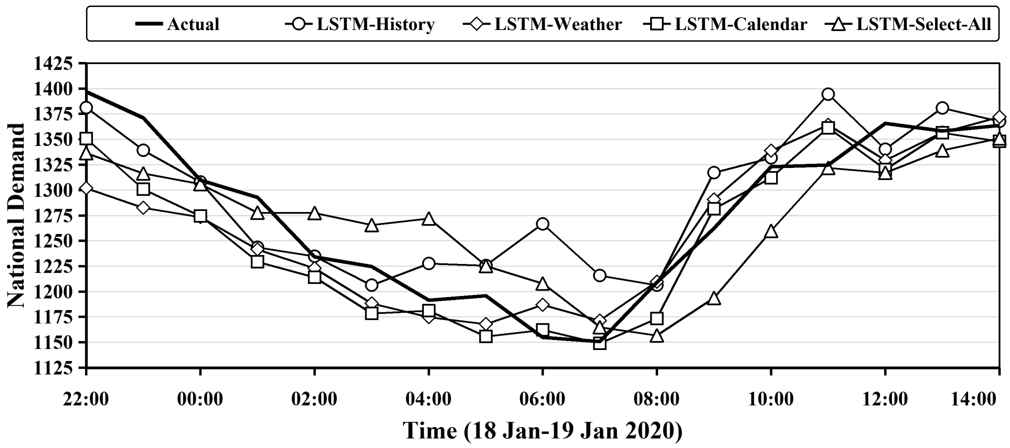

Figure 3. Enlarged 17-hour excerpt selected from the complete test-set prediction curves to show differences among the four LSTM feature groups (original)

The excerpt reveals different responses to the overnight decline and morning increase. LSTM-History is initially closest to the high observed demand, while LSTM-Weather and LSTM-Calendar generally predict a lower trough. During the morning rise, LSTM-History, LSTM-Weather, and LSTM-Calendar respond relatively quickly, whereas LSTM-Select-All responds more slowly. LSTM-History then briefly overpredicts before the curves converge later in the interval. These are local behavioural differences; the overall ranking is based on all 7,204 test observations.

4. Discussion

Overall, the comparison results suggest that the predictive value of a set of features depends on whether it provides more information beyond what the recent load alone provides. Therefore, a possible mechanism will be explored in the following paragraphs.

DOI: 10.54254/2755-2721/2026.BA36234

The enhanced performance of LSTM-Weather can be regarded as evidence that meteorological information provides a supplementary signal. Variables such as temperature and humidity assist the model in discriminating similar load patterns, and these patterns may display different evolution trends under various weather conditions. This accords with previous studies where temperature information was claimed to affect prediction accuracy [10, 11].

LSTM-Calendar has a weaker performance, maybe because calendar information was represented by binary indicators, which can not comprehensively capture the hourly variations of time and the patterns of working days. Recent evidence suggests that explicit calendar-state encoding can improve load forecasting accuracy [12].

The fact that LSTM-Select-All does not provide additional gains indicates that adding more variables does not lead to better prediction. Some of the inputs may be redundant. Using the same dataset, Yang et al. reported an overall mean variance inflation factor (VIF) of 6.33, where the VIFs for T2M of David, Santiago, and Tocumen were 15.26, 20.28, and 11.31, respectively [13]. These values exceed the commonly used VIF = 10 benchmark for very strong multicollinearity [14]. Furthermore, they may increase the difficulty for the model to learn patterns from the existing data. This exploration is consistent with Zhang et al., who found that selecting an appropriate number of weather variables may be more effective than using all available meteorological inputs [15].

5. Conclusion

Short-term electricity load forecasting supports reliable dispatch and reserve planning, but adding more inputs does not automatically improve LSTM performance. This study compared four LSTM models with different input settings. Among the four controlled feature groups, LSTM-Weather achieved the lowest test errors, with a MAPE of 1.599% and an RMSE of 26.333. LSTM-History, utilizing only past demand, remained competent. However, LSTM-Calendar performed the worst, and LSTM-Select-All did not outperform the weather-based model. The results show that recent demand carries most of the useful information for one-hour-ahead forecasting, while meteorological data supplies a measurable complementary signal.

The results of this ablation experiment indicate that integrating more parameters does not always improve the accuracy of predictions and helps to clarify which variety of information genuinely benefits the prediction accuracy. The finding is beneficial for grid operators to reduce unnecessary variables while maintaining accuracy. For other researchers, the finding indicates that selecting appropriate inputs is another way to improve model performance. Nevertheless, this study is limited to one dataset, a chronological test period, a model architecture, and five random seeds. Therefore, future work should apply rolling-origin evaluation and Diebold-Mariano tests to examine whether the performance among feature groups is statistically significant. Additionally, more model architectures and combinations of inputs and datasets should be explored.

References

[1] Hong, Y.-Y., & Chan, Y.-H. (2023). Short-term electric load forecasting using particle swarm optimization-based convolutional neural network. *Engineering Applications of Artificial Intelligence, 126*, Article 106773. https://doi.org/10.1016/j.engappai.2023.106773

[2] Fan, G.-F., Peng, L.-L., & Hong, W.-C. (2022). Short-term load forecasting based on empirical wavelet transform and random forest. *Electrical Engineering, 104*, 4433–4449. https://doi.org/10.1007/s00202-022-01628-y

[3] Wang, Y., Zhang, N., & Chen, X. (2021). A short-term residential load forecasting model based on LSTM recurrent neural network considering weather features. *Energies, 14*, Article 2737. https://doi.org/10.3390/en14102737

DOI: 10.54254/2755-2721/2026.BA36234

[4] Chitalia, G., Pipattanasomporn, M., Garg, V., & Rahman, S. (2020). Robust short-term electrical load forecasting framework for commercial buildings using deep recurrent neural networks. *Applied Energy, 278*, Article 115410. https://doi.org/10.1016/j.apenergy.2020.115410

[5] Eskandari, H., Imani, M., & Parsa Moghaddam, M. (2021). Convolutional and recurrent neural network-based model for short-term load forecasting. *Electric Power Systems Research, 195*, Article 107173. https://doi.org/10.1016/j.epsr.2021.107173

[6] Mansoor, H., Gull, M. S., Rauf, H., Shaikh, I. u. H., Khalid, M., & Arshad, N. (2024). Graph convolutional networks based short-term load forecasting: Leveraging spatial information for improved accuracy. *Electric Power Systems Research, 230*, Article 110263. https://doi.org/10.1016/j.epsr.2024.110263

[7] Perez Guillen, A., & Milanovic, J. (2026). Machine learning-based long-term forecasting of load composition considering exogenous variables. *IEEE Transactions on Smart Grid*. https://doi.org/10.1109/TSG.2026.3660327

[8] Burg, L., Gürses-Tran, G., Madlener, R., & Monti, A. (2021). Comparative analysis of load forecasting models for varying time horizons and load aggregation levels. *Energies, 14*, Article 7128. https://doi.org/10.3390/en14217128

[9] Aguilar Madrid, E. (2021). Short-term electricity load forecasting (Panama case study) [Data set]. *Mendeley Data*. https://doi.org/10.17632/byx7sztj59.1

[10] Pelekis, S., Seisopoulos, I.-K., Spiliotis, E., Pountridis, T., Karakolis, E., Mouzakitis, S., & Askounis, D. (2023). A comparative assessment of deep learning models for day-ahead load forecasting: Investigating key accuracy drivers. *Sustainable Energy, Grids and Networks, 36*, Article 101171. https://doi.org/10.1016/j.segan.2023.101171

[11] Shering, T., Alonso, E., & Apostolopoulou, D. (2024). Investigation of load, solar and wind generation as target variables in LSTM time series forecasting, using exogenous weather variables. *Energies, 17*, Article 1827. https://doi.org/10.3390/en17081827

[12] He, B., Cai, C., Diao, R., Han, J., Qian, B., & Wu, S. (2026). A calendar-aware frequency-decoupled framework for day-ahead substation load forecasting using SHAP-based interpretation. *Applied Sciences, 16*, Article 5815. https://doi.org/10.3390/app16125815

[13] Yang, Z., Li, J., Wang, H., & Liu, C. (2025). An Informer model for very short-term power load forecasting. *Energies, 18*, Article 1150. https://www.mdpi.com/1996-1073/18/5/1150

[14] Jović, A. S., Korunović, L. M., & Djokic, D. S. (2021). Application of meteorological variables for the estimation of static load model parameters. *Energies, 14*, Article 4874. https://www.mdpi.com/1996-1073/14/16/4874

[15] Zhang, E., Mirallès, F., Dellacherie, S., Wu, D., & Boulet, B. (2025). Causal feature selection for weather-driven residential load forecasting. *arXiv*. https://doi.org/10.48550/arXiv.2511.20508