Proceedings of the 4th International Conference on Functional Materials and Civil Engineering
DOI: 10.54254/2755-2721/2026.33612

Load-Memory and Weather-Calendar Feature-Driven Short-Term Load Probabilistic Forecasting with Split Conformal Calibration

Ruihan Zhao

College of Agricultural Engineering and Food Science, Shandong University of Technology, Zibo, China
2948189935@qq.com

Abstract. Short-term load forecasting is essential for power system scheduling, reserve allocation, demand response, and risk-aware operation. However, deterministic point forecasts cannot fully describe load uncertainty under weather variability, calendar effects, and non-stationary demand patterns. This paper proposes a probabilistic short-term load forecasting framework that integrates historical load-memory features, weather variables, calendar indicators, degree-based temperature proxies, lag variables, and rolling statistics. Linear Regression, Random Forest, and Histogram-based Gradient Boosting Regression are first compared for point forecasting, and quantile regression models are then used to construct raw 90% prediction intervals. To improve interval reliability, a split conformal calibration layer is further introduced as a post-processing step. Experiments are conducted on an hourly synthetic load dataset covering 2019–2022. The results show that Linear Regression achieves the lowest RMSE, while Histogram-based Gradient Boosting Regression obtains slightly better MAE and MAPE. Feature ablation confirms that historical load-memory features provide the main forecasting basis, while weather and calendar variables offer complementary information. For probabilistic forecasting, split conformal calibration improves PICP from 0.8290 to 0.8686 and reduces the Winkler Score, indicating better interval reliability and overall quality. Nevertheless, the calibrated coverage remains below the nominal 90% level, suggesting that adaptive calibration is still needed under distribution shift and extreme load fluctuations.

Keywords: short-term load forecasting, probabilistic forecasting, load-memory features, split conformal calibration, prediction interval

1. Introduction

Short-term load forecasting (STLF) plays a central role in day-ahead scheduling, reserve allocation, demand response, electricity trading, and operational risk assessment. Conventional STLF studies usually formulate the task as deterministic point forecasting and compare models using mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). However, power system operators do not face a single deterministic future load value.

DOI: 10.54254/2755-2721/2026.33612

Instead, future demand is driven by weather, calendar effects, holidays, socioeconomic activities, and user behavior, all of which introduce uncertainty. A point forecast alone cannot answer operational questions such as how far the realized load may deviate from the expected value or how much reserve should be retained during peak periods.

Recent high-impact studies have moved beyond pure accuracy competition and have paid increasing attention to interpretability, reproducibility, probabilistic forecasting, and robustness under distribution shift. Pinheiro et al. emphasized that STLF methods should be assessed not only by predictive accuracy but also by applicability, interpretability, and reproducibility across system levels [1]. Eren and Kucukdemiral reviewed recent deep learning approaches for STLF and pointed out that model complexity does not automatically guarantee generalization and practical value [2]. These findings suggest that a rigorous STLF study should focus on feature mechanisms, data boundaries, and evaluation protocols, rather than simply stacking more complicated algorithms.

Probabilistic load forecasting provides a direct way to quantify future demand uncertainty by producing quantiles, prediction intervals, or full predictive distributions. He et al. proposed a nonparametric probabilistic load forecasting method based on quantile combination [3], while Xu et al. developed a curve-to-curve quantile regression approach for French half-hourly electricity loads [4]. Massidda and Marrocu further investigated probabilistic and causal machine learning methods for total and thermal load forecasting in residential communities [5]. These studies demonstrate that probabilistic forecasting is becoming an important pathway from accuracy-oriented forecasting to risk-aware decision support.

Nevertheless, three limitations remain important. First, many studies still focus mainly on model ranking, while the boundary roles of weather, calendar, and historical load-memory features are not sufficiently examined. Second, raw quantile models can produce prediction intervals, but the empirical coverage of these intervals may deviate from the nominal confidence level, particularly during seasonal transitions, heat waves, cold snaps, or abnormal peak-load events. Third, experiments based only on synthetic data must clearly explain the data-generation mechanism and should be interpreted as controlled proof-of-concept evidence rather than as a substitute for validation on public utility datasets. Recent studies on concept drift and distribution shift have shown that load forecasting models can suffer from reliability degradation in non-stationary environments [6,7]. This observation is consistent with the experimental results in this paper, where split conformal calibration improves coverage but does not fully reach the nominal 90% level.

To address these issues, this paper constructs a reliability-aware probabilistic STLF framework that integrates load-memory, weather, and calendar features with quantile forecasting and split conformal calibration. The contributions are threefold. First, a unified feature system is developed by integrating historical load lags, rolling statistics, meteorological variables, calendar indicators, and cooling/heating degree proxies. Second, point forecasting, quantile interval forecasting, and split conformal calibration are incorporated into one evaluation pipeline, shifting the focus from deterministic accuracy to reliability-aware forecasting. Third, point-model comparison, feature ablation, long-window visualization, reliability-sharpness analysis, and monthly coverage diagnostics are jointly used to identify effective operating conditions and failure boundaries.

2. Methodology

2.1. Experimental pipeline and data splitting

The proposed method consists of four components: multi-source feature engineering, point forecasting model comparison, quantile interval forecasting, and split conformal calibration. The

DOI: 10.54254/2755-2721/2026.33612

objective is not to claim that a single model is universally optimal, but to establish a complete pipeline that evaluates accuracy, reliability, and interval sharpness simultaneously. Unlike deterministic forecasting methods, the proposed framework reports both point forecasts and prediction intervals, which can provide richer information for risk-aware scheduling.

The data are split chronologically into training, calibration, and testing subsets. The training set is used to fit point and quantile models, the calibration set is used to estimate the split conformal adjustment, and the test set is reserved for final evaluation. This chronological splitting strategy avoids future information leakage and is closer to the practical deployment logic of STLF. The hourly synthetic case covers 2019-2022 and includes seasonal variations, daily peaks and valleys, weekend and holiday effects, heat waves, cold snaps, storm disturbances, and random operational events. After lag-feature construction, the dataset contains 34,728 valid samples.

2.2. Synthetic data generation and reproducibility

The synthetic hourly load dataset is generated to emulate common STLF operating patterns rather than to represent a specific utility system. The load sequence combines a base-load component, daily and weekly cycles, annual seasonality, weekend and holiday effects, temperature-sensitive cooling and heating components, storm-related disturbances, random operational events, and stochastic noise. Meteorological variables are generated with seasonal temperature trends, humidity and wind-speed variations, and weather disturbance indicators. This setting follows the benchmark logic of probabilistic energy forecasting studies, where load, weather, calendar, and post-processing components are jointly evaluated [8]. Therefore, the synthetic experiment should be interpreted as controlled proof-of-concept evidence, while external validation on public or utility datasets remains necessary for stronger generalization claims.

2.3. Multi-source feature engineering

The feature system contains three groups. The first group includes weather and physical proxy variables, such as temperature, humidity, wind speed, storm index, cooling degree, heating degree, and absolute temperature deviation. Cooling and heating degrees are designed to represent the load pressure induced by high-temperature air-conditioning demand and low-temperature heating demand. The second group includes calendar variables, such as hour, day of week, month, day of year, weekend indicator, holiday indicator, and bridge-day indicator. Sine-cosine encoding is applied to cyclic variables to avoid artificial discontinuities at period boundaries. The third group consists of historical load-memory features, including load lags at 1, 2, 3, 6, 12, 24, 48, 72, 168, and 336 hours, as well as rolling means and standard deviations over 6, 12, 24, 72, 168, and 336 hours.

2.4. Point and interval forecasting models

For point forecasting, Linear Regression, Random Forest, and Histogram-based Gradient Boosting Regressor (Hist GBR) are selected as representative models. Linear Regression serves as a transparent baseline and tests whether feature engineering has already linearized the major load patterns. Random Forest represents a nonlinear ensemble baseline capable of capturing local nonlinear structures. Hist GBR is a boosting-tree model that can model high-dimensional interactions among engineered features. MAE, RMSE, and MAPE are used jointly for point forecast evaluation. RMSE is more sensitive to large errors and is therefore particularly relevant for peak-load assessment.

DOI: 10.54254/2755-2721/2026.33612

For interval forecasting, quantile loss is used to train the 0.05, 0.50, and 0.95 quantile models, yielding a raw 90% prediction interval. Since raw quantile intervals can be systematically under-covered or over-conservative, split conformal calibration is introduced as a model-agnostic post-processing layer inspired by conformalized quantile regression [9]. On the calibration set, the nonconformity scores are computed from the deviations outside the raw interval. A quantile of these scores is then used as a global adjustment to expand the lower and upper bounds on the test set. This procedure does not modify the base forecasting model and can be attached to different probabilistic forecasting algorithms. Under chronological splitting, however, the calibrated coverage should be interpreted empirically because temporal dependence and distribution shift may weaken exact finite-sample coverage.

2.5. Evaluation metrics

The evaluation system includes both point and interval metrics. MAE, RMSE, and MAPE are used for point forecasts. PICP, ACE, MPIW, Winkler Score, and average Pinball Loss are used for interval forecasts. PICP measures the empirical coverage probability, ACE quantifies the deviation from nominal coverage, MPIW reflects interval sharpness, and Winkler Score penalizes both excessive width and missed observations. Therefore, interval quality is evaluated as a trade-off between reliability and sharpness rather than by interval width alone.

3. Results and discussion

3.1. Point forecasting performance

Table 1 reports the point forecasting results on the test set. Linear Regression achieves MAE, RMSE, and MAPE values of 4.1405, 5.3685, and 3.5997%, respectively. Hist GBR achieves 4.0203, 5.4893, and 3.4041%, while Random Forest achieves 4.0975, 5.5383, and 3.4862%. Linear Regression performs best in terms of RMSE, while Hist GBR performs slightly better in terms of MAE and MAPE. This result does not support the simplistic claim that nonlinear models are always superior. Instead, it indicates that well-designed lag, rolling, cyclic, and degree-based features can already express the dominant load patterns effectively.

Fig. 1 further compares the point forecasting errors. The key interpretation is that different loss functions highlight different model advantages. If large errors and peak deviations are more important, Linear Regression is preferred because of its lower RMSE. If average absolute error and relative error are prioritized, Hist GBR is slightly stronger. This interpretation is more useful for engineering applications than a single model ranking, because system operators often care about different loss functions under different operational objectives.

Table 1. Point forecasting performance comparison

Model	MAE	RMSE	MAPE(%)
Linear Regression	4.1405	5.3685	3.5997
Hist GBR	4.0203	5.4893	3.4041
Random Forest	4.0975	5.5383	3.4862

DOI: 10.54254/2755-2721/2026.33612

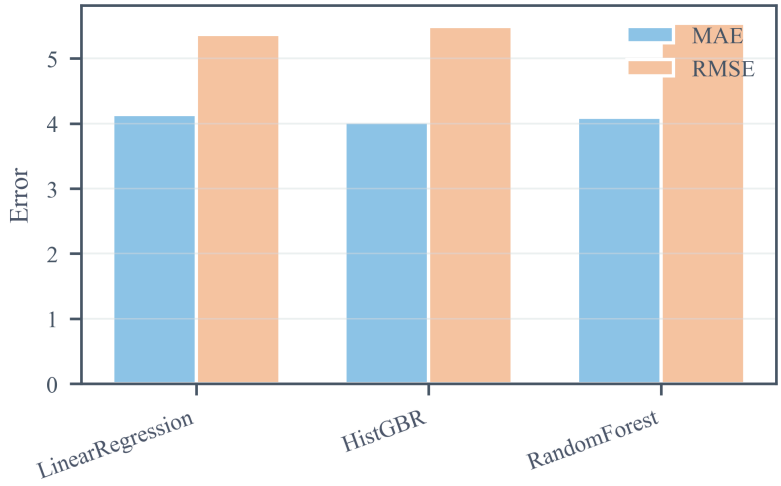

Figure 1. Error comparison of point forecasting models

3.2. Feature ablation analysis

The feature ablation results are shown in Fig. 2. The full 64-dimensional feature set obtains an RMSE of 5.4732. The history-only feature set achieves an RMSE of 6.0430. The weather-only, weather-calendar, and calendar-only feature sets lead to RMSE values of 9.8837, 10.0643, and 13.1236, respectively. These results show that historical load memory is the most important foundation for hourly STLF, while weather and calendar variables provide additional explanations for cooling/heating demand, holiday effects, and high-volatility periods.

This finding helps avoid overstatement of the role of weather-calendar features. The correct conclusion is not that weather and calendar information alone determines forecasting performance. Rather, historical load features provide the short-term inertia and periodic baseline, while weather and calendar features enrich the model with external driving information. Without historical load memory, weather and calendar variables alone cannot stably capture the short-term continuity of hourly demand.

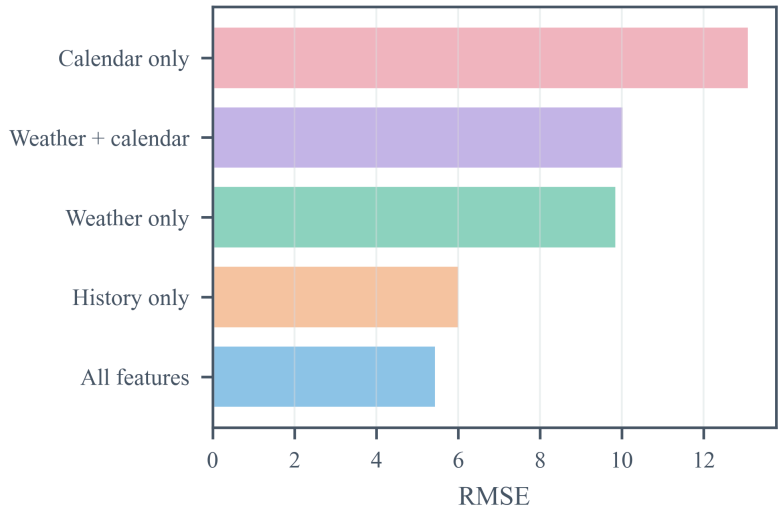

Figure 2. RMSE-based feature ablation results

DOI: 10.54254/2755-2721/2026.33612

3.3. Long-window forecast visualization

Fig. 3 presents the 90-day forecast window together with the split conformal calibrated prediction interval. The point forecast generally tracks daily load cycles and the seasonal increase in demand. The prediction interval covers the actual load during most normal periods. However, when load volatility increases after May, several realized peaks approach or exceed the upper bound of the interval. This suggests that the model has partial uncertainty-awareness but still under-expands intervals during extreme peaks and sudden fluctuations.

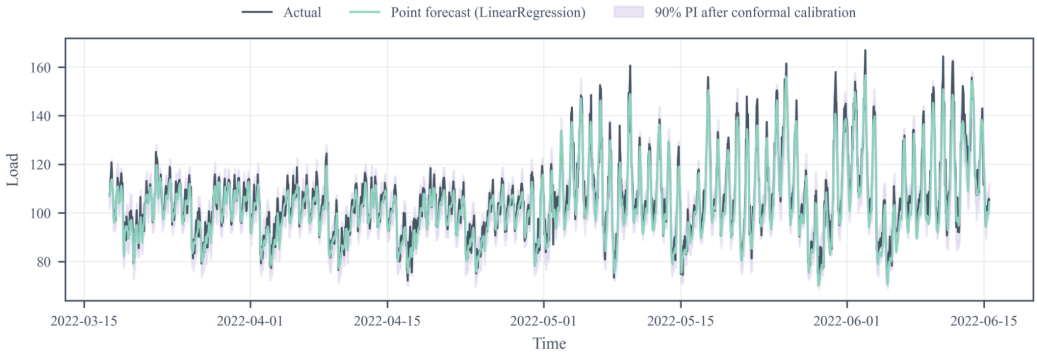

Figure 3. 90-day forecast trajectory and split conformal calibrated interval

3.4. Reliability-sharpness trade-off

Fig. 4 compares the reliability-sharpness trade-off between raw and calibrated prediction intervals. The raw quantile interval reaches a PICP of 0.8290, which is clearly below the nominal 90% level. After split conformal calibration, PICP increases to 0.8686 and ACE improves from -0.0710 to -0.0314, indicating that under-coverage is mitigated. At the same time, MPIW increases from 14.0441 to 15.5032, meaning that reliability improvement is achieved at the cost of wider intervals. More importantly, the Winkler Score decreases from 24.2819 to 23.5646, and the average Pinball Loss decreases from 2.1420 to 2.1181, implying that the overall reliability-sharpness balance is improved. However, the calibrated PICP still remains below the nominal 90% level and should not be interpreted as fully reliable coverage.

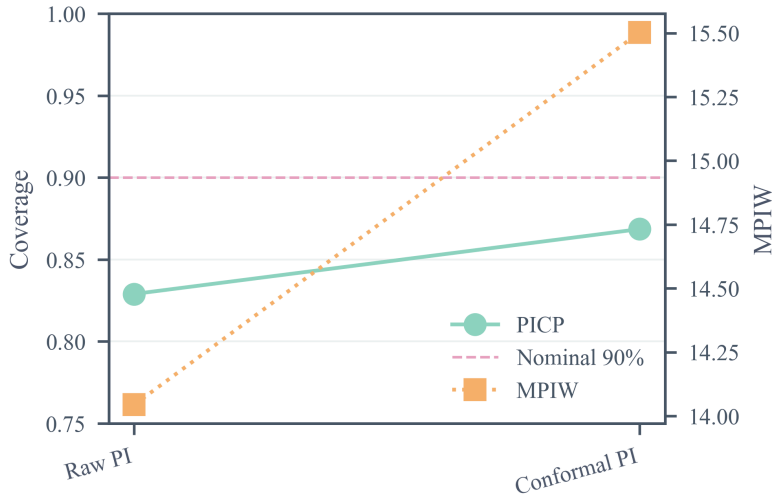

Figure 4. Reliability-sharpness trade-off between raw and calibrated intervals

DOI: 10.54254/2755-2721/2026.33612

The remaining coverage gap should be explicitly discussed. Possible reasons include distributional changes between calibration and test periods, long-tailed errors caused by extreme weather and operational events, and the limitation of using a single global adjustment in static split conformal calibration. Recent studies on concept drift and distribution shift also demonstrate that static probabilistic models may experience reliability degradation in non-stationary load series.

Table 2. Interval forecasting performance comparison

Method	PICP	ACE	MPIW	Winkler
Raw quantile 90% PI	0.8290	-0.0710	14.0441	24.2819
Quantile + split conformal 90% PI	0.8686	-0.0314	15.5032	23.5646

3.5. Monthly coverage and contribution boundary

Fig. 5 reports the monthly coverage stability. The calibrated intervals outperform the raw intervals in most months, but the coverage still drops below the nominal level in high-volatility summer and winter months. This indicates that under-coverage is not purely random; it is associated with seasonal load variation, weather sensitivity, and abnormal peaks. Future work should therefore investigate rolling calibration, adaptive conformal prediction, seasonal calibration, or weather-regime-based calibration.

From the perspective of contribution positioning, this study is best framed as a load-memory, weather, and calendar feature-driven probabilistic forecasting and reliability-diagnosis framework, rather than as a simple model-comparison paper. Its value lies in linking point forecasting, interval forecasting, coverage calibration, feature ablation, and stability diagnostics within a single experimental chain. The residual coverage gap is also informative because it shows that static calibration improves risk expression but still has limitations under non-stationary and high-volatility conditions.

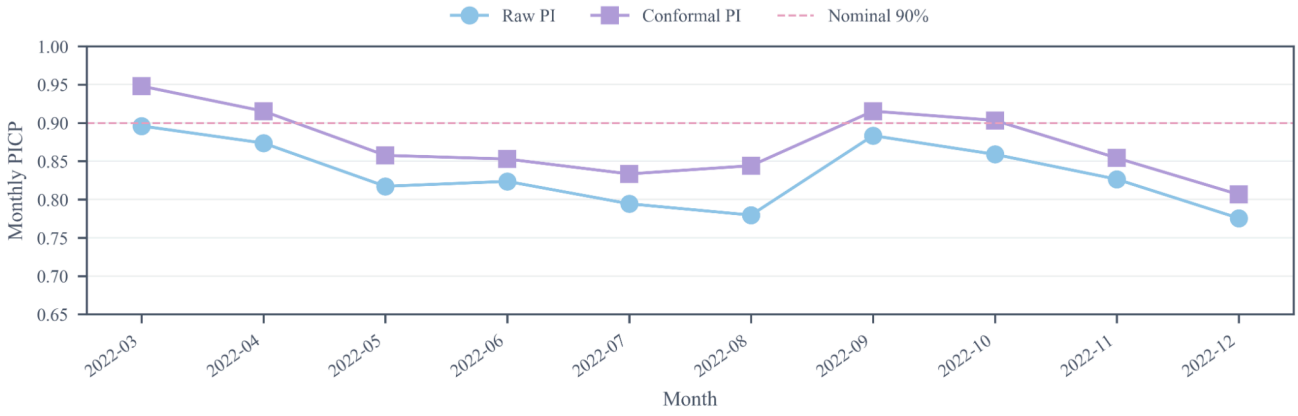

Figure 5. Monthly coverage stability of raw and calibrated intervals

4. Conclusion

This paper developed a reliability-aware short-term load probabilistic forecasting framework that integrates historical load-memory features, weather variables, and calendar indicators, and then calibrates quantile prediction intervals using split conformal prediction. In the 2019-2022 hourly synthetic case, Linear Regression achieved the lowest RMSE, while Hist GBR obtained slightly

DOI: 10.54254/2755-2721/2026.33612

better MAE and MAPE. The feature ablation analysis showed that historical load memory is the core basis of hourly STLF, while weather and calendar variables provide additional value for explaining cooling/heating demand, holiday effects, and high-volatility periods.

For probabilistic forecasting, the raw quantile interval suffered from clear under-coverage. Split conformal calibration improved PICP from 0.8290 to 0.8686 and reduced both Winkler Score and average Pinball Loss, showing the practical benefit and limitation of calibration in improving interval quality. However, the calibrated coverage still remained below the nominal 90% level, indicating the limitation of static global calibration under distribution shift and extreme fluctuations. Future research should validate the framework on real public load datasets and compare rolling calibration, adaptive conformal prediction, scenario-specific calibration, and online quantile ensembling methods for more reliable probabilistic STLF in non-stationary environments.

References

[1] Pinheiro, M.G., Madeira, S.C. and Francisco, A.P. (2023) Short-Term Electricity Load Forecasting—A Systematic Approach from System Level to Secondary Substations. Applied Energy, 332, 120493.
[2] Eren, Y. and Kucukdemiral, I. (2024) A Comprehensive Review on Deep Learning Approaches for Short-Term Load Forecasting. Renewable and Sustainable Energy Reviews, 189, 114031.
[3] He, Y., Cao, C., Wang, S. and Fu, H. (2022) Nonparametric Probabilistic Load Forecasting Based on Quantile Combination in Electrical Power Systems. Applied Energy, 322, 119507.
[4] Xu, X., Chen, Y., Goude, Y. and Yao, Q. (2021) Day-Ahead Probabilistic Forecasting for French Half-Hourly Electricity Loads and Quantiles for Curve-to-Curve Regression. Applied Energy, 301, 117465.
[5] Massidda, L. and Marrocu, M. (2023) Total and Thermal Load Forecasting in Residential Communities through Probabilistic Methods and Causal Machine Learning. Applied Energy, 351, 121783.
[6] Cao, C., He, Y., Zhou, Y. and Wang, S. (2025) An Online Probabilistic Combination Framework for Power Load Forecasting under Concept-Drifting Scenarios. Applied Energy, 399, 126518.
[7] Qin, D., Wu, X., Sun, D., Liang, Z. and Zhang, N. (2025) Load Forecasting under Distribution Shift: An Online Quantile Ensembling Approach. Applied Energy, 401, 126812.
[8] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond, " International Journal of Forecasting, vol. 32, no. 3, pp. 896-913, 2016.
[9] Y. Romano, E. Patterson, and E. J. Candes, "Conformalized quantile regression, " Advances in Neural Information Processing Systems, vol. 32, 2019.