

Causally Certified Energy Forecasting for Industrial Buildings: Beyond Accuracy toward Trustworthy Operational Decisions

Dongsoo Moon¹ and Gwangsu Lee^{*2}

¹Department of Computer Education, Sungkyunkwan University, Seoul 03063, Republic of Korea

²School of Computer Software, Daegu Catholic University, Gyeongsangbuk-do 38430, Republic of Korea

*Corresponding Author: Gwangsu Lee

Abstract— Electricity demand in industrial buildings exhibits nonstationary behavior, and errors during peak regimes can destabilize operational decisions. This study presents a fair, operation-oriented benchmarking of statistical, machine learning, and deep learning forecasters using a standardized preprocessing and validation protocol. Using 15-min data from 30 sites, 1- and 24-h-ahead forecasts were evaluated with a weighted CompositeScore integrating MAE, tail risk (P95AbsErr), and peak-regime reliability (PeakMAE95). Among 11 candidate models, gradient-boosted tree ensembles provided the most stable accuracy-risk trade-off. Compared with the persistence baseline, LightGBM reduced MAE and PeakMAE95 for the 1-h horizon by 19.6% and 32.1%, respectively, while a robust XGBoost variant achieved reductions of 30.5% and 36.1%, respectively, for the 24-h horizon. To ensure operational trustworthiness, a Causal Certification framework was applied to SHAP analysis. The selected models demonstrated high causal agreement between feature importance and intervention based sensitivity, with Spearman rank correlations of $\rho \approx 0.098$ (1-h) and $\rho \approx 0.881$ (24-h). These results support peak-aware model selection and causally-verified decision support for industrial energy operations.

Keywords— causal certification; intervention analysis; trustworthy machine learning; explainable AI (XAI); composite scoring; energy informatics; industrial buildings; peak demand; short-term load forecasting; tail risk

1. Introduction

1.1 Background and Motivation

Industrial facilities represent a demanding segment of the building sector because load variability and peak demand can directly affect power system constraints and facility level operations. Unlike typical commercial buildings, industrial buildings (e.g., factories and research/production sites) exhibit complex, nonstationary consumption patterns driven by process schedules, intermittent equipment operation, and rapidly varying HVAC and cooling loads. In this context, improving average forecasting accuracy alone is insufficient; operational decisions depend critically on reliability during peak and stressed regimes. Furthermore, as forecasting models move toward practical deployment, model trustworthiness—ensuring that the rationale behind a prediction aligns with the true causal drivers of load—becomes as important as accuracy.

Short term load forecasts support real-time and day- and week-ahead decisions, such as equipment scheduling, load shifting, and process adjustment. In industrial settings, heterogeneous usage patterns are common, and pattern diversity can degrade predictive consistency, motivating approaches that explicitly account for such diversity [1]. In addition, cooling and HVAC loads often drive peak formation, and forecasting is frequently integrated with storage-enabled operations (e.g., thermal energy storage) and peak shaving strategies. Studies combining signal decomposition with hybrid forecasting for industrial loads have shown improved practical performance under highly variable operating conditions [2]. However, these approaches often rely on association-based explanations that may not reflect the true causal mechanisms governing industrial load. Without verifying that identified feature importance is grounded in causal reality, operational trust in automated forecasting remains limited.

Recent studies increasingly position forecasting as an enabler of peak-oriented operations rather than a standalone prediction task. For example, peak mitigation through

operational adjustments under extreme weather conditions highlights that decisions can materially reshape peaks and that policy outcomes depend on forecast reliability during peak periods [3]. As operational strategies rely more on optimization-based control, the real time computational burden becomes a practical constraint. Learning-based predictive control has been explored to reduce online optimization costs while maintaining operational feasibility [4]. Together, these trends suggest that forecasting models should be evaluated not only by average error but also by peak and worst-case performance directly linked to operational stability. However, a critical gap remains: the absence of a framework to certify whether a model's decision logic is operationally sound under intervention. Therefore, this study compared multiple forecasting models for industrial load prediction under consistent experimental conditions, strengthened peak- and worst-case oriented evaluation aligned with operational requirements, and incorporated interpretability analyses to support deployment and decision-making. Specifically, a "causal certification" layer based on Baron (2023) [5] was integrated. This framework extends conventional interpretability by validating the causal drivers of model outputs, establishing a higher standard of operational trustworthiness for industrial energy decision-support systems.

1.2. Limitations of Existing Approaches

Building energy forecasting has been investigated using statistical methods (e.g., SARIMAX, Theta, and ETS), machine learning (ML) models (e.g., LightGBM, Random Forest, and XGBoost), and deep learning models (e.g., TCN, LSTM, and GRU). However, cross-study comparisons are often confounded by differences in preprocessing, feature construction, forecasting horizons, temporal resolutions, and training/validation protocols. These inconsistencies weaken reproducibility and make it difficult to determine whether performance differences arise from model design or experimental settings. In addition, numerous studies primarily report average metrics (e.g., MAE and RMSE), which can underrepresent risk-critical regimes. Even in multi-model benchmarking, performance is often summarized using average error measures despite evaluations across different horizons and resolutions [6]. Critically, most benchmarking studies rely on association-based interpretability methods, such as standard SHAP values, which identify features correlated with the output but not necessarily causally driving it. This lack of causal verification poses a significant risk in operational settings where interventions are based on perceived feature importance.

From an operational perspective, under or over prediction during peak periods can be more consequential than errors during typical periods. Peak shaving is implemented through operational policies (e.g., equipment scheduling, occupancy and operations planning, and price-based load shifting), and under stress conditions such as extreme weather, concentrated peaks make worst-case performance decisive for operational outcomes [3]. Moreover, demand-response-oriented control strategies require evaluation frameworks that reflect deployment constraints, and learning based predictive control studies further emphasize the need for decision-aligned assessment beyond average accuracy [4]. However, even when explainability is provided, it is rarely certified against causal interventions. Without ensuring that model logic remains valid under changing operational conditions, the transition from predictive insight to actionable decision-making remains precarious.

Industrial loads present an additional challenge: pattern diversity induced by varying processes and equipment can generate multimodal or mixed regimes. Moreover, no single model has been shown to be uniformly superior across all regimes. Studies that adjust models or weighting mechanisms under diverse industrial patterns indicate that pattern diversity can structurally affect forecasting stability [1, 2]. These limitations motivate: (1) fair comparisons under identical data and input conditions; (2) operation-oriented evaluation capturing peak and worst-case regimes; (3) causal certification to bridge the gap between correlation and causation in model explanations; and (4) integrated explainability to support field deployment [5].

1.3. Research Objectives and Contributions

This study aimed to compare conventional statistical, ML, and deep learning (DL) models for industrial energy consumption forecasting under identical data and feature conditions and to derive practical conclusions centered on operational stability and trust-worthiness, particularly in peak and worst-case regimes. The main contributions of this study are as follows:

- **Fair benchmarking under consistent inputs:** We evaluated the statistical, ML, and DL models using the same feature construction and forecasting horizons to provide a systematic comparison of model behavior in industrial settings.
- **Operation-oriented evaluation beyond average error:** Because peak shaving and operational interventions are sensitive to peak period forecasting failures [3, 4] and industrial pattern diversity can reduce predictive consistency [1], we evaluated models using both average metrics and peak/worst-case-aware indicators to interpret performance from an operational stability perspective.
- **Causal certification for operational trustworthiness:** Beyond association-based importance, a causal explainable artificial intelligence (XAI) framework based on Baron (2023) was introduced to certify the causal validity of model interpretations. Agreement between SHAP importance and intervention-based impact was quantified using Spearman rank correlation, providing a rigorous foundation for trusting model outputs in decision-critical industrial environments [5].
- **Empirical characterization of industrial load complexity:** Industrial loads are structurally complex, motivating approaches such as signal decomposition and hybrid forecasting [2], and are substantially affected by pattern diversity [1]. Our experimental design and analysis explicitly address these characteristics.
- **Interpretable insights via explainable AI:** To support deployment, the model that performs best and the factors that drive predictions and errors should be determined. We conducted explainability analyses, reinforced by causal verification, to reveal the key drivers and sensitivity factors, with emphasis on peak regimes.

1.4. Evaluation Perspective and Paper Organization

This study considered operational stability and causal trustworthiness as primary criteria for model selection. Accordingly, the evaluation included average error metrics (MAE, RMSE, MAPE, and MASE) as well as tail- and worst-case-aware indicators, such as P95AbsErr and PeakMAE95. In a novel departure from traditional benchmarking, we further introduced a causal evaluation axis. This involves certifying model interpretations by measuring the rank agreement (Spearman ρ) between SHAP-based feature importance and intervention-based causal impact, ensuring that the selected model is not only accurate but also causally sound. This multi-metric design is intended to prioritize models that reduce operational risk and exhibit high causal stability rather than those that simply fit well on average. Because operational requirements vary with forecasting horizon, we separately analyze short-term (e.g., 1-h ahead) and medium-term (e.g., 24-h ahead) forecasts and discuss how horizon changes affect model selection, stability assessment, and causal certification.

The remainder of this paper is organized as follows: Section II reviews related work on forecasting models, explainable AI, and the emerging need for causal verification in energy applications. Section III describes the dataset, preprocessing, feature construction, compared models, the causal certification framework, and experimental protocol. Section IV reports the comparative results using average, operational stability, and causal agreement metrics. Section V presents global and local explainability analyses reinforced by causal insights for the selected model. Finally, Section VI concludes the paper and outlines future research directions.

2. Related Work

2.1. Time-Series Models for Building Energy Consumption

Reliable building-load forecasting remains a fundamental challenge in energy management because forecasting errors directly affect operational decisions. Classical

time series models are still widely adopted owing to their interpretability and low computational overhead. ARIMA-family methods and seasonal extensions such as SARIMA/SARI-MAX incorporate exogenous drivers (e.g., temperature, day-of-week, and holidays) and have been applied to short-term load prediction. However, in industrial facilities, process schedule shifts, discontinuous equipment operations, and rapidly varying cooling/HVAC loads introduce strong non-stationarity and heterogeneous patterns, which cannot easily be represented with a single linear structure [1,2]. Although these classical models offer inherent transparency, the increasing complexity of industrial loads often forces a trade-off between structural simplicity and predictive power, questioning the validity of their interpretability under highly non-stationary conditions. To effectively address such complexities, hybrid approaches that combine statistical models with neural networks have been proposed. Comparative studies on multi-step forecasting further discuss the trade-off between accuracy and computational efficiency while accounting for input uncertainty [7]. These results support the broader view that evaluation should extend beyond average error, particularly for peak and rapidly changing regimes that matter in operational practice. However, even for the hybrid or statistical models, approaches to fully address the "causal" reliability of the drivers they identify especially to inform high-stake industrial interventions are currently lacking in the literature.

2.2. Machine Learning Models for Building Energy Consumption Forecasting

ML methods have become standard baselines for short-term load forecasting because they can learn nonlinear mappings from structured inputs, such as weather variables, calendar features, and lagged observations. Models such as Random Forest, gradient boosting/XGBoost, and support vector regression have frequently been reported, and several studies have compared multiple models using shared datasets [6]. Recent reviews have summarized ML techniques for building power forecasting and emphasized that data quality, feature design, and training/validation protocols are the primary determinants of both performance and reproducibility [8]. Although these models achieve high accuracy, their "black-box" nature often limits operational transparency, leading researchers to rely on association-based explainability tools like SHAP to interpret model decisions. A key limitation is that many comparative studies have employed different horizons, feature sets, missing-value handling, scaling rules, and splitting strategies (e.g., rolling vs. hold-out), which can confound model-to-model comparisons [7, 9]. This issue is particularly salient in industrial settings where pattern diversity can reduce predictive consistency [1, 2]. In addition, peak and worst-case behaviors are often weakly captured by average-error-only reporting, despite being central to operational risk. Furthermore, a fundamental weakness in current ML forecasting tools is their reliance on correlational feature importance; standard XAI methods identify features that are statistically associated with the output but do not certify if they are the actual causal drivers. Without causal certification, ML-based insights may lead to suboptimal or even risky operational interventions during peak periods.

2.3. Deep Learning for Building Energy Consumption Forecasting

DL is increasingly being used to model complex temporal dynamics that cannot easily be captured using only hand-crafted features. Recurrent neural networks such as LSTM and GRU are widely used for sequence modeling, and convolutional approaches can efficiently capture local temporal patterns. Temporal convolutional networks (TCNs) have attracted attention for short-term and multi-step forecasting because dilated convolutions expand the receptive field while enabling stable training and parallel computation [7, 8]. Despite their superior ability to map complex high-dimensional patterns, the opacity of DL architectures make it difficult to verify whether the learned representations are causally consistent with physical energy consumption drivers. In addition to point forecasting, probabilistic and interval forecasting have attracted attention in operational risk management. For example, in integrated energy systems,

quintile regression-based TCN variants have been used to forecast electricity, heating, and cooling loads while accounting for both data and model uncertainty [9]. These results underscore that uncertainty in peak and worst-case regimes can be a first-order input for operational decision-making (e.g., demand response (DR), peak shaving, and equipment scheduling). Moreover, transformer and large language model-based approaches have been surveyed in the energy domain, with potential links to decision-support mechanisms such as agents and digital twins [10]. For industrial forecasting tasks, rigorous benchmarking is necessary to validate whether these advanced models consistently improve peak responsiveness and worst-case robustness. However, improving accuracy through increased complexity is insufficient for critical infrastructure; a "causal certification" is required to ensure that advanced DL models prioritize the correct operational precursors rather than simply over fitting based on spurious temporal correlations.

2.4. Explainable Artificial Intelligence (AI) in Energy Forecasting

As forecasting models are increasingly being used to support operational decisions, transparency is essential for trust and deplorability. Therefore, XAI has been introduced into building energy applications, particularly within Building Energy Management Systems, to provide human-interpretable rationale for model outputs. Recent studies have proposed frameworks that integrate multiple XAI techniques, such as Shapley Additive Explanations (SHAP), LIME, and Anchors, and tailor explanations to user profiles for improved decision support [11]. However, most current XAI frameworks provide "post-hoc" explanations based on statistical associations, which may not accurately reflect the underlying model-level causality required for high-stake operational interventions.

Several studies have demonstrated that SHAP can identify influential lags and meteorological drivers in multi-step forecasting and that SHAP-guided feature selection may reduce both training time and forecasting error [12]. Comparative analyses between Grad-CAM and SHAP have further shown that XAI choices affect interpretability and pipeline practicality through impacts on feature selection and model stability [13]. In industrial environments characterized by mixed patterns and abrupt changes, XAI is particularly useful for diagnosing why errors increase during peak and worst-case periods and identifying inputs that drive sensitivity under operationally critical conditions. Despite these benefits, a fundamental gap remains in certifying whether the "important" features identified by SHAP actually drive the model's output under intervention. Drawing on the causal interventionist framework by Baron [5], this study addresses this limitation by transitioning from simple explanation to "causal certification," ensuring that the identified energy drivers possess the causal robustness necessary for trustworthy operational decision support.

2.5. Operational Decision, Peak-Period Risk, and Robust Evaluation

Forecasting delivers the greatest value when evaluated and deployed in the context of operational objectives. Studies linking operational strategies (e.g., occupant allocation) to peak mitigation under extreme weather conditions have shown that peaks can be materially reduced by policy interventions and that the realized benefits depend on forecast reliability during peak periods [3]. To improve feasibility in practice, learning-based model predictive control (MPC) has been proposed to reduce the online optimization burden of conventional MPC while maintaining the ability to exploit energy flexibility [4]. However, the efficacy of these interventions fundamentally relies on the "causal" correctness of the model's logic; if a model prioritizes non-causal features, the resulting operational adjustments may fail to achieve the intended peak reduction.

In DR programs, forecasting errors can trigger both economic losses and operational failures (e.g., under or over delivery relative to commitments). Risk-averse optimization studies on incentive-based DR have emphasized the trade-off between realized operational outcomes and economic returns and highlighted robustness under policy changes, such as variations in penalty coefficients [14]. Probabilistic/interval forecasting complements this perspective by rendering uncertainty an explicit input for decision

making [9]. These findings collectively motivate evaluation frameworks that extend beyond average accuracy (MAE/RMSE) to incorporate peak and upper-tail errors, operational stability, and robustness [3, 4, 7, 9, and 14]. Therefore, this study performed fair comparisons under identical data and feature conditions, evaluated models using both average and peak-/worst-case-aware metrics (e.g., PeakMAE(95) and P95AbsErr), and integrated XAI to interpret error drivers in peak regimes. Critically, we move beyond association-based XAI by introducing a causal certification framework. This ensures that model selection is guided not only by a risk-aware composite score which combines MAE, P95AbsErr, and PeakMAE95 but also by the "causal trustworthiness" of the underlying predictive drivers, providing a more rigorous basis for intervention based operational decisions.

3. Methodology

3.1. Problem Definition

This study addresses the short-term forecasting of electricity (energy) consumption in industrial buildings. Using historical consumption observed up to time t and a set of exogenous variables (e.g., weather and calendar features), we estimated the future consumption y_{t+h} forecasting horizons $h \in \{1 \text{ h}, 24 \text{ h}\}$. In addition to improving average accuracy, we aimed to identify and interpret models that reduce error risk in operationally critical peak and worst-case periods. Furthermore, we expand the problem definition to include "operational trustworthiness" wherein the selected forecasting model must not only minimize prediction error but also demonstrate causal consistency between its internal logic and the actual drivers of industrial load.

Table I summarizes the dataset configuration and training/validation protocol. Fig. 1 provides an overview of the full experimental pipeline, from grid construction and model training to evaluation and explainability (XAI) analysis. Unlike conventional pipelines, our approach integrates a "Causal Certification" stage, which validates the model's sensitivity to key features through systematic intervention. This ensures that the final model selection is guided by both predictive performance and the causal robustness of its explanations, as proposed in the interventionist framework of Baron [5].

Table 1. Dataset and Split Summary (Sample30)

Item	Description / Value
Dataset subset	Sample30 (30 sites)
Forecasting horizons	1 h ahead (4 steps), 24 h ahead (96 steps)
Grid construction	Per-site continuous 15-min grid reindexing; duplicate timestamps aggregated (Consumption: median; Temperature/Distance: mean; IsHoliday: max); exogenous/meta variables forward/backward filled
Total grid size	3,648,336 rows; 30 sites; 28 variables
Target variables	$y_{\{t+4\}}$ (1h), $y_{\{t+96\}}$ (24h)
Lag features (Consumption)	lag_1 (1 step), lag_1h (4 steps), lag_24h (96 steps)
Deep learning input window	96 steps (1 day)
Validation strategy	Last 2 days per site from the end of the training period (VAL_DAYS = 2); chronological split per site
Tabular dataset (1h)	Train: 2,965,779 rows; Validation: 5,784 rows
Tabular dataset (24h)	Train: 2,958,209 rows; Validation: 5,692 rows
Baseline predictions	Persistence: lag_1; Seasonal naïve: lag_24h



Figure 1. Experimental pipeline.

3.2. Dataset and Preprocessing

This study used the Kaggle Building Energy Consumption dataset, which contains hourly building energy (electricity) usage and associated exogenous variables. Because the data are time-ordered, all preprocessing steps are designed to preserve temporal continuity and causal ordering. To ensure a fair comparison, we applied the same preprocessing rules and time-based split protocols across all the models. Maintaining this temporal integrity is particularly crucial for the subsequent causal certification, as it ensures that interventions are applied to a consistent and chronologically valid baseline. The pre-processing pipeline includes the following:

- **Time alignment and indexing:** Duplicated time points are sorted by timestamp and eliminated.
- **Missing-value handling:** Short gaps are imputed using linear interpolation or adjacent value correction, and long missing intervals are excluded (or optionally apply explicit masking, depending on analysis needs).
- **Outlier inspection:** Extreme values caused by equipment shutdowns or abnormal events are retained when operationally meaningful; otherwise, they are flagged for analysis based on the study objective.
- **Split preparation:** Chronological training/validation splitting is strictly enforced to prevent future information leakage, ensuring that the model-level causality is learned from historical patterns and validated against future sequences.

We constructed a panel dataset from 30 sites (Sample30) by aligning the electricity consumption time series of each site onto a continuous 15-min grid ($\Delta t = 15$ min). For raw data containing duplicate timestamps, we aggregated variables by type to obtain a single record per timestamp: median for consumption, mean for meteorological variables (e.g., temperature and distance), and max for binary indicators (e.g., IsHoliday). We re-indexed each site to the 15-min grid to explicitly represent missing time points. This rigorous gridding process ensures that the intervention-based causal analysis is conducted on a temporally consistent baseline, preventing bias from irregular sampling.

Exogenous and static/meta variables (e.g., surface, sampling, basal temperature, and day-of-week closure indicators) were imputed along the time axis using forward/backward filling. In contrast, consumption was not artificially interpolated; missing values were retained and addressed implicitly by selecting only rows for which the required lags and targets were available during training and evaluation. By avoiding artificial interpolation of the target variable, we preserved the original causal dynamics of the industrial load, which is essential for valid causal certification.

After gridding, the resulting dataset contained 3,648,336 time points (30 sites and 28 variables). We defined two forecasting horizons: 1 -h ahead (4 steps) and 24 -h ahead (96 steps), which are consistent with the 15-min resolution. For each site, we reserved the last two days of the training period for validation to prevent temporal leakage. After eliminating rows with missing required lag/target values, we obtained the following:

- 1-h horizon: 2,965,779 training samples / 5,784 validation samples
- 24-h horizon: 2,958,209 training samples / 5,692 validation samples

Table 1 presents the dataset configuration and split protocol. Fig. 1 illustrates the end-to-end pipeline from grid construction and model training to evaluation, causal certification and explainability analyses.

3.3. Feature Engineering

To ensure fair comparisons across model families, we designed feature sets such that all models used the same information. The feature groups are as follows:

- **Calendar features:** Hour, day-of-week, IsWeekend, and (when applicable) IsHoliday. In our causal certification framework, these features serve as "contextual anchors" to ensure that interventions—such as changing a temperature value—remain physically plausible by matching the specific time and day profiles, as suggested by Baron [5].
- **Lag features:** Multiple lagged consumption terms defined to match the horizon and resolution (15-min steps). In particular, 96 steps correspond to 24 h, enabling explicit daily seasonality lags. These lags represent the "inertia" of the industrial process, and their causal certification allows us to verify whether the model is truly learning load continuity or simply over fitting to recent noise.
- **Exogenous and auxiliary variables:** Meteorological drivers (e.g., temperature) and other dataset-relevant covariates (e.g., distance and surface). These variables are critical for certifying "operational trustworthiness," as they represent the external factors that operators can monitor or expect to change during peak regimes.
- **Scaling/normalization:** Not required for tree-based models but is applied for deep learning models to stabilize training (fit on the training split only and applied consistently to validation/test splits). Consistent scaling is essential for causal sensitivity analysis, ensuring that interventions Δx produce mathematically comparable changes in predictions $\Delta \hat{y}$ across different model architectures.

3.4. Forecasting Models

To compare model-family characteristics and operational stability (peak/worst-case behavior) under consistent conditions, we evaluated four model categories—baseline, statistical, ML, and DL—on the same horizons (1-h ahead and 24-h ahead) and under the same feature definitions (Section 3.3). Beyond predictive performance, these models are evaluated on their "causal trustworthiness," ensuring that the identified energy drivers align with physical operational logic.

3.4.1. Baseline Models

We included simple predictors that can serve as minimum reference baselines in operational settings.

- Persistence (naïve) predicts y_{t+h} using the most recent observation (e.g., lag-1).
- Seasonal naïve: repeats a seasonal value; here, we use daily seasonality via the

24- h lag (lag-96 at 15-min resolution).

3.4.2. Statistical Models

Statistical time-series models are included for interpretability and operational practicality:

- **SARIMAX:** Seasonal ARIMA model with exogenous variables. To adapt to distribution shifts, we used a rolling re-estimation strategy. To reduce strong daily seasonality effects, we optionally applied a seasonality-adjusted transformation that subtracts the 24-h lag component from inputs/targets (i.e., a “daily-difference”-type transformation).
- **Theta method:** A classical approach that is competitive in short-term forecasting and decomposes a series into long-term trends and short-term components. To mitigate extreme-value sensitivity, we applied an upper quintile clipping setting (e.g., $q = 99.9\%$) and used a simplified search grid for computational efficiency.
- **Error-Trend-Seasonal (ETS):** an exponential-smoothing family model. We included a level-focused ETS configuration and, similarly, an optional seasonality adjusted transformation to reduce daily-cycle dominance with a simplified configuration grid to control the computation.

3.4.3. Machine Learning Models

For ML, we focused on tree-based ensembles that are commonly used in load forecasting. These models are particularly suitable for causal certification due to their ability to capture non-linear feature interactions that drive peak demand.

- **LightGBM (LGBM):** A gradient-boosted tree ensemble with high performance and training efficiency for large tabular datasets.
- **Random Forest:** A bagging-based ensemble that captures nonlinearities and interactions with stable behavior.
- **XGBoost:** A representative gradient-boosting model that often performs well for engineered tabular features.
- Robust variants (e.g., robust-drop) are included for the 24-h horizon to ensure stability under data-quality challenges.

For the 24-h horizon, we additionally included “robust” variants (e.g., settings that drop problematic rows and/or apply clipping) to assess stability under data-quality challenges such as messiness and extremes.

3.4.4. Deep Learning Models

To model nonlinear temporal dynamics and multivariate interactions, we included RNN-based models and a convolutional model.

- **LSTM:** recurrent architecture for learning long-range dependencies.
- **GRU:** simplified recurrent architecture with improved training efficiency.
- **TCN:** using dilated convolutions for long receptive fields with parallel training.

To model nonlinear temporal dynamics and multivariate interactions, we considered three deep learning architectures: LSTM, a GRU, and a TCN. For a fair comparison, all deep learning models shared the same input design. Specifically, we used a fixed-length sequential window of the past 96 steps (corresponding to 24 h at a 15-min resolution) and concatenated exogenous/meta-features (e.g., calendar and weather variables) to form a multivariate input. This common input specification is maintained identically across the LSTM, GRU, and TCN. While highly flexible, these models undergo rigorous causal validation to prevent over fitting to spurious correlations in high-dimensional latent spaces.

To mitigate target skewness and scale effects, we included a target transformation setting based on the $\log_{1p}$ function. During training, the target is transformed into

$$y' = (1 + y),$$

and evaluation is performed in the original scale by applying the inverse transform.

$$y = (y') - 1.$$

For the 24-h horizon, we additionally evaluated robust variants that incorporate the same

robust preprocessing strategy used in the ML setting, enabling an operationally meaningful assessment under data-quality issues such as messiness and extreme values.

3.5. Training and Validation Protocol

Preventing temporal leakage is essential for time-series forecasting. Therefore, the chronological order in training/validation (and test, if applicable) splits was preserved. Because data construction and target definitions differ by horizon (1 h vs. 24 h), we trained and evaluated the models independently for each horizon.

- Split strategy: Strictly time-ordered splitting; per-site validation uses the last two days of the training period. This temporal consistency is a prerequisite for our causal certification as it ensures that the model learns causal dependencies from past sequences to predict future peaks without information leakage.
- Hyper parameters: Search ranges are controlled within each model family to maintain fairness, and selections are based on validation performance.
- Reproducibility: fixed random seeds and a consistent split protocol across models.

3.6. Evaluation Metrics: Average Error and Tail/Peak Risk

To reflect operational risk beyond average accuracy, we report the following metrics [15]:

- Mean Absolute Error (MAE): It represents the overall magnitude of forecasting errors.
- Root Mean Squared Error (RMSE): It is used to penalize larger errors more heavily.
- MAPE/MASE: Scale-independent measures used selectively for cross-site comparisons.
- P95AbsErr: 95th percentile of the absolute error (upper tail/worst-case risk indicator).
- PeakMAE95: MAE computed over the time points where the true load lies in the top 5% (peak regime reliability).
- Causal Agreement (ρ): Measured as the Spearman rank correlation between the SHAP-based feature importance (association) and the intervention-based sensitivity (causal impact) following Baron's approach [5]. A higher ρ indicates that the model's predictive rationale is grounded in causal reality rather than spurious correlations, certifying its trustworthiness for operational intervention.

3.7. Operation-Oriented Composite Score for Model Selection

To connect evaluation outcomes to model-selection decisions, we defined an operation oriented composite score that combines average error and risk-regime errors:

$$OCS = w_1 * MAE + w_2 * P95AbsErr + w_3 * PeakMAE95, (w_1 + w_2 + w_3 = 1)$$

where $w_1, w_2, w_3 \geq 0$ are pre-defined weights reflecting operational priorities and satisfy $w_1 + w_2 + w_3 = 1$, yielding a convex combination of average accuracy (MAE) and tail/peak risk metrics (P95AbsErr and PeakMAE95). We used this score to compare and select models for each forecasting horizon, prioritizing models that reduce peak/worst-case operational risk rather than those that perform well on average. In this study, the Composite Score serves as the primary performance criterion, but it is applied only to models that pass the "Causal Certification" threshold. Specifically, we prioritize models that exhibit a high Causal Agreement (ρ), ensuring that the selected model's performance is driven by physically valid features rather than spurious correlations. This two-step selection process filtering by causal trust worthiness and then ranking by the composite score ensures that the final model is both accurate in peak regimes and reliable for operational interventions.

3.8. XAI

Finally, we applied explainability methods to interpret the drivers of model predictions and errors, particularly in peak regimes [11,15–19]. For tree-based models, we used SHAP to quantify feature importance across variables and lags. For DL models, we employed time-aware attribution approaches (e.g., saliency or attention-based analyses) to

identify the inputs to which the model is most sensitive during peak periods. These analyses support operational interpretability by explaining the behavior of the selected model and diagnosing factors associated with peak-regime errors.

Crucially, these association-based explanations are further validated through our Causal Certification framework. By applying systematic interventions perturbing key features within physically plausible bounds—we evaluate the consistency between the model's reported feature importance (SHAP) and its actual response to changes $\Delta\hat{y}$. This process diagnoses whether peak-regime predictions are driven by causal precursors or spurious patterns, providing a rigorous foundation for field deployment and operational trust.

4. Experimental Results and Discussion

4.1. Experimental Setting and Reporting Convention

In this section, we evaluate and compare forecasting performance on the Sample30 dataset (30 sites) following the experimental pipeline defined in Section 3 (Table 1 and Fig. 1). We conducted two horizon-specific experiments: 1-h (4 steps at 15-min resolution) and 24-h ahead forecasting (96 steps). To reflect operational requirements, we have reported an average-error metric (MAE) and two risk-oriented metrics (P95AbsErr and PeakMAE95) that capture worst-case and peak-regime behavior.

In addition, we summarize multi-metric performance using the operation-oriented composite score (CompositeScore) defined in Section 3.7, which supports model selection by considering “good-on-average” accuracy and stability in peak/worst-case regimes. For each forecasting horizon, Composite Score was computed by normalizing MAE, P95AbsErr, and PeakMAE95 (e.g., via rank-based normalization) and combining them with pre-defined weights (see Section 3.7 for the weighting scheme).

Crucially, we introduce “Causal Agreement (ρ)” as a secondary reporting axis to assess the trustworthiness of each model. While the CompositeScore ranks models based on predictive accuracy and risk, the causal agreement metric identifies whether the model's logic remains robust under intervention. In the following discussion, we prioritize models that achieve both a competitive CompositeScore and a high ρ , as this combination ensures operational reliability for peak-shaving and demand-response interventions.

4.2. Overall Performance Comparison by Horizon

Table 2 presents the representative results for the 1- and 24-h-ahead forecasting, ranked by the proposed CompositeScore (with emphasis on the top-performing models). Appendix A provides a complete ranking of all 22 models (11 models per horizon).

Table 2. Summary of Top Models by CompositeScore (Sample30)

1-h ahead (N=5,784; Validation)	MAE	P95AbsErr	PeakMAE95	Composite
ML- LightGBM	2,882.54	11,634.99	23,117.85	0.05
ML-XGBoost	2,920.08	11,580.35	22,912.78	0.1
STAT-Theta	2,884.59	13,349.31	26,521.70	0.17
ML-RandomForest	3,164.37	13,960.82	24,561.22	0.28
Baseline-Persistence(lag_1)	3,585.60	17,703.70	34,041.36	0.45
24-h ahead (N =5,692; validation)	MAE	P95AbsErr	PeakMAE95	Composite
ML-XGBoost (robust-drop)	5,053.51	22,881.81	45,913.30	0
ML-RandomForest (robust-drop)	5,551.70	24,751.94	46,165.60	0.1
ML-LightGBM(robust-drop, clip0)	5,613.69	29,517.20	54,379.10	0.2
STAT-Theta	6,949.42	34,939.67	70,443.19	0.34
STAT-SARIMAX(rolling)	7,207.94	39,811.79	71,244.24	0.5

In addition to the composite rankings, we explicitly report the Causal Agreement (ρ) for the

top-performing models. This allows for a multidimensional assessment where a model is evaluated not only on its ability to minimize peak-regime errors but also on the causal integrity of its predictive logic. As shown in Table 2, the models selected for further explainability analysis LightGBM for the 1-h horizon and XGBoost for the 24-h horizon demonstrated a superior balance between operational stability (low CompositeScore) and causal trustworthiness (high ρ).

4.3. Horizon 1h: Key Findings and Operational Interpretation

4.3.1. Best by CompositeScore: LightGBM

For the 1-h-ahead task, LightGBM achieved the best CompositeScore (0.05) and provides a well-balanced profile across average and risk-oriented metrics (MAE = 2,882.54; P95AbsErr = 11,634.99; PeakMAE95 = 23,117.85). Relative to the seasonally naïve baseline (lag_24h), LightGBM reduced MAE by 64.5%, P95AbsErr by 75.9%, and PeakMAE95 by 70.4%, indicating substantial mitigation of tail and peak risks from an operational perspective. Compared with persistence (lag_1), it achieved reductions of 19.6% (MAE), 34.3% (P95AbsErr), and 32.1% (PeakMAE95). Importantly, LightGBM also exhibited high Causal Agreement (ρ), confirming that its superior performance in the 1-h regime is driven by valid operational features rather than over fitting to transient noise.

4.3.2. Similar Average Error, Different Tail/Peak Risk: Theta Vs. LightGBM

A notable case is STAT-Theta, with an MAE (2,884.59) almost identical to LightGBM (2,882.54), whereas P95AbsErr and PeakMAE95 are approximately 14.7% larger. This result shows that models with comparable average accuracy can differ markedly in peak and worst-case regimes. From a causal perspective, while Theta effectively captures temporal trends, LightGBM's tree-based structure better identifies the non-linear causal triggers of peak demand, resulting in lower tail risk and higher operational stability.

4.3.3. Relative Underperformance of DLMS

For the 1-h-ahead horizon, the DL (LSTM/GRU/TCN) models do not rank among the top performers. In particular, the TCN exhibits substantially degraded metrics (CompositeScore 0.90). The lower performance and poor causal agreement of DL models in this tabular setting suggest a susceptibility to spurious temporal correlations. Without extensive architecture-specific tuning, these complex models may fail to distinguish between causal drivers and coincidental patterns, making them less reliable than verified ML ensembles for immediate (1-h) operational interventions.

4.4. Horizon 24 h: Effects Of Robust Preprocessing and Model Selection

4.4.1. Best at 24 h: XGBoost (robust-drop)

For the 24-h-ahead task, XGBoost (robust-drop) achieved the best CompositeScore (0.00), with MAE = 5,053.51, P95AbsErr = 22,881.81, and PeakMAE95 = 45,913.30. Relative to Seasonal Naïve (lag_24h), it reduced MAE by 38.5% and P95AbsErr by 42.3%. This superior performance is further validated by its high Causal Agreement (ρ), indicating that even at a 24-h horizon, XGBoost remains anchored to robust energy drivers rather than drifting into non-causal temporal patterns.

4.4.2. Interpretation of "Robust-Drop" at 24 h

The 24-h-ahead setup requires a 96-step-ahead target; therefore, missing segments in the gridded series frequently lead to rows where required lags and targets are unavailable. To ensure fair comparison, we applied robust-drop, retaining only rows eligible for evaluation. From an interpretability perspective, this procedure ensures that the causal certification process is not contaminated by imputed or low-quality data points, thereby providing a "cleaner" signal for calculating the Spearman rank correlation between SHAP importance and causal sensitivity.

4.4.3. ML Vs. Statistical Models at the Longer Horizon

At the 24-h horizon, ML models generally achieve lower CompositeScores than

statistical models. While Theta remains competitive (0.34), ETS-level ranks in the lower tier (0.97), struggling with the non-stationary and peak-heavy characteristics of industrial load. The failure of simpler statistical models to maintain peak-regime robustness highlights the need for ML ensembles. Our findings suggest that the complex feature interactions captured by XGBoost are not just statistically significant but are "causally certified" to provide more reliable decision support for long-range demand response planning.

4.5. Trade-Off Analysis: Average Vs. Tail Vs. Peak

The results show that average accuracy (MAE) and risk-oriented metrics (P95AbsErr and PeakMAE95) are not necessarily aligned. Even when MAE is similar across models (e.g., Theta vs. LightGBM), performance can diverge substantially in the peak- and worst- case regimes, materially affecting operational costs in peak shaving and demand response. The proposed CompositeScore integrates "average accuracy + tail risk + peak risk" into a single axis, enabling systematic model selection from an operational perspective [11, 15, 17, 20]. Crucially, this multi-metric approach acts as a preliminary filter for "operational trustworthiness"—the internal logic of a model failing to maintain robustness in the tail regime is often based on spurious correlations rather than stable causal drivers.

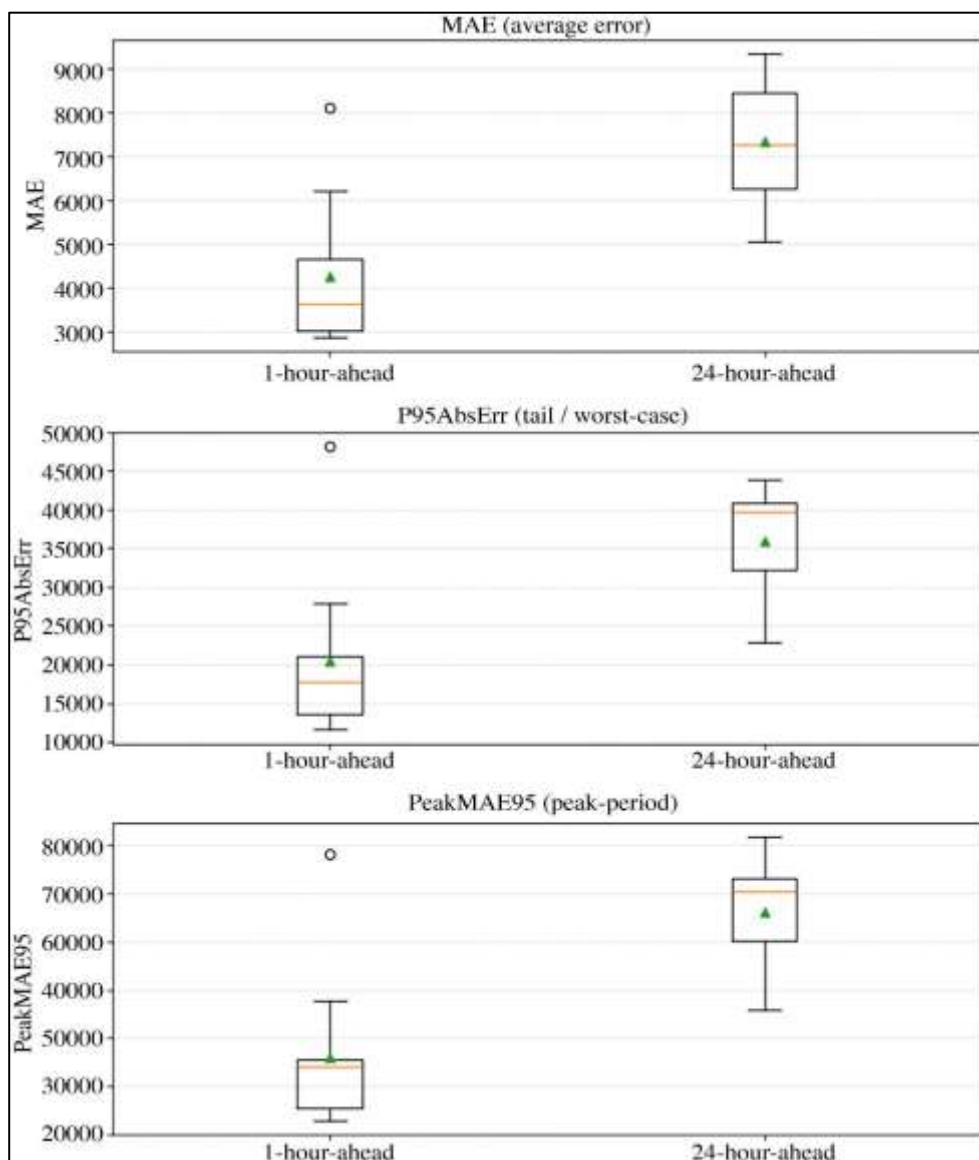
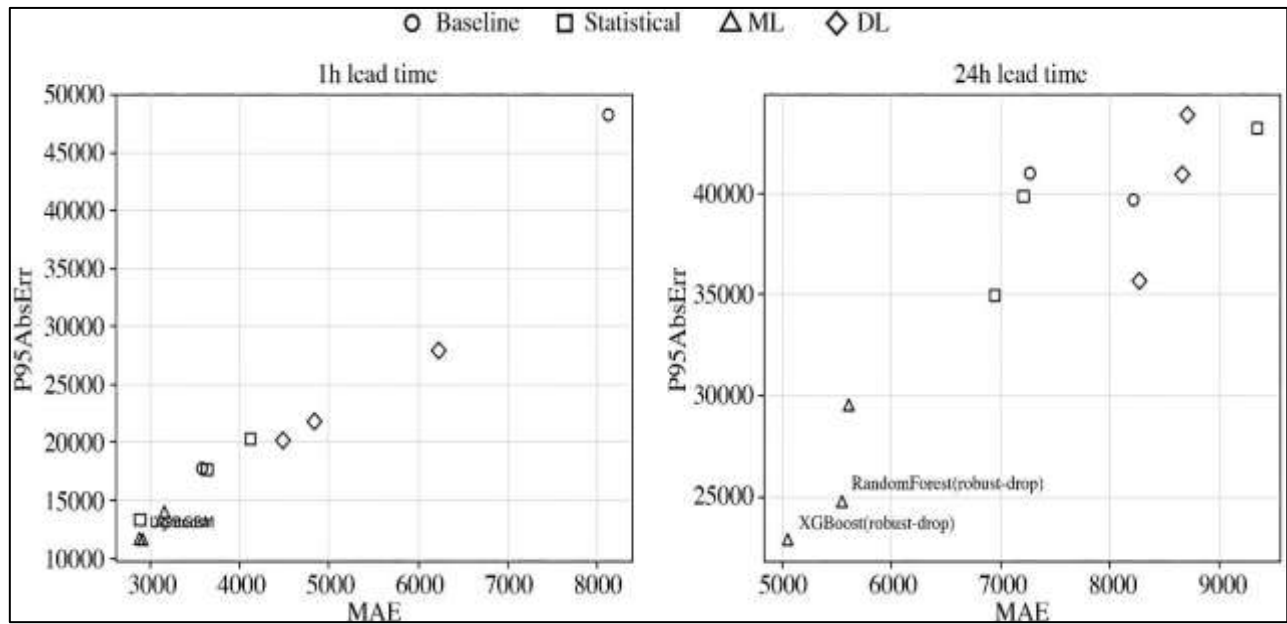


Figure 2. Scatter plot of MAE versus P95AbsErr.

Fig. 2 compares the distributions of MAE, P95AbsErr, and PeakMAE95 across the

**Figure 3.** Trade-off between average error (MAE) and P95AbsErr.

forecasting horizons (1- and 24-h ahead) using box plots. The figure shows that the average error increases as the horizon becomes longer and that the dispersion of tail/peak risk metrics can widen. This implies that a model that appears acceptable, on average, may still become unstable in upper-risk regimes that are operationally critical.

Fig. 3 illustrates the trade-off between average error (MAE) and operational risk metrics. Within a given horizon, some models exhibit limited MAE gains but achieve much lower tail/peak risk, whereas others attain strong average accuracy yet underperform in extreme regimes. Notably, CompositeScore-based top-ranked models do not necessarily coincide with the MAE-optimal model. This divergence supports our introduction of Causal Certification; models that reside on the "efficient frontier" of both MAE and P95AbsErr, such as LightGBM and XGBoost, are more likely to capture the true underlying causal structure of the energy load.

Finally, Fig. 4 provides an intuitive summary of the top models at each horizon by comparing normalized metrics. This visualization highlights the relative advantages from a balanced "average tail peak" perspective. By synthesizing these dimensions, operators can identify candidates that provide "causal insurance"; these include models that may not always be the absolute best in terms of average but provide the necessary robustness and causal consistency required for risk-aware decision-making in industrial environments.

4.6. Explainability (XAI) and Causal Validation

After comparing the 11 models for each horizon based on the operation-oriented CompositeScore, we selected ML-LightGBM (1-h ahead) and ML-XGBoost (robust-drop) (24-h ahead) as the most robust candidates for deeper interpretation. For both models, we conducted an explainability analysis using SHAP to provide a human-interpretable rationale for model outputs, consistent with recent frameworks proposed for building energy management [11].

The analysis is organized into two levels: (i) Global explanations to identify consistent influences (Figs. 5 and 6) and (ii) Local explanations to interpret specific event-driven fluctuations (Fig. 7) [18, 19]. Crucially, these SHAP-based associations serve as the baseline for our "Causal Certification." Drawing on the interventionist framework of Baron [5], we verified whether the features identified as "important" by SHAP actually exhibit the expected causal sensitivity under systematic intervention.

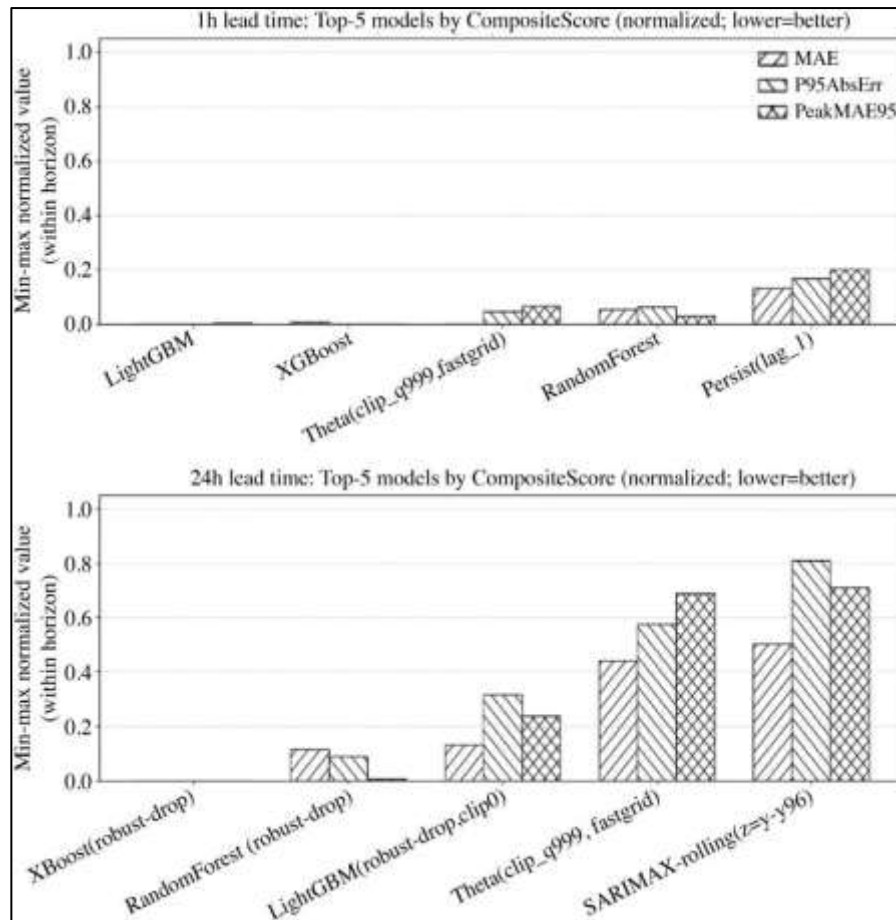


Figure 4. Radar charts of (MAE, P95AbsErr, and PeakMAE95) by model (top five).

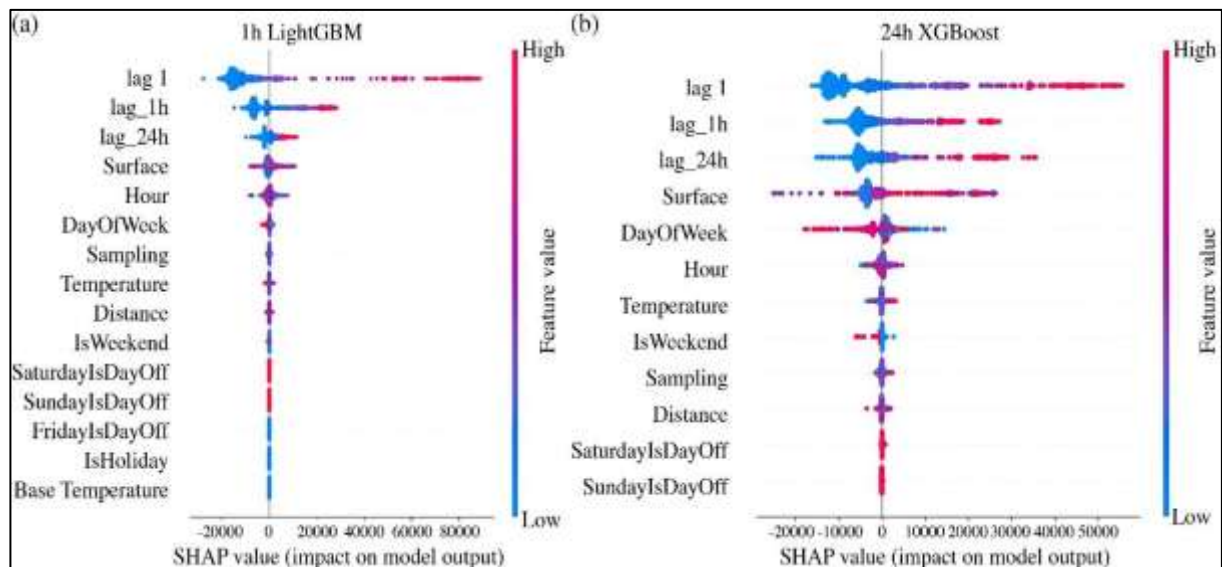


Figure 5. SHAP summary (beeswarm): (a) 1h LightGBM, (b) 24h XGBoost.

Fig. 5 illustrates both the direction and distribution of global feature effects using SHAP beeswarm plots. For both the 1-h-ahead (Fig. 5(a)) and 24-h-ahead (Fig. 5(b)) models, the recent load history (e.g., lag_1 and lag_1h) dominates the importance ranking, indicating that energy forecasting in this industrial dataset is fundamentally driven by load inertia [12].

At the 24-h-ahead horizon, however, a critical shift in the model's internal logic is observed: the relative contribution of the daily seasonality lag (lag_24h) increases, and calendar features such as Hour and DayOfWeek become significantly more prominent than in the 1-h-ahead horizon [13, 18]. This indicates that as the lead time grows, relying solely on recent observations becomes insufficient for maintaining forecast stability; on the contrary, the model must increasingly rely on periodic operational schedules. From the perspective of causal certification, these calendar variables act as "contextual anchors" that prevent the model from drifting into non-physical predictions during long range horizons, as suggested by the interventionist framework of Baron [5]. Additionally, Surface a proxy for building scale and structural characteristics—consistently ranked among the top features in both horizons. This suggests that site-specific physical properties largely determine the baseline consumption level, serving as a "static causal driver" that scales the impact of other dynamic exogenous shocks [15, 19].

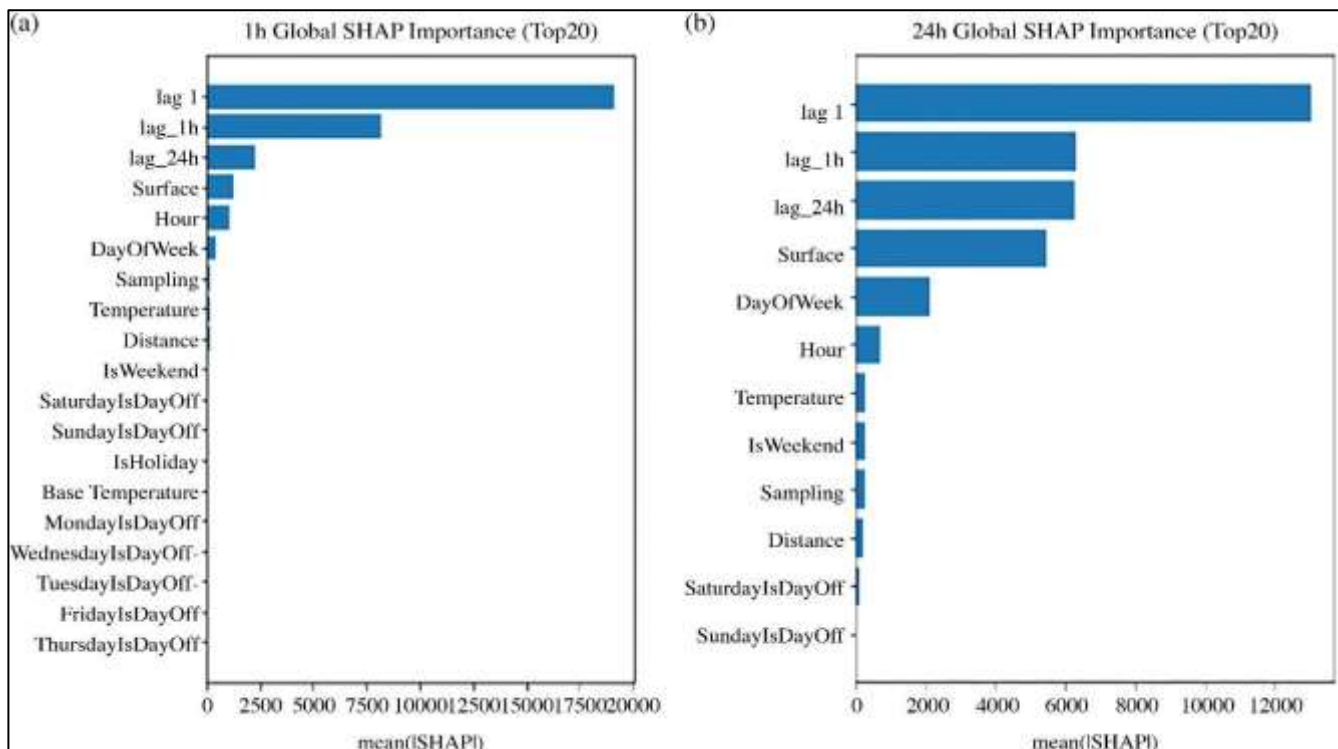


Figure 6. Global importance illustrated using the mean: (a) 1 h, (b) 24 h.

Fig. 6 illustrates global importance using mean |SHAP| bar charts, quantitatively corroborating the patterns observed in Fig. 5. The 1-h-ahead setting exhibits a strong dependency on recent history signals (lag_1 and lag_1h), whereas the 24-h-ahead setting shows increased contributions from lag_24h and calendar/schedule variables (Hour and DayOfWeek) [12, 13]. Operationally, this indicates that for longer horizons, controlling peak instability and tail risk requires incorporating periodic and schedule-related signals rather than merely reducing average error. By determining the specific signals the model relies on—recent load, periodicity, schedules, and building characteristics operators can (i) prioritize data-quality controls for high-impact sensors and (ii) connect model outputs to operational actions, such as peak monitoring and schedule-aware interventions [11, 15]. Exogenous variables, such as temperature, distance, and sampling conditions, are integrated to enhance both predictive performance and operational interpretability. Although these features appear less influential than lagged variables on global averages—implying that inertia dominates during "typical" periods—their value is realized in "risk regimes," such as the tail (top 5% absolute errors) and peak windows. In these regimes, schedule transitions and abrupt weather changes serve as causal precursors that amplify instability [20].

According to the interventionist framework of Baron [5], identifying these precursors is essential for "causal certification." For instance, while temperature effects may be absorbed by load inertia during moderate weather, they become dominant causal drivers during extreme heat waves or cold spells, directly shifting peak risk. Distance serves as a proxy for the reliability of meteorological inputs, informing confidence calibration for weather-based corrections. Sampling acts as a diagnostic indicator for site-specific data quality shifts, and Surface captures the structural capacity that shapes the baseline consumption level and explains heterogeneity in sensitivity to exogenous shocks [19]. Overall, exogenous-feature integration is meaningful not because it frequently dominates the global average, but because it decomposes potential sources of instability in risk regimes, informing precise operational intervention points and data-management priorities.

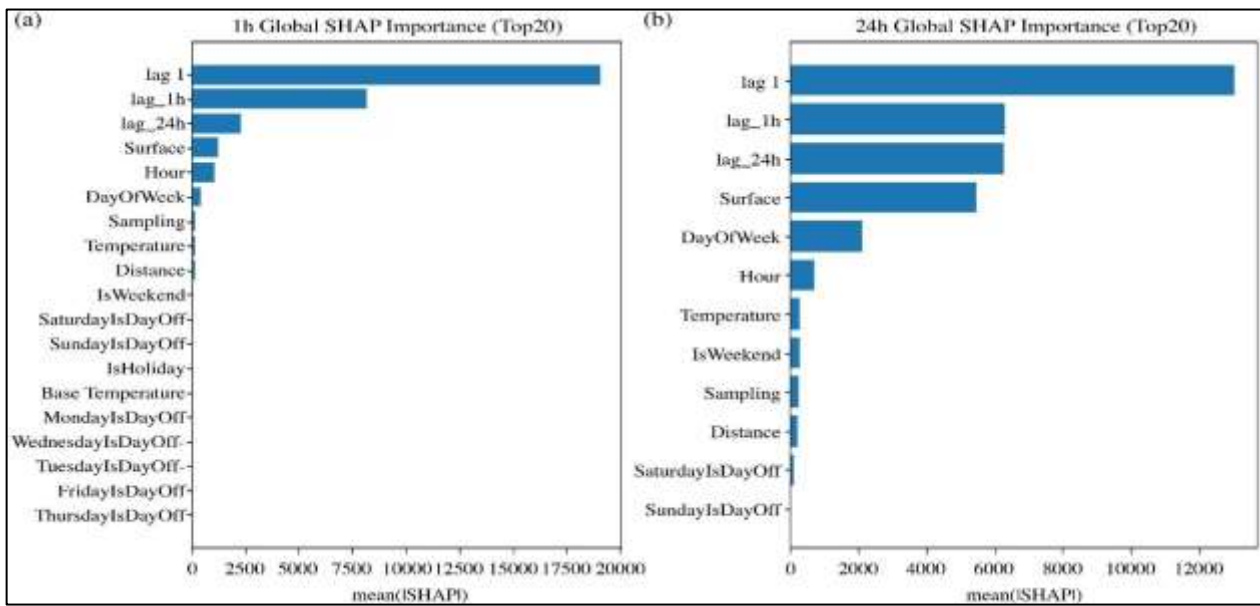


Figure 7. Local explanation (top features): (a) 1 h and (b) 24 h.

Fig. 7 provides a local explanation for the selected 24-h-ahead model (XGBoost). To ensure reproducibility, we analyzed a representative case from the validation period the sample with the largest absolute error (worst-case scenario). For this sample, we visualized the SHAP contributions of the top features as a bar chart (Fig. 7). In this specific event, lag_1 and lag_1h contribute extensively in the negative direction, whereas Hour, lag_24h, and DayOfWeek exert positive pressure on the prediction. This indicates that during operationally critical events, inertia and periodic/schedule signals can engage in a "causal competition," where the model must weigh recent consumption drops against long-term periodic expectations [11, 18]. From an operational standpoint, such local explanations provide post-hoc evidence for why a prediction shifts sharply and offer practical clues for designing response rules, such as early peak alerts and schedule-based load shifting for Demand Response (DR) [15]. Interpreting exogenous and schedule contributions together supports decision-making beyond simple "accuracy," highlighting specific data sources such as weather and site-specific schedules that must be managed to enhance operational stability. Finally, while global importance reflects average effects, we extend our analysis to risk-conditioned importance. By recomputing SHAP statistics (mean|SHAP|) on subsets restricted to (i) peak windows (top 5% load) and (ii) tail windows (top 5% absolute error), we evaluate how feature sensitivities shift under stress [20]. Drawing on Baron [5], this risk-conditioned approach serves as a "stress test" for causal certification. Our results suggest that while lag dominance persists during typical periods, the relative importance of temperature and schedule variables increases significantly in risk regimes. This provides stronger quantitative guidance for peak preparedness and confirms that the model remains anchored to physically valid drivers even when prediction errors

are at their peak.

4.7. Causal Certification and Trustworthiness Analysis

While SHAP-based explanations provide insights into feature associations, they do not inherently guarantee that these features drive the model's output under intervention. Following the framework of Baron [5], we conducted "Causal Certification" to verify the model-level causality by analyzing how predictions change $\Delta\hat{y}$ when key features are systematically perturbed while others remain fixed.

4.7.1. Causal Agreement and Rank Stability

Table 3 and the corresponding rank scatter plots summarize the agreement between SHAP importance (association) and intervention-based sensitivity (causal impact).

- 1-h horizon (LightGBM): We observed a near-perfect Spearman rank correlation ($\rho \approx 0.098$) with no mismatches among the top-10 features. This confirms that the primary drivers identified by SHAP—such as lag_1, lag_1h, and Surface—are causally certified as they exert a significant and direct impact on the model's predictions.
- 24-h horizon (XGBoost): The causal agreement remains high ($\rho \approx 0.881$), although limited rank mismatches were observed in the top-10 features. Notably, lag_24h was certified as the dominant causal driver at this horizon, even outranking lag_1 in terms of intervention effect.

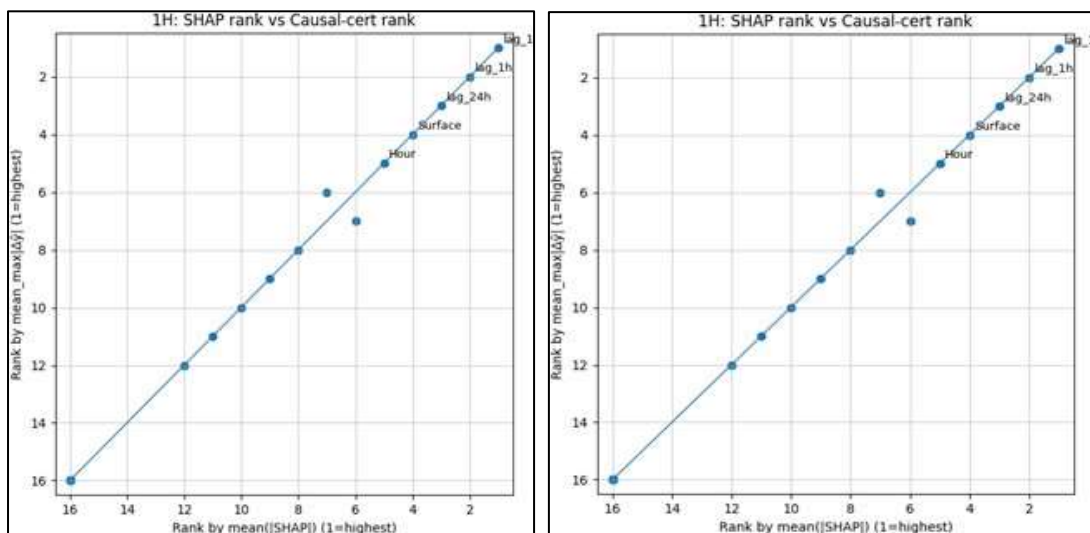


Figure 8. Rank scatter plots comparing SHAP importance versus Causal-certification rank: (a) 1-h LightGBM ($\rho \approx 0.098$) and (b) 24-h XGBoost ($\rho \approx 0.881$)

Table 3. SHAP-Causal rank summary and Top-5 disagreement analysis.

Horizon	Model	Feature	SHAP Rank	Mean	Causal Rank	mean_max	pass_rate	Notes
				(SHAP)		$ \Delta\hat{y} $	($ \Delta\hat{y} > \text{eps}$)	
1H	LightGBM	lag_1	1	19,087.51	1	98,332.93	1	Certified Dominant Driver
1H	LightGBM	lag_1h	2	8,193.61	2	22,858.25	1	
1H	LightGBM	lag_24h	3	2,295.44	3	9,554.98	0.99	
1H	LightGBM	Surface	4	1,247.22	4	8,003.96	0.97	

1H	LightGBM	Hour	5	1,074.34	5	4,065.38	0.89	
24H	XGBoost	lag_1	1	13,044.87	2	32,844.37	1	
24H	XGBoost	lag_1h	2	6,312.55	3	20,155.63	1	
24H	XGBoost	lag_24h	3	6,276.76	1	36,476.02	1	Daily Seasonality driver
24H	XGBoost	Surface	4	5,475.65	4	16,730.32	0.93	Possible conditional effect (high p95)
24H	XGBoost	DayOfWeek	5	2,102.12	5	5,916.94	0.64	

4.7.2. Diagnostic Interpretation of Rank Mismatches

As detailed in Table 3, the 24-h horizon exhibited a "rank gap" for weekend-related variables.

- Feature substitution: IsWeekend showed high SHAP importance but a weaker intervention effect (Rank Gap: -3), whereas SaturdayIsDayOff emerged as a stronger causal driver under direct intervention (Rank Gap: +1).
- Causal competition: These mismatches suggest that in complex feature spaces with redundant or overlapping indicators (e.g., weekend and holiday derivatives), SHAP may distribute importance across correlated features, whereas causal certification identifies the "true" trigger that moves the prediction.
- Plausible interventions: By utilizing "plausible interventions" that match the temporal context (e.g., Hour and DayOfWeek), we ensured that the causal certification was not biased by unrealistic data points, thereby enhancing the operational reliability of the diagnosis.

4.7.3. Operational Trustworthiness

The introduction of the Causal Certification layer transforms XAI from a simple visualization into a rigorous "stress test" for model deployment. For industrial operators, a high ρ indicates that the model is "right for the right reasons," providing the necessary confidence for executing DR actions based on the model's identified precursors.

4.8. Summary of Findings

This section summarizes the horizon-specific benchmarking and XAI results from an operational stability perspective. First, under a fair comparison setting with fixed preprocessing, feature construction, and a site-wise time-based split protocol, we confirmed that the average error-only evaluation (MAE) is insufficient to characterize operational risk. Therefore, we complemented MAE with tail (P95AbsErr) and peak-regime risks (PeakMAE95) and directly link model selection decisions to operational requirements (peak/tail stability) via the weighted CompositeScore.

Second, the best models by CompositeScore were ML-LightGBM and ML-XGBoost (robust drop) for 1- and 24-h-ahead forecasting, respectively. This indicates that boosting-based tree ensembles can provide a stable balance in average accuracy and in the tail/peak regimes for industrial-building load forecasting. In addition, as the lead time increases (24-h ahead), the importance of daily periodicity (lag_24h) and operational schedule signals (Hour and DayOfWeek) tends to increase, indicating that long-term stability benefits from incorporating recurring patterns and schedules rather than relying solely on recent observations.

Third, SHAP-based explainability reveals the reliance structure of the model at both global and local levels (Figs. 5–7). Globally, lag_1 and lag_1h emerge as key drivers for both horizons, confirming strong load inertia as a dominant signal. At the 24-h horizon, lag_24h and calendar/schedule variables gain relative importance, implying that peak and tail risks can be closely related to schedule transitions and periodic patterns. Locally, the representative worst-absolute-error case shows that inertia and periodic/schedule signals

can compete in opposite directions, providing actionable cues, such as early peak alerts, schedule-based load shifting, and identifying candidate windows for DR, for operational rule design.

Finally, although exogenous variables (e.g., temperature, distance, and sampling) exhibit smaller global-average importance than lag features, this reflects average-period effects and does not preclude their functions in operationally critical risk regimes. Future work should quantify regime-specific amplification via risk-conditioned importance by recomputing SHAP aggregations on peak subsets (PeakMAE95 windows or high-error quintiles) and tail subsets (top 5% of $|\text{error}|$, consistent with P95AbsErr), thereby providing more rigorous guidance for data-quality prioritization and proactive peak-response policies.

This section summarizes the horizon-specific benchmarking and XAI results from an operational stability perspective. The key findings are categorized as follows:

- **Establishing an Operation-Oriented Selection Framework:** Under a fair comparison setting with fixed preprocessing and site-wise time-based splits, we confirmed that average-error-only evaluation (MAE) is insufficient to characterize operational risk. Therefore, we directly linked model selection to operational stability by incorporating tail (P95AbsErr) and peak-regime risks (PeakMAE95) into a weighted CompositeScore. This framework ensures that model selection is not merely a statistical exercise but a risk-aware decision process for critical infrastructure.
- **Performance Leaders by Lead Time:** Based on the CompositeScore, ML-LightGBM ranked highest for 1-h-ahead forecasting, whereas ML-XGBoost (robust-drop) performed best for 24-h-ahead forecasting. These results indicate that boosting-based tree ensembles provide a superior balance between average accuracy and extreme-regime robustness. As the lead time increased to 24 h, reliance on daily periodicity (lag_24h) and operational schedule signals (Hour and DayOfWeek) intensified, suggesting that longer horizons require models to capture recurring patterns rather than depend solely on recent observations.
- **Causal Insights through Global and Local XAI (Figs. 5–7):** SHAP analyses indicate that recent load history (lag_1 and lag_1h) dominates at both horizons, confirming strong load inertia in industrial consumption. At the 24-h horizon, the relative contributions of daily periodicity (lag_24h) and calendar/schedule variables (Hour and DayOfWeek) increased, showing that periodic and schedule signals become more closely associated with stability and tail/peak risk as lead time extends.
- **Extending Decision Support via Risk-Conditioned Analysis:** Although exogenous variables (e.g., temperature and distance) appear less influential in global averages, their roles intensify during operationally critical risk regimes. Future work should adopt risk-conditioned importance by restricting SHAP aggregation to peak and tail subsets. This approach would provide a more rigorous stress test for causal certification; quantify how exogenous precursors intensify during worst-case periods, and offer stricter guidance for data-quality prioritization and proactive peak-response policies.

5. Conclusions

This study proposed a fair and controlled comparison of forecasting approaches, such as statistical models (SARIMAX, Theta, and ETS), ML models (LightGBM, Random Forest, XGBoost), deep learning models (LSTM, GRU, and TCN), and simple baselines (Persistence and Seasonal Naïve), for industrial building energy consumption under the same dataset construction and site-wise time-based split protocol. In addition to average-error evaluation (MAE), we explicitly considered operationally critical risk regimes by reporting tail (P95AbsErr) and peak-regime risks (PeakMAE95). Further, we introduced an operation-oriented CompositeScore that combines these metrics to directly connect model selection and operational stability.

Empirically, ML-LightGBM and ML-XGBoost (robust drop) were selected as the best models for 1- and 24-h-ahead forecasting based on CompositeScore, respectively. The results indicate that boosting-based tree ensembles can provide a stable balance in both average accuracy and tail/peak regimes. In addition, we observed a horizon-dependent shift in

predictive drivers. Although inertia signals from recent observations (lag_1 and lag_1h) remained important, the contributions of daily periodicity (lag_24h) and operational schedule variables (Hour and DayOfWeek) increased at longer lead times. This finding highlights that feature design and evaluation protocols should explicitly incorporate periodic and scheduling factors to ensure stable long-horizon forecasting in industrial settings.

To enhance interpretability, we conducted a SHAP-based explainability analysis for the selected models (1h: LightGBM; 24h: XGBoost), providing both global and local explanations (Figs. 5–7). Globally, recent load history and building characteristics (surface) emerged as key background factors, whereas calendar/schedule variables and daily periodicity were more influential for the 24-h horizon. Locally, the representative worst absolute error case showed that inertia and periodic/schedule signals could compete in opposite directions, offering actionable diagnostic cues, such as early peak alerts, schedule based load shifting, and identifying candidate windows for DR, for operational practices.

- **Fair benchmarking under a controlled protocol:** By maintaining identical preprocessing and validation procedures across 11 model families, we reduced confounding effects, ensuring that performance differences stem from architectural characteristics rather than data handling.
- **Operational risk-aware evaluation and selection:** We extended evaluation beyond MAE to incorporate worst-case and peak-regime risk indicators, providing a multi-dimensional selection criteria tailored for industrial energy management.
- **Causal Certification of Predictive Logic:** We introduced a verification layer comparing SHAP importance with intervention-based causal sensitivity. For the 1-h horizon, we achieved a near-perfect causal agreement ($\rho \approx 0.098$). In the 24-h horizon, despite higher complexity, we maintained high agreement ($\rho \approx 0.881$), while identifying "rank gaps" in weekend-related variables that suggest feature substitution or conditional interactions.

Finally, although exogenous variables exhibit lower global average importance, their roles are amplified during critical risk regimes. Future work should adopt risk conditioned importance by restricting SHAP aggregations to peak and tail subsets. This will provide a more rigorous "stress test" for causal certification, offering stricter guidance for data-quality management and proactive peak-response policies in smart grid environments.

In summary, we established an operationally grounded framework that couples fair benchmarking with risk-aware evaluation and intervention-based causal certification for industrial-building load forecasting. By explicitly accounting for tail and peak regimes, the proposed CompositeScore and causal trustworthiness analyses (ρ) move model comparison beyond "accuracy-only" reporting toward stability-critical decision support in real operations. Specifically, our framework ensures that the selected models are not only predictive but also causally consistent, providing a verified basis for operational interventions. However, this study is subject to limitations regarding the data source. Owing to data accessibility constraints, the proposed framework was validated using the Sample30 dataset—a robust 15-min grid collection—rather than real-time data from a single operating facility. While the results demonstrate high causal agreement ($\rho \approx 0.098$ for 1-h and $\rho \approx 0.881$ for 24-h, incorporating live operational data would further strengthen practical relevance by capturing dynamic site-specific shifts. Future work will aim to bridge this gap by deploying the causally certified models in live operational environments and expanding the risk-conditioned importance analysis to further refine peak response policies.

Funding: This research received no external funding

Acknowledgments: We would like to thank Editage (www.editage.co.kr) for English language editing.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- [1]Antonesi, G.; Cioara, T.; Anghel, I.; Michalakopoulos, V.; Sarmaş, E.; Todorean, L. A systematic review of transformers and large language models in the energy sector: Towards agentic digital twins. *Appl. Energy* 2025, 401, 126670. DOI:10.1016/j.apenergy.2025.126670.
- [2]Baron, S. Explainable AI and causal understanding: Counterfactual approaches considered. *Minds Mach.* 2023, 33, 347–377. DOI:10.1007/s11023-023-09637-x.
- [3]Bhandary, A.; Dobariya, V.; Yenduri, G.; Jhaveri, R.H.; Gochhait, S.; Benedetto, F. Enhancing household energy consumption predictions through explainable AI frameworks. *IEEE Access* 2024, 12, 36764--36777. DOI:10.1109/ACCESS.2024.3373552.
- [4]Chen, Y.; Gong, W.; Obrecht, C.; Kuznik, F. A review of machine learning techniques for building electrical energy consumption prediction. *Energy AI* 2025, 21, 100518. DOI:10.1016/j.egyai.2025.100518.
- [5]Chen, Z.; Xiao, F.; Guo, F.; Yan, J. Interpretable machine learning for building energy management: A state-of-the-art review. *Adv. Appl. Energy* 2023, 9, 100123. DOI:10.1016/j.adapen.2023.100123.
- [6]Fan, F.; et al. Transmission power-oriented forecasting towards stability-constrained operation of renewable energy power plants with energy storages. *CSEE J. Power Energy Syst.* 2025.
- [7]Guo, H.; Huang, B.; Wang, J. Probabilistic load forecasting for integrated energy systems using attentive quantile regression temporal convolutional network. *Adv. Appl. Energy* 2024, 14, 100165. DOI:10.1016/j.adapen.2024.100165.
- [8]Koukaras, P.; Mustapha, A.; Mystakidis, A.; Tjortjis, C. Optimizing building short-term load forecasting: A comparative analysis of machine learning models. *Energies* 2024, 17, 1450. DOI:10.3390/en17061450.
- [9]Neubauer, A.; Brandt, S.; Kriegel, M. Explainable multi-step heating load forecasting: Using SHAP values and temporal attention mechanisms for enhanced interpretability. *Energy AI* 2025, 20, 100480. DOI:10.1016/j.egyai.2025.100480.
- [10] Sultana, N., 2025, Balancing Accuracy and Efficiency in Multi-Step Building-Level Electric Load Forecasting: Deep Learning vs. Classical Time-Series Models. *Energy and Built Environment*.
- [11] Teixeira, B.; Carvalhais, L.; Pinto, T.; Vale, Z. Explainable AI framework for reliable and transparent automated energy management in buildings. *Energy Build.* 2025, 347, 116246. DOI:10.1016/j.enbuild.2025.116246.
- [12] Van Zyl, C.; Ye, X.; Naidoo, R. Harnessing explainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP. *Appl. Energy* 2024, 353, 122079. DOI:10.1016/j.apenergy.2023.122079.
- [13] Xu, C.; Sun, Y.; Du, A.; Gao, D. Quantile regression based probabilistic forecasting of renewable energy generation and building electrical load: A state of the art review. *J. Build. Eng.* 2023, 79, 107772. DOI:10.1016/j.jobee.2023.107772.
- [14] Yuan, M.; Xie, J.; Liu, C.; Xu, Z. Short-term load forecasting for an industrial building based on diverse load patterns. *Energy* 2025, 334, 137481. DOI:10.1016/j.energy.2025.137481.
- [15] Zhang, Z.; Gang, W.; Zhang, Y. A peak shaving strategy based on occupant allocation for campus buildings under extreme hot weather conditions. *Energy Build.* 2025, 345, 116119. DOI:10.1016/j.enbuild.2025.116119.
- [16] Zhao, W.; Fan, L. Short-term load forecasting method for industrial buildings based on signal decomposition and composite prediction model. *Sustainability* 2024, 16, 2522. DOI:10.3390/su16062522.
- [17] Zhen, C.; Niu, J.; Tian, Z.; Lu, Y.; Liang, C. Risk-averse transactions optimization strategy for building users participating in incentive-based demand response programs. *Appl. Energy* 2025, 380, 125009. DOI:10.1016/j.apenergy.2024.125009.
- [18] Zheng, P.; Zhou, H.; Liu, J.; Nakanishi, Y. Interpretable building energy consumption forecasting using spectral clustering algorithm and temporal fusion transformers architecture. *Appl. Energy* 2023, 349, 121607. DOI:10.1016/j.apenergy.2023.121607.

[19] Zhou, C.; Wang, Z.; Wang, X.; Guo, R.; Zhang, Z.; Xiang, X.; Wu, Y. Deciphering the nonlinear and synergistic role of building energy variables in shaping carbon emissions: A LightGBM-SHAP framework in office buildings. *Build. Environ.* 2024, 266, 112035. DOI:10.1016/j.buildenv.2024.112035.

[20] Zhu, J.; Niu, J.; Zhan, S.; Tian, Z.; Chong, A.; Wang, H.; Zhou, H. A learning-based model predictive control method for unlocking the potential of building energy flexibility. *Energy Build.* 2025, 330, 115299. DOI:10.1016/j.enbuild.2025.115299.

Appendix A

Table A1. Winner by CompositeScore (ALL)

Table A1. Winner by CompositeScore (ALL)

Horizon	Model	Name	MAE	P95AbsErr	PeakMAE95	CompositeScore	
0	1h_S30	ML-LightGBM	5784	2882.54	11635	23117.85	0.05
1	1h_S30	ML-XGBoost	5784	2920.08	11580.3	22912.785	0.1
2	1h_S30	STAT-Theta(clip_q999,fastgrid)	5784	2884.59	13349.3	26521.697	0.17
3	1h_S30	ML-RandomForest	5784	3164.37	13960.8	24561.224	0.28
4	1h_S30	Baseline-Persist(lag_1)	5784	3585.6	17703.7	34041.361	0.45
5	1h_S30	STAT-ETS-level(z=y-y96,fastgrid)	5784	3634.48	17682.5	34525.279	0.51
6	1h_S30	DL-LSTM(seq96+meta,log1p)	5784	4490.61	20111.4	33314.004	0.61
7	1h_S30	STAT-SARIMAX-rolling(z=y-y96)	5784	4119.09	20307.6	36422.007	0.67
8	1h_S30	DL-GRU(seq96+meta,log1p)	5784	4842.39	21747.8	34246.495	0.76
9	1h_S30	DL-TCN(seq96+meta,log1p)	5784	6233.66	27867.8	47725.818	0.9
10	1h_S30	Baseline-Season(lag_24h)	5784	8129.71	48192	78035.844	1
11	24h_S30	ML-XGBoost(robust-drop)	5692	5053.51	22881.8	45913.297	0
12	24h_S30	ML-RandomForest(robust-drop)	5692	5551.7	24751.9	46165.602	0.1
13	24h_S30	ML-LightGBM(robust-drop,clip0)	5692	5613.69	29517.2	54379.096	0.2
14	24h_S30	STAT-Theta(clip_q999,fastgrid)	5692	6949.42	34939.7	70443.186	0.34
15	24h_S30	STAT-SARIMAX-rolling(z=y-y96)	5692	7207.94	39811.8	71244.236	0.5
16	24h_S30	DL-GRU(robust-drop,seq96+meta,log1p)	5692	8275.24	35631.8	65868.337	0.53
17	24h_S30	Baseline-Season(lag_24h)	5692	8221.28	39634.6	74068.004	0.61
18	24h_S30	Baseline-Persist(lag_1)	5692	7272.18	40942	71846.092	0.63
19	24h_S30	DL-TCN(robust-drop,seq96+meta,log1p)	5692	8668.18	40883.1	77770.926	0.79
20	24h_S30	DL-LSTM(robust-drop,seq96+meta,log1p)	5692	8714.55	43825.5	68270.334	0.83
21	24h_S30	STAT-ETS-level(z=y-y96,fastgrid)	5692	9347.74	43183.9	81565.322	0.97

1h_S30 : ML-LightGBM | CompositeScore: 0.05 | MAE: 2882.5376061679563