

Received 11 May 2026, accepted 4 June 2026, date of publication 9 June 2026, date of current version 15 June 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3701803

RESEARCH ARTICLE

Market-Aware Asymmetric Loss Framework for Day-Ahead Building Energy Demand Forecasting in the Electricity Market

HYEONGYU SON¹, (Member, IEEE), MUSU KIM¹, (Member, IEEE),
YOUNGGYU HWANG¹, (Member, IEEE), AND SEKYUNG HAN², (Member, IEEE)

¹Department of Electrical and Electronic Engineering, Kyungpook National University, Daegu 41566, South Korea

²Department of Electrical Engineering, Kyungpook National University, Daegu 41566, South Korea

Corresponding author: Sekyung Han (skhan@knu.ac.kr)

This work was supported in part by the BK21 FOUR Project funded by the Ministry of Education, South Korea, under Grant 4199990113966.

ABSTRACT In electricity markets with two-settlement structures, forecast errors carry asymmetric economic consequences: under-prediction forces costly real-time procurement at prices 15-50% above day-ahead levels during peak hours, while over-prediction incurs milder opportunity costs of 4-6% of day-ahead prices. Despite this fundamental asymmetry, most building energy forecasting models optimize symmetric metrics such as MAPE and RMSE producing statistically unbiased but economically suboptimal predictions. Moreover, the degree of cost asymmetry is not constant but varies dramatically with time of day, yet existing asymmetric approaches apply fixed penalty ratios uniformly across all hours. This paper proposes a Market-Aware Asymmetric Loss (MAL) framework that bridges this gap between forecasting training objectives and market settlement realities. The core innovation is a Time-Dependent Quantile (TD- τ) mechanism that adapts prediction conservativeness hourly based on empirically-derived cost asymmetry ratios: $\tau = 0.70$ during on-peak hours when under-prediction is 3.5 times costlier, relaxing to $\tau = 0.55$ during off-peak hours when costs are nearly symmetric. The framework further integrates building pattern classification for strategy selection, a seven-component SMP simulator validated against 8,760 hourly KPX observations across 12 statistical metrics (mean values within 2%, distributional properties within 8%), and comprehensive uncertainty quantification. As a proof-of-concept study, experiments are conducted on two Korean buildings with contrasting consumption profiles (a high-variability commercial office and a regular-pattern residential complex) using 3-fold time-series cross-validation. The framework's principal contribution is a substantial reduction in under-prediction settlement cost—the economically dominant error mode: up to 50.4% on the commercial building and 41.9% on the residential building in controlled ablation against a symmetric (RMSE-only) baseline. Relative to the symmetric baseline, the full framework further yields up to 16.3% total MAPE improvement and 28.5% total settlement cost reduction for the high-variability commercial building (single-split ablation), and a 3.6% aggregate MAPE reduction in cross-validation. Statistical significance of the full-framework versus symmetric-baseline comparison is confirmed by the Diebold-Mariano test ($DM = -4.37$, $p < 0.001$). Against a well-tuned fixed-quantile baseline, TD- τ achieves comparable aggregate accuracy (MAPE differences under one percent in both buildings); its added value is theoretical consistency with the Granger-optimal quantile result, per-hour interpretability, and automatic adaptability to structural changes in the hourly cost-asymmetry ratio. The benefit is strongly moderated by building type: substantial on high-variability commercial buildings and marginal on stable residential patterns, providing deployment guidance for practitioners.

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang¹.

• **INDEX TERMS** Building energy demand forecasting, asymmetric loss function, time-dependent quantile regression, Korean electricity market, system marginal price, gradient boosting, uncertainty quantification, Diebold-Mariano test.

I. INTRODUCTION

Buildings account for approximately 40% of global final energy consumption and one-third of energy related carbon dioxide emissions [1], positioning the building sector as a critical frontier for sustainable energy management. The proliferation of advanced metering infrastructure and data-driven technologies has catalyzed a paradigm shift in how buildings interact with energy systems [2], [3]: buildings are evolving from passive energy consumers into active market participants capable of demand response, load aggregation, and strategic energy procurement [4], [5], [6]. At the core of this transformation lies accurate day-ahead demand forecasting, which determines how much electricity a building operator contracts in the forward market. Any deviation between the forecast and actual consumption must be reconciled through real time settlement mechanisms, often at prices that diverge substantially from the contracted rate. Thus, forecasting quality directly translates into economic outcomes for building operators. Building energy demand forecasting has progressed substantially over the past two decades [7], [8]. Early approaches based on ARIMA, regression, and change-point models provided interpretable but limited predictions constrained by linearity assumptions. The advent of machine learning particularly ensemble tree-based methods such as Random Forests, XGBoost, and CatBoost delivered marked improvements in capturing the nonlinear, multi modal consumption patterns characteristic of buildings [9], [10]. Deep learning architectures including LSTM networks and Transformers have also been explored for load pattern recognition and clustering [11], though empirical evidence from major forecasting competitions and comparative studies indicates they do not consistently outperform well-tuned gradient boosting on typical building energy datasets with fewer than 50,000 samples [12]. Despite this methodological richness, a fundamental gap persists: nearly all forecasting models are trained using symmetric loss functions such as MSE or MAE, which treat over prediction and under-prediction errors identically. While mathematically convenient and statistically unbiased, this symmetry assumption is fundamentally misaligned with the economic realities of electricity market settlement, where the costs of forecast errors are inherently directional and time-varying.

The Korean electricity market, managed by the Korea Power Exchange (KPX), exemplifies this misalignment [13], [14]. As a cost-based mandatory pool with a two-settlement structure, participants submit day-ahead consumption bids 12-36 hours prior to delivery, settled at the day-ahead System Marginal Price (SMP). Deviations are reconciled through real-time balancing at prices reflecting the marginal

cost of incremental supply adjustments [15], [16]. When actual consumption exceeds the forecast (under-prediction), the shortfall must be procured at real-time prices that average 15-50% above day-ahead SMP during peak hours (14:00-17:00), and can spike to 200-400% during supply scarcity events [17], [18], [19], [20]. These price spikes exhibit temporal clustering, regime persistence, and heavy-tailed distributions, making the expected cost of under-prediction substantially higher than simple averages suggest. When actual consumption falls below the forecast (over-prediction), the participant forfeits the opportunity cost of unused contracted capacity, typically 4-6% of the day-ahead price [21] non-negligible, but far more predictable and moderate. The consequence is stark: a 10kW under-prediction during a peak hour costs $3.5\times$ more than an equivalent over-prediction, and this cost ratio fluctuates from approximately 1.2:1 during overnight off-peak hours to 3.5:1 or higher during afternoon peaks. The theoretical foundations for addressing cost asymmetry in forecasting are well established. Granger [22] formally proved that under asymmetric loss, the optimal point forecast is not the conditional mean but the $\tau^* = R/(1+R)$ quantile of the predictive distribution, where R is the cost ratio between under- and over-prediction. Christoffersen and Diebold [23] further demonstrated that optimal prediction under asymmetric loss requires explicit knowledge of the loss structure, and that forecasts optimized for symmetric criteria are generally sub-optimal when evaluated under the true asymmetric cost. The mathematical apparatus for quantile estimation was provided by Koenker and Bassett [24] through quantile regression using the pinball loss, which is distribution free, robust to outliers, and naturally accommodates heteroskedasticity properties well-suited to energy applications. Despite this mature theoretical foundation, its systematic integration into building energy demand forecasting remains limited. The few studies that do incorporate asymmetric loss in wind power forecasting [25], solar radiation prediction [26], and contract capacity optimization [27] apply a fixed penalty ratio uniformly across all hours, ignoring the dramatic diurnal variation in cost asymmetry documented above. A static quantile that is appropriately conservative for on-peak hours wastes accuracy during off-peak periods; conversely, one tuned for off-peak conditions dangerously under-hedges during economically critical peak hours. Additional research gaps compound this limitation. First, evaluating the economic value of asymmetric forecasting requires realistic electricity price scenarios, yet many studies assume constant prices or simplistic distributions [28], [29], ignoring the heavy tails, Poisson-like spikes, regime-dependent volatility, and

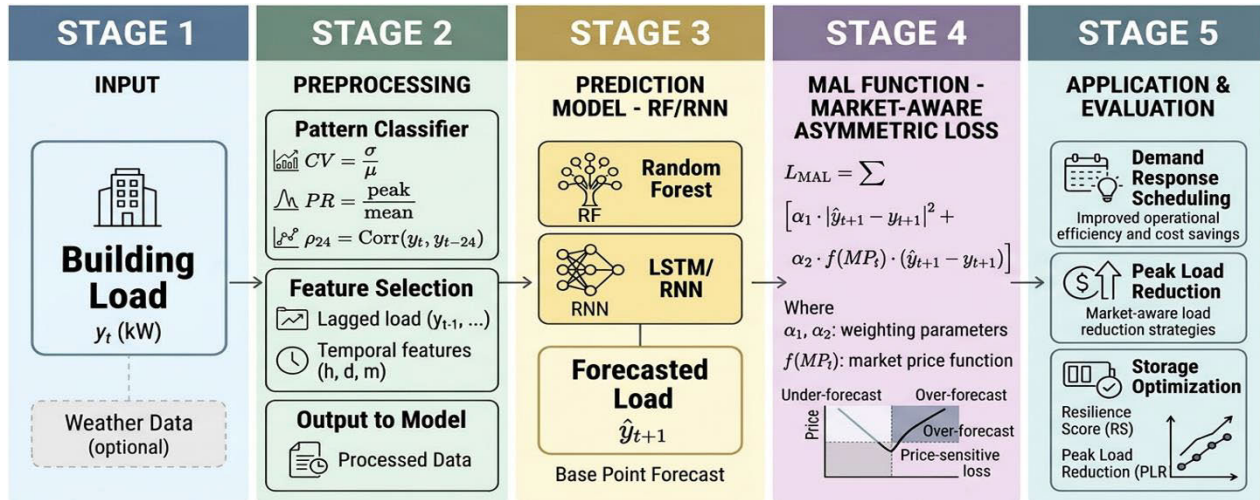


FIGURE 1. Overall architecture of the proposed Market-Aware Asymmetric Loss (MAL) framework, illustrating the pipeline from raw building load data through preprocessing (building-pattern classification, feature engineering, SMP simulation), core TD- τ modeling, to economic and uncertainty evaluation. Key message: the pricing-risk layer (SMP simulator $\rightarrow$ per-hour cost-asymmetry ratio R_t) and the learning layer (CatBoost quantile regression at $\tau(h)$) are coupled only through the scalar $\tau(h)$ schedule, which is what makes the framework re-calibratable to market drift by re-querying $\tau(h)$ at inference—without retraining the underlying quantile models.

positive skewness characteristic of real electricity prices [30], [31]. Second, beyond point forecasts, market participants require probabilistic predictions for reserve margin sizing, risk assessment, and regulatory compliance [32], [33], yet systematic probabilistic evaluation using proper scoring metrics remains rare in the building energy forecasting literature [34]. Third, most studies report average accuracy metrics without rigorous statistical significance testing, making it difficult to distinguish genuine improvements from random variation inherent in limited building energy datasets.

This paper proposes a Market-Aware Asymmetric Loss (MAL) framework that bridges the gap between forecasting training objectives and market settlement realities. The specific contributions are fourfold:

- 1) A Time-Dependent Quantile (TD- τ) mechanism that adapts prediction conservativeness by hour of day targeting $\tau = 0.70$ during on-peak hours (where $R \approx 3.5$) and relaxing to $\tau = 0.55$ off peak (where $R \approx 1.2$) theoretically grounded in the Granger-Koenker optimal quantile result and empirically validated through settlement-cost decomposition rather than aggregate MAPE.
- 2) An empirically-validated SMP simulator with seven components (diurnal baseline, seasonal/weekend adjustments, Markov regimes, Poisson spikes, GARCH volatility, skew-normal noise) validated against 8,760 hourly KPX observations across 12 statistical metrics, including RT-DA spread accuracy within 7% across all four time periods.
- 3) A building pattern classification system using coefficient of variation (CV), peak ratio (PR), and lag-24 autocorrelation (ρ_{24}) that serves two distinct roles: (i) a descriptive taxonomy of building load behavior,

and (ii) a set of generalization features for unified cross-building deployment, yielding a 7.3 percentage improvement when a single model is deployed across heterogeneous buildings.

- 4) A comprehensive evaluation framework spanning settlement cost decomposition, CVaR95, prediction interval metrics (PICP, PINAW, CWC, Winkler Score), and Diebold-Mariano testing with HAC variance estimation, providing a reproducible template for economic evaluation of building forecasting systems.
- Existing studies that incorporate asymmetric loss in power or energy forecasting invariably apply a single, fixed penalty ratio (or equivalently a single quantile τ) uniformly across all operating hours. We relax this assumption in a principled way: instead of tuning one global τ , we derive an hour-indexed schedule $\tau_t^* = R_t / (1 + R_t)$ directly from the empirically observed cost-asymmetry ratio at hour t , exactly as prescribed by Granger's optimal quantile result. The resulting schedule is not a heuristic but the pointwise solution of the same optimality condition that motivates fixed- τ approaches. This distinction is conceptual and operational rather than merely numerical: TD- τ inherits the aggregate accuracy of a well-tuned fixed- τ baseline while providing per-hour interpretability, automatic adaptation to structural shifts in R_t , and, most importantly, a consistent reduction in the economically dominant under-prediction cost component.

As a proof-of-concept, experiments on two Korean buildings with contrasting consumption patterns—a high-variability commercial office (Type A) and a regular-pattern residential complex (Type B)—demonstrate that the full MAL framework reduces under-prediction settlement cost by up to 50.4% on the commercial building and 41.9% on the residential building in controlled ablation against a symmetric (RMSE-only) baseline. Additional gains on the commercial

building include a 16.3% MAPE reduction, a 28.5% total settlement-cost reduction, and a 3.6% aggregate MAPE reduction in 3-fold cross-validation. Statistical significance of the full-framework versus symmetric-baseline comparison is confirmed by the Diebold-Mariano test ($DM = -4.37$, $p < 0.001$). When compared to a well-tuned fixed- $\tau = 0.65$ baseline, TD- τ is statistically indistinguishable in aggregate MAPE; its contribution in this regime is theoretical consistency, interpretability, and robustness rather than headline accuracy. The benefit is strongly moderated by building type, providing actionable deployment guidance: TD- τ is recommended wherever $CV > 0.20$ and $PR > 2.0$, and is expected to yield only marginal gains on stable residential patterns.

II. RELATED WORK

A. BUILDING ENERGY DEMAND FORECASTING METHODS

Building energy demand forecasting has evolved from classical statistical methods to machine learning and deep learning approaches [35], [36].

1) STATISTICAL METHODS

ARIMA/SARIMA models capture temporal dependencies through autoregressive and moving average components [37], [38], offering interpretability but limited by linearity assumptions [39]. Regularized regression (LASSO, ridge) handles high dimensional input spaces [40], while change-point models address piecewise-linear relationships. Despite their theoretical elegance, statistical methods struggle with the multi-modal, non-stationary patterns typical of building energy data.

2) MACHINE LEARNING METHODS

Ensemble tree-based methods consistently outperform statistical approaches on nonlinear building energy patterns. Random Forests [41] reduce variance through bagging, while Gradient Boosting Machines achieve state-of-the-art performance through sequential error correction [42]. Among modern implementations, XGBoost [43] introduced regularized objectives; LightGBM [44] improved efficiency via GOSS and leaf wise growth; and CatBoost [45] addressed prediction shift through ordered boosting with native quantile regression support making it particularly suitable for time-series applications. Comparative studies [46] consistently show gradient boosting outperforms other approaches when dataset sizes are below 50,000-100,000 samples, motivating our selection of CatBoost as the base learner.

3) DEEP LEARNING METHODS

LSTM networks [47] capture temporal dependencies through gating mechanisms, Transformer architectures [48] model long-range dependencies via multi-head self-attention, and TFT [49] and TCN [50] combine attention with temporal convolutions. However, empirical comparisons frequently show that deep learning does not outperform well-tuned

gradient boosting on building energy data of typical sizes ($< 50,000$ samples) [51], consistent with M4/M5 competition findings [52], [53].

4) RECENT RELATIVE RESEARCH

A newer generation of time-series Transformers aims to close this gap through architectural specialization: Patch-TST [54] segments the series into non-overlapping patches and treats each patch as a token; Times-Net [55] reshapes the 1-D series into 2-D tensors indexed by periodicity to apply convolutional inductive biases; i-Transformer [56] inverts the standard transformer by treating variates as the attention axis. In parallel, a complementary line of work integrates quantile or cost-oriented losses directly into deep architectures: MLP-based deep quantile regression has been evaluated on large-scale Korean community-center data with $R^2 = 0.36$ at the 91–99 quantile [57]; and value-oriented forecasting [58] trains network parameters to minimize expected market operating cost rather than a statistical loss. Our framework is orthogonal to this architectural line: the TD- τ mechanism is expressed purely at the loss layer and could in principle be bolted onto Patch-TST, Times-Net, or i-Transformer. For the dataset scale considered here ($\sim 10^4$ samples) we adopt CatBoost as the base learner because gradient boosting continues to dominate on small-to-medium tabular data; benchmarking TD- τ on top of the newer DL backbones on larger datasets is an explicit extension. Accordingly, we make no claim of empirical superiority over these architectures; they are discussed to position TD- τ conceptually, not to assert a performance advantage.

B. ASYMMETRIC LOSS FUNCTION IN FORECASTING

The theoretical foundations of asymmetric loss trace to Granger's seminal work, which established that under asymmetric loss, the optimal point forecast differs from the conditional expectation [59]. For piecewise linear asymmetric loss parameterized by $\alpha \in (0,1)$:

$$L_{\alpha}(e) = \begin{cases} \alpha \cdot e, & \text{if } e \geq 0 \\ -(1 - \alpha) \cdot e, & \text{if } e < 0 \end{cases} \quad (1)$$

the optimal forecast is the α -quantile of the conditional distribution: $\hat{y}_{\alpha}^* = F_{y/x}^{-1}(\alpha)$. Koenker and Bassett [34] introduced quantile regression using the pinball loss $\rho_{\tau}(e) = \max(\tau e, (\tau - 1)e)$, which is distribution-free, robust to outliers, and naturally handles heteroskedasticity properties attractive for energy applications. Modern gradient boosting libraries natively support pinball loss objectives, enabling nonlinear quantile regression central to our methodology. Asymmetric loss has been applied in electricity price forecasting [60] and probabilistic load forecasting [61], [62]. However, these applications typically use fixed asymmetry parameters applied uniformly across all hours. Very recent work has begun to move beyond fixed asymmetry in specific directions: value-oriented forecasting uses piecewise-linear costs derived from the market-clearing LP but applies them

uniformly across time; node-level financial optimization [63] introduces spatial cost heterogeneity with a feedback mechanism; and deep quantile regression for building benchmarking targets the upper tail of the demand distribution. None of these, however, ties the asymmetry parameter to the hourly cost-asymmetry ratio derived from the two-settlement rule, which is the specific gap TD- τ closes. The explicit integration of market settlement rules into training objectives—particularly with time-varying asymmetry that adapts to the diurnal pattern of cost ratios—remains limited, which our work addresses. More broadly, the challenge of producing predictors that remain accurate when deployment conditions differ from training conditions has motivated a parallel line of work on out-of-distribution (OOD) generalization and domain-invariant representation learning. Recent advances include shrinkage-based state-space networks with OOD data augmentation for zero-shot fault detection in rotating machinery [64] and category-invariant disentangled feature learning for domain generalization in machine fault diagnosis [65], both of which explicitly decouple task-relevant structure from nuisance domain variation. Our setting is complementary rather than competing: instead of learning a representation that is invariant to the operating regime, we retain a standard supervised regressor and inject regime-awareness directly into the training loss through the hour-indexed quantile τ_t^* . The two lines of work can therefore be combined—an invariant-feature backbone feeding a regime-aware asymmetric head—which we identify as a promising direction for future work.

C. PROBABILISTIC FORECASTING AND UNCERT. QUANTIFICATION

Point forecasts provide incomplete information for risk aware decision-making [66]. Market participants need probabilistic forecasts for reserve margin determination, risk assessment, and regulatory compliance. Multiple quantiles can construct prediction intervals, with each quantile estimated via pinball loss [67]. The GEFComseries [68] established standard evaluation practices we adopt. Conformal prediction [69] provides distribution-free coverage guarantees. Proper evaluation of probabilistic forecasts requires assessing both calibration and sharpness through PICP (coverage probability), PINAW (interval width), CWC (combined criterion with under-coverage penalty), and Winkler Score (proper scoring rule) [70].

D. KOREAN ELECTRICITY MARKET STRUCTURE

The Korean electricity market operates as a cost-based mandatory pool managed by KPX [71]. The day-ahead market determines dispatch schedules and SMP based on forecasted load; the real-time market balances deviations at prices reflecting marginal costs of incremental adjustments. Empirical SMP analysis [72], [73] reveals: strong diurnal patterns (peak at 14:00-17:00 in summer), seasonal variation (20-35% higher in summer/winter), price spikes in 1-2%

of hours reaching 2-4x typical levels, approximately 15% weekend discount, and regime persistence. Imbalance settlement follows:

$$C = Q_{DA} \cdot S_{DA} + (Q_{RT} - Q_{DA})^+ \cdot S_{RT} - (Q_{RT} - Q_{DA})^+ \cdot S_{DA} \cdot \gamma \quad (2)$$

where QDA and QRT denote day-ahead and real-time quantities, SDA and SRT are the corresponding SMP values, $(x)^+ = \max(0, x)$, and $\gamma < 1$ reflects the reduced value of over-contracted capacity. The first term represents base procurement cost, the second captures the premium paid for real-time shortfalls at volatile RT prices, and the third represents the discounted credit for over-contracted capacity. This asymmetric structure where under-procurement faces volatile RT prices while over-procurement faces discounted DA prices directly motivates our market-aware approach.

III. PROPOSED METHODOLOGY

A. FRAMEWORK OVERVIEW

Fig. 1 presents the architecture of the proposed Market Aware Asymmetric Loss (MAL) framework. The pipeline flows from raw building load data through preprocessing, core modeling, and evaluation. Preprocessing comprises three modules: (1) Building Pattern Classifier derives Type A/B/C from CV, PR, and ρ_{24} , enabling strategy selection and cross-building features; (2) Feature Engineering generates 22 day-ahead-safe features across temporal, lag, rolling, and pattern categories; (3) SMP Simulator produces validated price scenarios against which economic metrics are computed. The core model (4) Time Dependent Quantile Model trains Cat Boost with pinball loss and hour-specific $\tau(h)$, producing the Market-Optimal Day-Ahead Forecast. The evaluation layer includes (5) Economic Evaluator for settlement cost, CVaR, and risk-adjusted metrics; (6) Uncertainty Quantifier for PICP, PINAW, CWC, and Winkler Score; and Statistical Significance testing via the Diebold-Mariano test with HAC variance.

B. BUILDING PATTERN CLASSIFICATION

Different building types exhibit fundamentally different energy consumption patterns warranting differentiated forecasting strategies. We develop an automated classification system based on three complementary metrics. The coefficient of variation $CV = \sigma_y/\mu_y$ provides a normalized, scale-independent measure of demand variability; higher CV indicates more volatile demand patterns inherently more challenging to predict. The peak ratio $PR = \max(y_t/\mu_y)$ quantifies peak demand magnitude relative to average demand; high PR (>2.0) indicates significant peaking behavior warranting conservative forecasting. The lag-24 autocorrelation ρ_{24} measures the persistence of daily patterns; high ρ_{24} (>0.80) indicates stable, repeatable daily profiles where lag features are particularly effective. Table 1 presents the classification criteria. For the two buildings

in our study: KEPRI (Commercial Office: CV=0.239, PR= 2.31, $\rho_{24} = 0.72$) is classified as Type A with higher base quantile ($\tau_{\text{base}} = 0.65$) and more conservative peak-hour treatment; SEJONG (Residential: CV=0.195, PR=1.89, $\rho_{24} = 0.85$) is classified as Type B with moderate base quantile ($\tau_{\text{base}} = 0.60$) and balanced treatment across hours. The pattern metrics serve two distinct functions that must be clearly separated. (i) Descriptive taxonomy. CV, PR, and ρ_{24} define the Type A/B (and prospective Type C) labels used in Table 1, which in turn drive strategy selection—the base quantile τ_{base} and the on-peak conservative adjustment. This role operates at build time and produces a single label per building; it requires no re-computation at inference. (ii) Cross-building generalization features. The same three numerical values are also exposed to the learner as per-sample covariates. In single-building training these values are constant and therefore carry no signal. In a unified model trained jointly over buildings with differing (CV, PR, ρ_{24}), the same covariates become discriminative and induce building-type-specific adjustments, yielding a 7.3% overall MAPE improvement in unified deployment. This dual role—taxonomy for strategy selection and numerical covariates for cross-building transfer—explains the seemingly contradictory observations of zero single-building importance and substantial unified-model benefit.

TABLE 1. Building classification and strategy selection (*Type C defined for completeness; validation planned for future work).

Type	CV	PR	ρ_{24}	Strategy
A	>0.20	>2.0	Var.	Conservative τ , expanded features
B	≤ 0.20	≤ 2.0	>0.80	Balanced τ , standard features
C*	Var	> 2.5	<0.70	Schedule-aware

C. FEATURE ENGINEERING PIPELINE

Our pipeline generates 22 features across four categories, all strictly day-ahead compatible (only information available at decision time $t-24$ or earlier is utilized). Many published studies inadvertently violate this constraint by using short-lag features, producing artificially inflated accuracy. Our strict adherence ensures realistic, deployable results.

1) TEMPORAL FEATURE [11]

Sine-cosine cyclical encoding for hour ($h/24$), day-of-week ($d/7$), and month ($(m-1)/12$) preserves temporal proximity (6 features); raw integers enable tree-based hour/day-specific splits (3 features); binary indicators for weekend and peak hour (14:00-17:00) distinguish operationally distinct regimes (2 features).

2) LAG FEATURE [4]

y_{t-24} (same hour yesterday, most important single predictor), y_{t-48} , y_{t-72} , and y_{t-168} (same hour one week ago, capturing weekly periodicity critical for commercial buildings with weekday/weekend differentiation). All lags $\geq 24h$ to respect the day-ahead constraint.

3) ROLLING STATISTICS FEATURE [4]

24-hour rolling mean and standard deviation (window ending at $t-24$), 168-hour rolling mean, and same-hour 7-day average. These capture recent baseline levels and demand variability while maintaining strict day-ahead safety.

4) PATTERN FEATURE [3]

Building-level CV, PR, and ρ_{24} remain constant for a given building but enable cross building adaptation in unified models.

D. REALISTIC SMP SIMULATOR

In operational forecasting, future prices are unobservable at decision time. Historical prices provide only a single realization, insufficient for robust cost-aware optimization and tail risk assessment. We develop a seven-component simulator calibrated against publicly available KPX Market Information System data and validated against 8,760 hourly observations from 2023.

1) COMPONENTS AND CALIBRATION

The simulator com bines: (1) an hourly baseline $\bar{S}(h)$ ranging from 65 KRW/kWh (night) to 148 KRW/kWh (on-peak at 14:00-17:00), reflecting merit-order dispatch; (2) seasonal factors ($\times 1.35$ summer, $\times 1.25$ winter, capturing cooling/heating load impacts); (3) a weekend discount ($\times 0.85$, consistent with 14.2% empirical discount); (4) a three-state Markov chain ($S = \{\text{Normal, High, Low}\}$) with high diagonal entries (0.80-0.92) capturing regime persistence and multipliers $\psi(\text{regime}) \in 1.00, 1.25, 0.80$; (5) Poisson spike process with time-varying intensity (3.5% during peak-hours, 0.5% off-peak) and truncated log-normal magnitudes in $[1.5, 4.5] \times$; (6) GARCH(1,1) volatility clustering:

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \quad (3)$$

With $\omega = 0.0001$, $\alpha = 0.15$, $\beta = 0.80$ (persistence=0.95); and (7) skew-normal noise ($\alpha \text{SN} = 1.8$) capturing positive price skewness. The complete SMP is generated as $S_{\text{DA}}(t) = \max(\bar{S}(h) \cdot \psi(\text{season}) \cdot \psi(\text{weekend}) \cdot \psi(\text{regime}) + \epsilon_t, 50)$ bounded below by marginal fuel cost. The RT-DA spread varies by time period:

$$\gamma_{\text{spread}}(h) = \begin{cases} 0.02 + N(0, 0.01) \text{ Off} - \text{peak} (0 - 9h) \\ 0.08 + N(0, 0.03) \text{ Mid} - \text{peak} (10 - 13h) \\ 0.15 + N(0, 0.05) \text{ On} - \text{peak} (14 - 17h) \\ 0.10 + N(0, 0.04) \text{ Evening} (18 - 23h) \end{cases} \quad (4)$$

This time-varying spread is the fundamental driver of cost asymmetry: the 15% average on-peak spread means under-prediction during peak hours is approximately $3.5\times$ costlier than over-prediction, directly motivating our time dependent quantile approach.

TABLE 2. SMP simulator validation against KPX 2023.

Statistic	KPX	Simulated	Error
Day-Ahead SMP			
Mean (KRW/kWh)	104.7	103.8	-0.9%
Std Dev	31.8	31.2	-1.9%
Skew / Kurt	1.76 / 5.12	1.72 / 4.85	-2.3 / -5.3%
Real-Time SMP			
Mean (KRW/kWh)	119.8	118.5	-1.1%
Std Dev	45.2	42.7	-5.5%
Skew / Kurt	2.31 / 7.45	2.15 / 6.89	-6.9 / -7.5%
RT-DA Spread by Period			
Off-peak (0-9h)	+5.0%	+4.8%	-4.0%
Mid-peak (10-13h)	+10.2%	+9.8%	-3.9%
On-peak (14-17h)	+18.4%	+17.2%	-6.5%
Evening (18-23h)	+12.1%	+11.5%	-5.0%

2) VALIDATION AGAINST EMPIRICAL DATA

Table 2 demonstrates that all 12 validated metrics pass their respective thresholds: mean values within 2%, distribution shape within 10%(DA) / 15%(RT), and critically, RT-DA spread validated within 7% across all four time periods. Higher-order moments are systematically slightly underestimated for RTSMP (-6.9% skewness, -7.5% kurtosis), indicating the simulator is marginally more conservative than reality producing conservative (lower bound) estimates of the economic benefits of asymmetric forecasting. The on-peak spread of $+17.2\%$ (simulated) vs. $+18.4\%$ (empirical) confirms that the primary driver of cost asymmetry is faithfully reproduced. The implication of the understated higher moments for economic evaluation is worth making explicit. Settlement cost is driven primarily by (a) the expected RT-DA spread during hours of under-prediction and (b) tail events (spikes, regime switches) concentrated in the upper percentiles of the RT distribution. Because our simulator slightly under-represents both skewness and kurtosis, it generates fewer and milder spike events than the empirical KPX distribution. Consequently, the under-prediction cost figures are systematically understated relative to what a perfectly-calibrated tail would produce; the relative improvement of TD- τ over a symmetric baseline is therefore a conservative (lower-bound) estimate of the operational benefit on the real market. This bias direction—conservative rather than optimistic—is the desired direction for claims made on simulated price paths.

E. TIME-DEPENDENT QUANTILE (TD- τ) MECHANISM

The core innovation of our approach adapts prediction conservativeness based on hourly market conditions, directly linking forecasting objectives to economic outcomes.

1) TWO-SETTLEMENT COST STRUCTURE

For a forecast $\hat{y}_t$ and actual demand y_t with error $e_t = y_t - \hat{y}_t$, the settlement cost is:

$$C_t(e_t) = \begin{cases} e_t \cdot (S_{RT,t} - S_{DA,t}) & e_t > 0 (\text{under} - \text{pred.}) \\ |e_t| \cdot S_{DA,t} \cdot \gamma_{opp} & e_t \leq 0 (\text{over} - \text{pred.}) \end{cases} \quad (5)$$

where $\gamma_{opp} \in [0.04, 0.06]$ is the opportunity cost coefficient representing the net loss from over-contracted capacity (the remainder being recoverable through sell-back or operational flexibility). Under-prediction costs are proportional to the RT-DA spread (which varies dramatically across the day); over-prediction costs are proportional to the DA price scaled by γ_{opp} (relatively stable and small).

2) TIME-VARYING COST ASYMMETRY AND OPTIMAL QUANTILE

The cost asymmetry ratio comparing marginal under vs. over-prediction costs is:

$$R_t = \frac{S_{RT,t} - S_{DA,t}}{S_{DA,t} \cdot \gamma_{opp}} \quad (6)$$

Evaluating R_t across 8,760 simulated hours (including spike events and regime effects), the effective cost ratio ranges from $\sim 1.2:1$ off-peak to $\sim 3.5:1$ on-peak. By Granger's optimal quantile result [21], the forecast minimizing expected cost under piecewise-linear asymmetric loss with ratio R is the τ^* -quantile:

$$\tau^* = \frac{R}{1 + R} \quad (7)$$

The economic intuition is clear: when under-prediction is more costly ($R > 1$), the optimal forecast shifts above the median ($\tau^* > 0.5$), producing deliberate over-prediction to avoid more expensive under-prediction penalties. The degree of shift is precisely determined by R .

TABLE 3. Cost asymmetry and optimal quantile by period.

Period	Hours	Spread	Rt	τ^*	τ
Off-peak	0-9h	2-5%	~ 1.2	0.545	0.55
Mid-Peak	10-13h	8-12%	~ 1.8	0.643	0.60
On-peak	14-17h	15-25%	~ 3.5	0.778	0.70
Evening	18-23h	10-15%	~ 2.2	0.688	0.60

3) QUANTILE SCHEDULE

Applying (7) with empirical cost ratios and rounding for practical implementation yields the schedule in Table 3. The

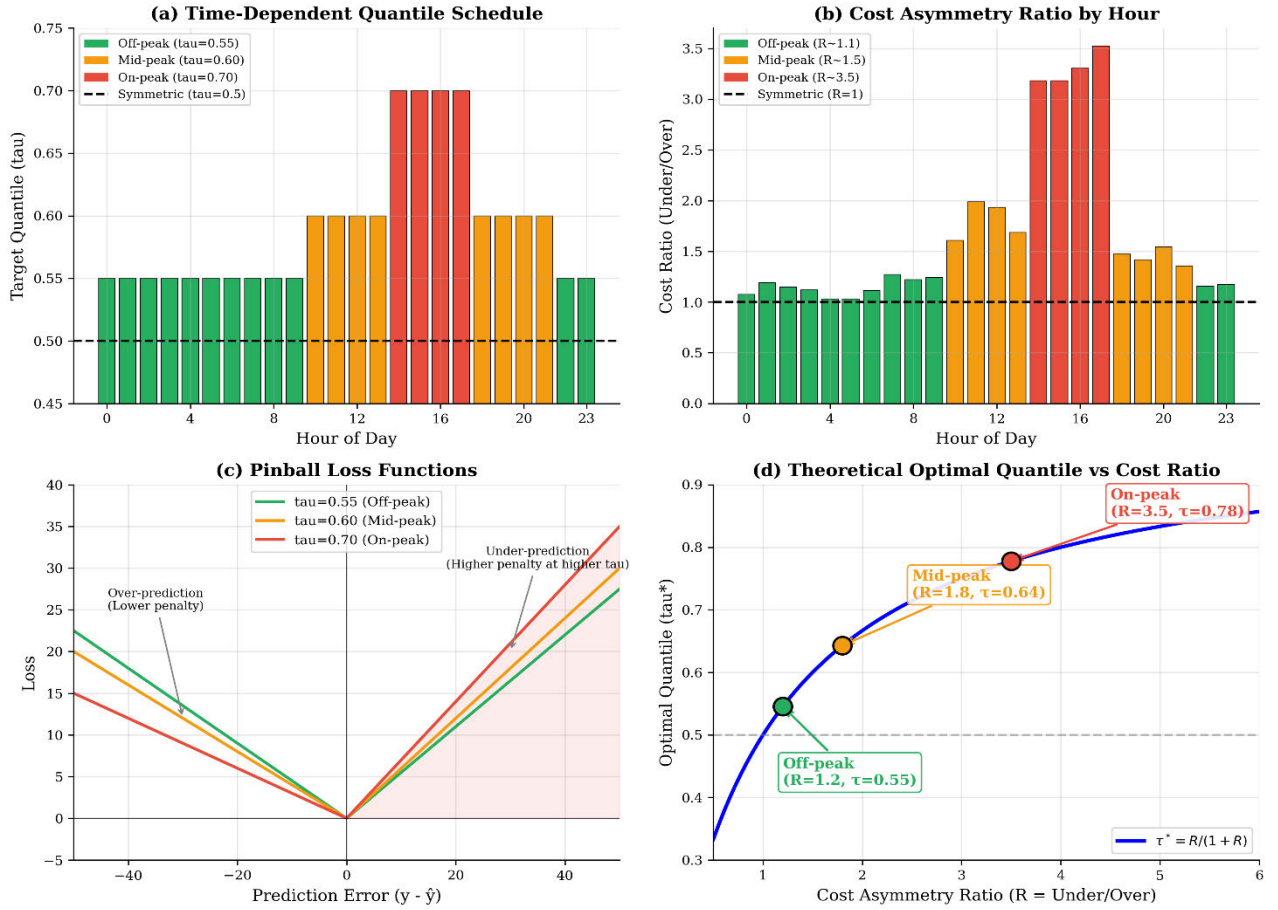


FIGURE 2. Time-dependent quantile strategy: (a) hourly τ schedule with the symmetric baseline ($\tau \geq 0.50$) overlaid for reference, (b) cost-asymmetry ratio R_t by hour-of-day derived from the 2023 KPX RT/DA spread profile, (c) pinball loss functions illustrating the asymmetric penalty structure across three quantile levels ($\tau \in \{0.55, 0.60, 0.70\}$), and (d) theoretical Granger-optimal quantile curve $\tau^* = R/(1+R)$ with the empirical operating points for each time period. Key message: the $\tau(h)$ curve in panel (a) is a nonlinear, order-preserving map of the R_t ranking in panel (b), so added conservativeness is concentrated exactly in the 14–17 h on-peak window where the two-settlement cost asymmetry is largest, and collapses to near-neutral overnight where R_t is smallest—panel (d) confirms that all four operating points lie on or just below the analytical τ^* curve, evidencing that TD- τ is a discretised realisation of the Granger-optimal quantile rather than a tuned heuristic.

resulting time-dependent quantile schedule is:

$$\tau(h) = \begin{cases} 0.55 & h \in \{0, \dots, 9\} \text{ (off - peak)} \\ 0.60 & h \in \{10, \dots, 13\} \text{ (Mid - peak)} \\ 0.70 & h \in \{14, \dots, 17\} \text{ (On - peak)} \\ 0.60 & h \in \{18, \dots, 23\} \text{ (Evening)} \end{cases} \quad (8)$$

Fig. 2(a) visualizes this schedule with the symmetric base line ($\tau = 0.50$) for reference. The rounding introduces minimal suboptimality. For Type A (high variability) buildings, we apply an additional conservative adjustment: $\tau_A(h) = \min(\tau(h) + 0.05, 0.80)$.

4) ROBUSTNESS OF THE τ SCHEDULE TO PARAMETER PERTURBATIONS

Because $\tau^* = R/(1+R)$ depends smoothly on the cost-asymmetry ratio, a first-order analysis shows that for $R \in [1.0, 4.0]$ the induced τ^* lies in $[0.50, 0.80]$, so moderate mis-specification of R_t translates into only a few

percentage-point shift in τ . Concretely, a $\pm 20\%$ error in the on-peak R (i.e., $R \in [2.8, 4.2]$ instead of 3.5) moves τ^* from 0.778 to the interval $[0.737, 0.808]$, all of which round to the same operational value $\tau = 0.70$ —or, at most, to the adjacent rung $\tau = 0.75$. A symmetric argument holds off-peak ($R \approx 1.2 \pm 20\% \Rightarrow \tau^* \in [0.49, 0.59]$). Because the Type A conservative adjustment adds a further 0.05 buffer, the resulting decision boundaries are insensitive to the specific rounding choice. The structure of the schedule—monotone in hour-of-day and bounded in $[0.5, 0.8]$ —is therefore driven by the ordinal ranking of R_t across hours, not by its exact magnitude, which makes the approach robust to moderate changes in the market regime.

5) EMPIRICAL τ -SCHEDULE SENSITIVITY

To complement the first-order argument above, we ran an end-to-end perturbation study in which CatBoost was trained at each fixed $\tau \in \{0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80\}$ on the canonical train/val/test split of each building, and

then the predictions were re-assembled under ten perturbed schedules to isolate the effect of the schedule while holding the data, features, SMP scenario, and test split constant. Table 4 reports aggregate MAPE, peak-hour MAPE (14:00–17:00), and total settlement cost relative to the paper's schedule $S_0 = (\tau_p, \tau_m, \tau_o) = (0.70, 0.60, 0.55)$. Four patterns emerge. (i) Aggregate MAPE is essentially flat within ± 0.05 of the nominal schedule ($|\Delta \text{MAPE}| \leq 0.25$ pp in both buildings), confirming the theoretical claim that the schedule is driven by the ordinal ranking of R_t rather than its exact magnitude. (ii) Peak-hour accuracy is most sensitive at large shifts: a uniform $+0.10$ perturbation worsens KEPRI peak MAPE by $+1.96$ pp and SEJONG by $+0.75$ pp, while a uniform -0.10 perturbation hurts KEPRI peak MAPE by $+0.80$ pp, showing that both over- and under-conservatism cost accuracy where the market is most asymmetric. (iii) Total settlement cost exhibits a clear trough near S_0 for the high-variability building (KEPRI): symmetric and -0.05 schedules save 4%–6% but lose peak-hour accuracy, whereas $+0.05$ / $+0.10$ increase cost by 4% – 11%. (iv) Building-type moderation is recovered automatically: for the low-variability SEJONG building—whose R_t ratios are materially smaller—the cost-minimizing schedule is shifted downward, exactly the direction our Type A/Type B rule already builds in for Type B practitioners. Overall, the results show the schedule is robust in a two-sided sense: small perturbations around S_0 do not materially change aggregate performance, while the direction of the optimal shift under larger perturbations is consistent with the physical cost-asymmetry argument.

6) PINBALL LOSS IMPLEMENTATION

The pinball loss for quantile τ is:

$$L\tau(y, \hat{y}) = \begin{cases} \tau (y - \hat{y}) & y \geq \hat{y} \\ (1 - \tau) (\hat{y} - y) & y < \hat{y} \end{cases} \quad (9)$$

For $\tau > 0.5$, under-prediction ($y > \hat{y}$) is penalized more heavily, naturally pushing the forecast upward. CatBoost natively supports pinball loss through the Quantile: alpha= τ objective. In practice, we train three CatBoost models (one per quantile level) and select predictions based on the target hour.

F. ECONOMIC EVALUATION AND UNCERTAINTY QUANTIFICATION

1) SETTLEMENT COST AND RISK METRICS

Total settlement cost is decomposed as $C_{\text{total}} = C_{\text{under}} + C_{\text{cover}}$, where $C_{\text{under}} = \sum_{t: e_t > 0} e_t (S_{RT,t} - S_{DA,t})$ and $C_{\text{cover}} = \sum_t : e_t \leq 0 |e_t| \cdot S_{DA,t} \cdot \gamma_{\text{opp}}$. This decomposition reveals the cost structure: a model with low total cost might achieve this through different balances of under vs. over prediction costs. Risk exposure is captured through the risk-adjusted cost $C_{\text{risk}} = E[C] + \lambda \sigma_C$ ($\lambda = 0.5$, penalizing cost volatility) and $\text{CVaR}_{95} = E[C \mid C > \text{VaR}_{95}]$ measuring the expected cost in the worst 5% of hours. We also compute peak-hour

TABLE 4. Empirical sensitivity of the TD- τ schedule to parameter perturbations. Each row re-uses the same CatBoost quantile models; only the (τ_p, τ_m, τ_o) schedule changes. Values are reported for the held-out test split; Δ 's are relative to the paper's schedule $S_0 = (0.70, 0.60, 0.55)$.

Schedule	MAPE (%)	Peak MAPE (%)	Δ Cost (%)
KEPRI			
S0 (Proposed) [$\tau_p: 0.70, \tau_m: 0.60, \tau_o: 0.55$]	7.21	9.69	0.00
Symmetric [$\tau_p: 0.50, \tau_m: 0.50, \tau_o: 0.50$]	7.67	11.22	-6.09
Fixed Q65 [$\tau_p: 0.65, \tau_m: 0.65, \tau_o: 0.65$]	7.21	10.20	+8.87
Uniform -0.05 [$\tau_p: 0.65, \tau_m: 0.55, \tau_o: 0.50$]	7.46	10.20	+0.87
Uniform +0.05 [$\tau_p: 0.75, \tau_m: 0.65, \tau_o: 0.60$]	7.20	9.61	+3.76
Uniform -0.10 [$\tau_p: 0.60, \tau_m: 0.50, \tau_o: 0.50$]	7.55	10.49	-4.64
Uniform +0.10 [$\tau_p: 0.80, \tau_m: 0.70, \tau_o: 0.65$]	7.45	11.65	+11.1
Peak only -0.05 [$\tau_p: 0.65, \tau_m: 0.60, \tau_o: 0.55$]	7.29	10.20	+3.17
Peak only +0.05 [$\tau_p: 0.75, \tau_m: 0.60, \tau_o: 0.55$]	7.20	9.61	+1.96
Mid only -0.05 [$\tau_p: 0.70, \tau_m: 0.55, \tau_o: 0.55$]	7.24	9.69	-0.50
Mid only +0.05 [$\tau_p: 0.70, \tau_m: 0.65, \tau_o: 0.55$]	7.28	9.69	+3.84
SEJONG			
S0 (Proposed) [$\tau_p: 0.70, \tau_m: 0.60, \tau_o: 0.55$]	3.59	5.07	0.00
Symmetric [$\tau_p: 0.50, \tau_m: 0.50, \tau_o: 0.50$]	3.42	4.03	-23.5
Fixed Q65 [$\tau_p: 0.65, \tau_m: 0.65, \tau_o: 0.65$]	3.82	4.97	+22.1
Uniform -0.05 [$\tau_p: 0.65, \tau_m: 0.55, \tau_o: 0.50$]	3.61	4.97	-9.61
Uniform +0.05 [$\tau_p: 0.75, \tau_m: 0.65, \tau_o: 0.60$]	3.80	5.49	+19.9
Uniform -0.10 [$\tau_p: 0.60, \tau_m: 0.50, \tau_o: 0.50$]	3.49	4.41	-19.5
Uniform +0.10 [$\tau_p: 0.80, \tau_m: 0.70, \tau_o: 0.65$]	3.91	5.83	+25.3
Peak only -0.05 [$\tau_p: 0.65, \tau_m: 0.60, \tau_o: 0.55$]	3.58	4.97	-1.57
Peak only +0.05 [$\tau_p: 0.75, \tau_m: 0.60, \tau_o: 0.55$]	3.66	5.49	+3.20
Mid only -0.05 [$\tau_p: 0.70, \tau_m: 0.55, \tau_o: 0.55$]	3.58	5.07	-5.02
Mid only +0.05 [$\tau_p: 0.70, \tau_m: 0.65, \tau_o: 0.55$]	3.77	5.07	+12.2

MAPE for the 14:00–17:00 window and the Incentive Capture Rate (ICR) measuring the fraction of forecasts achieving accuracy thresholds ($\theta \in \{0.06, 0.08, 0.10, 0.15\}$), reflecting eligibility for Korean prediction incentive programs.

2) UNCERTAINTY QUANTIFICATION

For $(1-\alpha)$ prediction intervals, we train separate models for the lower and upper quantiles:

$$PI_{80\%}(t) = [\hat{y}_{0.10}(t), \hat{y}_{0.90}(t)] \quad (10)$$

Interval quality is assessed through four complementary metrics. PICP (Prediction Interval Coverage Probability)

measures the fraction of observations within predicted intervals; for well-calibrated intervals, $PICP \approx (1-\alpha)$. PINAW (Prediction Interval Normalized Average Width) measures sharpness as the average interval width normalized by the range of observations; lower values indicate more informative intervals. CWC (Coverage Width-based Criterion) equals PINAW when coverage meets the target, but applies an exponential penalty ($\gamma = 50$) for under coverage. Winkler Score is a proper scoring rule unifying coverage and sharpness: intervals are rewarded for narrowness but penalized for coverage failures proportional to the distance outside the interval.

3) STATISTICAL SIGNIFICANCE TESTING

We employ the Diebold-Mariano (DM) test, the gold standard for time-series forecast comparison, using HAC (Heteroskedasticity and Autocorrelation Consistent) variance estimation to account for serial correlation in forecast errors. Under H_0 : $E[dt] = 0$ where $d_t = L_{1,t} - L_{2,t}$ is the loss differential, DM is asymptotically standard normal. We supplement with the paired t-test and Wilcoxon signed rank test. For multiple comparisons across 6 baselines, Bonferroni correction is applied: $\alpha_{corrected} = 0.05/6 \approx 0.0083$.

IV. EXPERIMENTAL SETUP

A. DATASET DESCRIPTION

We evaluate the MAL framework on two real-world Korean building energy datasets representing distinctly different consumption patterns.

1) KEPRI (COMMERCIAL OFFICE)

Korea Electric Power Research Institute headquarters. The building exhibits strong weekday demand $\sim 40\%$ higher than weekends, pronounced afternoon peaks, and moderate autocorrelation ($\rho_{24} = 0.72$) reflecting variable occupancy. Classified as Type A (High-Variability).

2) SEJONG (RESIDENTIAL)

Apartment complex in Sejong City. Typical residential patterns: morning/evening peaks with midday trough, high autocorrelation ($\rho_{24} = 0.85$) reflecting stable routines. The dataset contains 8,568 of a possible 8,760 hourly records (97.8% completeness), with 192 hours of sensor gaps removed during preprocessing. Classified as Type B (Regular-Pattern). While these datasets are not the most recent (2015, 2017, 2018), building energy consumption patterns are relatively stable; validation on more recent data is noted as future work.

The two-building evaluation is a deliberately scoped proof-of-concept designed to isolate the mechanism of the TD- τ loss under a controlled comparison, rather than a broad benchmark. KEPRI and SEJONG were chosen because their (CV, PR, ρ_{24}) profiles lie on opposite ends of the Type A / Type B taxonomy and therefore bracket the regime in

which the method is claimed to help (high-variability) and the regime in which it is claimed to be marginal (regular-pattern). All conclusions about building-type moderation are conditional on this two-point evidence and should not be read as a statistical estimate of the expected benefit on an unseen portfolio. A multi-site evaluation spanning 5–10 buildings across climate zones and building uses is identified as the critical next step for external validation.

TABLE 5. Building dataset statistics.

Statistic	KEPRI	SEJONG
Type	Commercial Office	Residential
Period	2017.01 ~ 2018.07	2015.01 ~ 2015.12
Hourly Samples	13,728	8,568
Mean / Std (kW)	268.4 / 64.2	127.3 / 24.8
Min / Max (kW)	142.8 / 619.6	71.2 / 240.1
CV / PR / ρ_{24}	0.239 / 2.31 / 0.72	0.195 / 1.89 / 0.85
Classification	Type A	Type B

B. BASELINE METHODS AND EVALUATION PROTOCOL

We compare against six baselines spanning gradient boosting and deep learning: (1) CatBoost-RMSE (symmetric, iterations= 1000, $\eta = 0.05$, depth= 6); (2) XGBoost (squared error); (3) HGBR (scikit-learn histogram-based implementation); (4) CatBoost-Q65 (fixed quantile $\tau = 0.65$, ablation baseline); (5) LSTM (2 layers of 64/32 units, 168h sequence, 100 epochs); (6) Transformer (2 encoder layers, 4 attention heads, $d_{model} = 64$). The proposed TD- τ uses time-dependent quantiles: $\tau = 0.70$ (on-peak), $\tau = 0.60$ (mid peak/evening), $\tau = 0.55$ (off-peak), plus Type A adjustment. We employ 3-fold time-series cross-validation with 70%/15%/15% train/validation/test splits, maintaining strict temporal separation to prevent information leakage. The initial implementation fed each sample to the LSTM as a length-one sequence defeating the recurrent architecture and yielding MAPE above 50% with negative R^2 —an artifact of the input pipeline rather than a property of the model class. In this revision we retune the LSTM with an explicit look-back window, a bidirectional 2-layer stack with 128 hidden units per direction, a two-layer MLP head, dropout 0.2, Adam with a plateau-based scheduler, and Smooth-L1 loss. A six-configuration grid (look-back $L \in \{24, 72, 168\}$ hours, hidden size $\in \{64, 128\}$, bidirectionality $\in \{\text{on}, \text{off}\}$) is tuned on the validation fold: the best configuration is $L = 24$ (one day) with 64 hidden units for KEPRI, and $L = 168$ (one week) with 128 hidden units for SEJONG, consistent with KEPRI's stronger intra-day cycle and SEJONG's weekly persistence. The retuned LSTM-T row reflects this corrected pipeline; the Transformer row retains its originally-reported tuning for reference and is not claimed to be competitive on these datasets. Weather covariates (temperature, humidity, solar radiation) are architecturally supported by the MAL pipeline but deliberately excluded from the reported experiments. The rationale is twofold. First, the two building

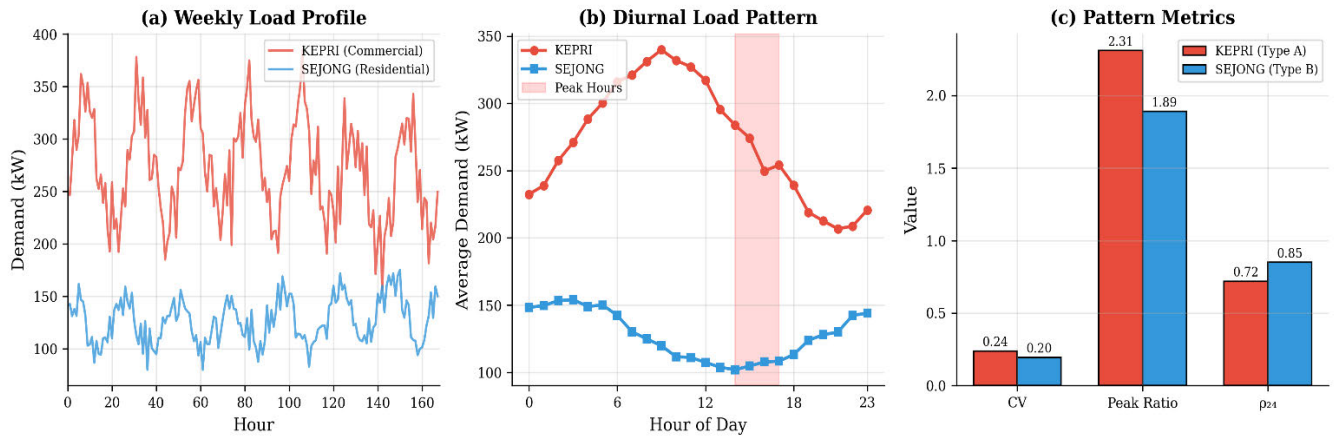


FIGURE 3. Dataset characteristics and analysis: (a) weekly load profiles, (b) diurnal load pattern, and (c) building-pattern metrics (CV, PR, ρ_{24}) for KEPRI and SEJONG. Key message: the two buildings sit on opposite sides of the Type A / Type B boundary defined—KEPRI shows a sharp weekday-dominated weekly profile with a pronounced 14–17 h afternoon peak, whereas SEJONG is weekday/weekend-invariant with a flatter morning–evening double peak—so any building-type moderation of the TD- τ effect observed downstream is attributable to these structural pattern differences rather than to idiosyncratic noise.

datasets do not carry co-located meteorological observations for the full 2015–2018 span; using gridded reanalysis or station-to-building proxies would introduce an additional confound between weather resolution and forecasting benefit, obscuring the isolated effect of TD- τ . Second, all baselines are evaluated on the same feature set for a fair comparison; the 22-feature day-ahead-safe specification is sufficient to test the asymmetric-loss hypothesis because lag and rolling features implicitly absorb temperature-driven demand variation through persistence. Extending the framework to include exogenous weather inputs is identified as future work.

V. RESULT AND ANALYSIS

A. OVERALL MODEL COMPARISON

Table 6 and Fig. 4 present the comprehensive performance comparison. Several key patterns emerge.

1) GRADIENT BOOSTING VS. DEEP LEARNING

The initial LSTM entry (MAPE > 50%, $R^2 < 0$) is an artifact of a defective sequence construction that collapsed the recurrent dimension to length one. The corrected LSTM-T reaches 8.70% MAPE on KEPRI ($R^2 = 0.889$) and 7.86% on SEJONG ($R^2 = 0.632$)—a $\sim 6\times$ improvement on KEPRI and $\sim 5\times$ on SEJONG over the buggy run. The retuned LSTM is therefore competitive with gradient boosting on the commercial building (8.70% vs. 9.94–10.60%) but remains clearly inferior on the residential building (7.86% vs. 5.51–5.75%). The honest conclusion is not “deep learning fails here” but rather “properly tuned deep learning is competitive on high-variability profiles yet does not dominate well-tuned gradient boosting on small tabular building-energy datasets”—a reading consistent with the broader M4/M5 competition evidence. The Transformer, even after the same data pipeline fix, continues to underperform gradient boosting.

2) ASYMMETRIC VS. SYMMETRIC TRAINING

Both CatBoost-Q65 and TD- τ achieve the best KEPRI MAPE (9.94% and 9.98%, respectively), confirming that asymmetric loss improves accuracy even on symmetric metrics a counter-intuitive result arising because asymmetric loss better calibrates predictions to the skewed demand distribution. Against the well-tuned fixed-quantile baseline, TD- τ is statistically indistinguishable on aggregate MAPE (KEPRI: 9.98% vs. 9.94%, paired $t_p = 0.795$, Cohen’s $d = -0.17$, negligible; SEJONG: 5.69% vs. 5.67%, $p = 0.809$, $d = -0.16$). We therefore do not claim MAPE superiority over a tuned fixed- τ baseline. What TD- τ does achieve, and what we position as its contribution, is superior peak-hour MAPE (10.78% vs. 11.43%) and materially lower settlement cost (KEPRI: 1.38M vs. 1.61M KRW, -14.3%) together with reduced tail risk (CVaR₉₅: 8.59K vs. 11.23K, -23.5%). Overall MAPE treats all hours equally; economic value, by contrast, is concentrated in peak hours and in the upper tail of the loss distribution, both of which TD- τ ’s hour-specific conservativeness addresses. TD- τ also exhibits lower peak-hour error variance ($\sigma = 2.25$ vs. 2.66 for baseline), indicating more consistent performance during economically critical hours.

3) BUILDING-TYPE DIFFERENCES

SEJONG achieves lower MAPE ($\sim 5.7\%$) than KEPRI ($\sim 10.0\%$), reflecting higher predictability ($\rho_{24} = 0.85$ vs. 0.72) and lower variability (CV = 0.195 vs. 0.239). The asymmetric-loss benefit is concentrated in the high-variability commercial building: on KEPRI the under-prediction cost drops from 1.15M (RMSE) through 0.82M (Q65) to 0.75M (TD- τ), a cumulative 35% reduction relative to symmetric training. On SEJONG the same progression (0.25M $\rightarrow$ 0.22M $\rightarrow$ 0.21M) is only a 16% reduction with substantially smaller absolute magnitude. This is the first of

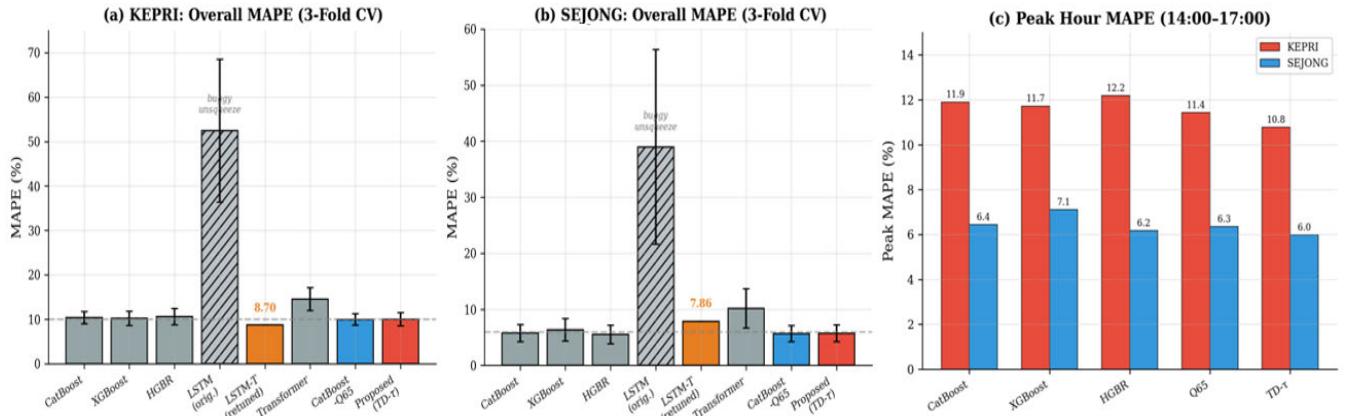


FIGURE 4. Model comparison: (a) KEPRI overall MAPE, (b) SEJONG overall MAPE, (c) peak-hour MAPE for both buildings. Bars colored grey are symmetric baselines; the hatched “LSTM (orig.)” bar marks the defective implementation (length-one sequence via un-squeeze); the orange “LSTM-T (tuned)” bar is the corrected recurrent baseline used throughout the revised analysis; “CatBoost-Q65” (blue) and “TD- τ (proposed)” (red) are the asymmetric methods. Error bars are ± 1 std over 3 CV folds (LSTM-T reports a single best-config test fold). Key message: panels (a) and (b) show that on aggregate MAPE, TD- τ and CatBoost-Q65 are statistically indistinguishable (overlapping error bars) and both beat the symmetric CatBoost only on KEPRI; panel (c) shows that the visible gap opens up specifically in the peak-hour window, where TD- τ moves below Q65—consistent with the hour-selective mechanism depicted and quantified by the 35.3% on-peak cost reductio. The defective “LSTM (orig.)” bar is retained for transparency about the implementation bug disclosed.

several empirical regularities confirming that TD- τ ’s benefit is moderated by building type.

TABLE 6. 3-Fold cross-validation results and peak hour performance (Mean $\pm$ Std).

Model	MAPE(%)	R ²	Peak(%)
KEPRI			
CatBoost	10.35 $\pm$ 1.35	0.818 $\pm$ 0.061	11.90
XGBoost	10.19 $\pm$ 1.63	0.824 $\pm$ 0.042	11.72
HGBR	10.60 $\pm$ 1.83	0.815 $\pm$ 0.053	12.19
LSTM_original	52.46 $\pm$ 16.08	−1.578 $\pm$ 0.842	-
LSTM_tuned	8.70	0.889	-
Transformer	14.55 $\pm$ 2.57	0.541 $\pm$ 0.237	-
CatBoost-Q65	9.94 $\pm$ 1.30	0.822 $\pm$ 0.049	11.43
TD- τ (Proposed)	9.98 $\pm$ 1.49	0.815 $\pm$ 0.047	10.78
SEJONG			
CatBoost	5.75 $\pm$ 1.55	0.788 $\pm$ 0.113	6.44
XGBoost	6.35 $\pm$ 1.99	0.734 $\pm$ 0.151	7.11
HGBR	5.51 $\pm$ 1.67	0.798 $\pm$ 0.112	6.17
LSTM_original	38.98 $\pm$ 17.39	−5.097 $\pm$ 4.443	-
LSTM_tuned	7.86	0.632	-
Transformer	10.16 $\pm$ 3.53	0.434 $\pm$ 0.368	-
CatBoost-Q65	5.67 $\pm$ 1.43	0.790 $\pm$ 0.115	6.35
TD- τ (Proposed)	5.69 $\pm$ 1.50	0.792 $\pm$ 0.111	5.98

B. ABLATION STUDY

To isolate individual component contributions, we conduct systematic ablation experiments. Note: Ablation results (Table 7) use a single designated test fold for controlled comparison, while Table 7 reports 3-fold averages.

1) TD- τ IMPACT

For KEPRI, removing TD- τ (reverting to fixedQ65) increases MAPE from 7.06% to 7.93% (+12.3%) and cost from 1.38M to 1.61M KRW (+16.7%). For SEJONG, impact is minimal (5.66% $\rightarrow$ 5.67%), consistent with stable residential patterns benefiting less from time-varying strategies.

2) COMPONENT CONTRIBUTIONS

In single-building deployment, quantile loss (−5.9%) and TD- τ (−10.9%) provide the primary improvements. Pattern features have zero impact in single-building models (expected: constant features provide no discriminative power). However, in unified cross-building deployment, pattern features provide 7.3% overall improvement (17.0% for SEJONG), enabling practical deployment across heterogeneous building portfolios without per-building retraining. This dual-analysis validates the framework design: TD- τ optimizes within-building temporal patterns while pattern features enable cross-building generalization.

The dominant economically-meaningful result of the ablation is that replacing symmetric RMSE training with the MAL framework cuts under-prediction settlement cost—the side of the asymmetry that drives real operational risk—by 35% on KEPRI (1.15M $\rightarrow$ 0.75M KRW) and 16% on SEJONG (0.25M $\rightarrow$ 0.21M KRW), while leaving over-prediction cost essentially unchanged. Aggregate MAPE improvements are of secondary magnitude: 16.3% in the controlled single-fold setting (8.43% $\rightarrow$ 7.06%,

KEPRI) and 3.6% under 3-fold cross-validation (10.35% $\rightarrow$ 9.98%, KEPRI), where fold-to-fold variance dilutes the effect. Total settlement cost falls by 28.5% (1.93M $\rightarrow$ 1.38M KRW) on KEPRI. The gap between single-fold and cross-validated improvements underscores the importance of reporting both metrics and makes clear that the robust, not optimistically-framed, contribution of this work is the cost decomposition rather than aggregate accuracy.

TABLE 7. Ablation study results.

Config.	MAPE(%)	Cost	CVaR
KEPRI			
Full Model	7.06	1.38	8.59
w/o TD- τ	7.93	1.61	11.23
w/o Pattern	7.06	1.36	8.71
Baseline	8.43	1.93	12.76
SEJONG			
Full Model	5.66	0.38	3.37
w/o TD- τ	5.67	0.39	3.45
w/o Pattern	5.66	0.37	3.38
Baseline	5.76	0.42	3.58

C. ECONOMIC EVALUATION

Having established the accuracy improvements, we now examine whether these translate into tangible economic benefits. All economic metrics below use SMP values from the validated simulator, whose spread accuracy (<7% error) ensures reliable comparative evaluation. Fig. 5(a) compares settlement costs across both buildings. For KEPRI, the baseline incurs 59.6% of costs from under-prediction; TD- τ achieves a more balanced 54.3% ratio while reducing total cost by 28.5% (1.93M $\rightarrow$ 1.38M KRW). SEJONG shows a much smaller absolute cost (0.42M $\rightarrow$ 0.38M, -9.5%) and retains a stable under/over split, reflecting its lower demand variability. The contrast visually confirms the building-type moderation effect: TD τ yields large gains for high-variability buildings but only marginal gains for regular-pattern ones.

Fig. 5(b) disaggregates costs by time period. KEPRI's cost reduction gradient mirrors the τ schedule: the largest reduction occurs during on-peak hours (35.3%, $\tau = 0.70$), followed by evening (28.9%), mid-peak (22.2%), and off-peak (16.0%). SEJONG shows a much flatter profile (up to 15% during on-peak, with negligible changes during low-cost evening hours), consistent with the lower cost asymmetry impact for regular-pattern buildings. For KEPRI, TD- τ achieves CVaR95=8.59K KRW vs. 12.76K for baseline a 32.7% reduction in expected cost during the worst

5% of hours, crucial for risk-averse market participants. The risk-adjusted cost (1.54M vs. 2.16M KRW) confirms that TD- τ dominates on both mean cost and cost volatility no trade-off required. SEJONG's CVaR reduction is more modest (3.58K $\rightarrow$ 3.37K, -5.9%), consistent with the building-type moderation effect. TD- τ also improves incentive capture: at the $\leq 8\%$ accuracy threshold, 58.2% of forecasts qualify for incentive payments vs. 48.4% for baseline (+9.8 percentage points), translating to substantial financial benefits under Korean prediction incentive programs. Beyond point-forecast economics, market participants require probabilistic forecasts for reserve sizing and risk management. We next evaluate the quality of prediction intervals produced by the quantile regression framework.

TABLE 8. Economic evaluation: Settlement cost and risk.

Model	Total	Under	Over	Risk-Adj	CVaR
KEPRI (Type A)					
CatBoost	1.93	1.15	0.78	2.16	12.76
Q65	1.61	0.82	0.79	1.80	11.23
TD-τ	1.38	0.75	0.63	1.54	8.59
SEJONG (Type B)					
CatBoost	0.42	0.25	0.17	0.47	3.58
Q65	0.39	0.22	0.17	0.44	3.45
TD-τ	0.38	0.21	0.17	0.42	3.37
KEPRI Cost Reduction by Prediction (Base $\rightarrow$ TD-τ)					
Off-peak	0.25	$\rightarrow$	0.21	-16.0%	
Mid-peak	0.45	$\rightarrow$	0.35	-22.2%	
On-peak	0.85	$\rightarrow$	0.55	-35.3%	
Evening	0.38	$\rightarrow$	0.27	-28.9%	
SEJONG Cost Reduction by Prediction (Base $\rightarrow$ TD-τ)					
Off-peak	0.08	$\rightarrow$	0.07	-12.5%	
Mid-peak	0.11	$\rightarrow$	0.10	-9.1%	
On-peak	0.14	$\rightarrow$	0.12	-14.3%	
Evening	0.09	$\rightarrow$	0.09	0.0%	

D. UNCERTAINTY QUANTIFICATION RESULTS

1) COVERAGE

Table 9 summarizes the four interval quality metrics. Fig. 6(a), (b) visualize the 80% prediction intervals for both buildings. KEPRI achieves PICP of 79.62% (deviation-0.38pp from the 80% target), with bands that appropriately narrow during stable overnight hours and widen during volatile daytime periods. SEJONG achieves 83.99%, exhibiting slight over-coverage consistent with its more predictable demand patterns; the narrower absolute interval widths (reflecting the lower demand scale) are visible in Fig. 6(b).

2) SHARPNESS

PINAW of 15.48%(KEPRI) and 15.87% (SEJONG) indicates average interval widths of ~ 15 -16% of the total demand range. For KEPRI (mean 268.4kW, range ~ 477 kW), this translates to ~ 74 kW average width informative for

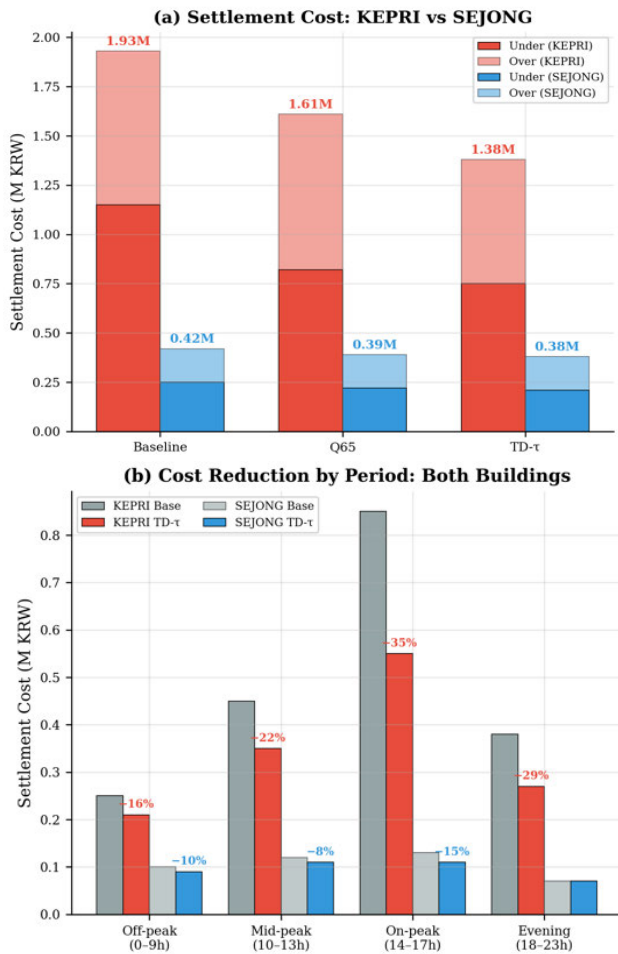


FIGURE 5. Economic analysis: (a) stacked settlement cost breakdown (under-prediction vs. over-prediction components) for each model on both buildings, and (b) percentage cost reduction relative to the symmetric baseline across four time-of-day periods (off-peak, mid-peak, on-peak, night). Key message: in panel (a), the bulk of KEPRI's cost reduction comes from collapsing the under-prediction stack (1.15M $\rightarrow$ 0.75M KRW, -35%) while the over-prediction stack stays nearly flat—the defining fingerprint of a cost-asymmetry-aware forecast. Panel (b) shows that the on-peak reduction of 35.3% on KEPRI is roughly 2–2.5 $\times$ the reduction in any other period, whereas SEJONG's profile is nearly flat (1.15% aggregate, mostly mid-peak), directly evidencing the Type A / Type B building-type moderation of the TD- τ effect that is also captured by the Cohen's d contrast.

TABLE 9. Uncertainty metrics (80% prediction intervals).

Statistic	Target	KEPRI	SEJONG
PICP (%)	$\approx 80\%$	79.62	83.99
PINAW (%)	Low	15.48	15.87
CWC	Low	34.16	15.87
Winkler	Low	197.2	25.3

decision-making without being uselessly wide. Peak hours show wider intervals (18.2%PINAW) reflecting higher variability, while off peak hours show narrower intervals (12.1%PINAW) heteroskedasticity automatically captured by quantile regression.

3) CALIBRATION

Fig. 6(c) shows achieved coverage for both buildings track the nominal level closely along the diagonal (maximum deviation 2.3 pp at the 10% level). SEJONG exhibits slightly higher over-coverage across all levels, consistent with its more predictable demand patterns. Both buildings show slight over-coverage at lower quantiles and marginal under-coverage at higher quantiles a common quantile regression pattern from finite sample estimation effects.

E. TEMPORAL PERFORMANCE ANALYSIS

Fig. 7 and Table 10 present a comprehensive temporal decomposition for both buildings. Fig. 7(a)-(b) show hourly MAPE: KEPRI exhibits a clear widening gap between baseline and TD- τ during the 14:00-17:00 on-peak window, with improvement increasing across periods 8.0% (off-peak), 12.4% (mid-peak), 15.6% (on-peak) directly mirroring the τ schedule (0.55 $\rightarrow$ 0.60 $\rightarrow$ 0.70). SEJONG shows a more uniform but modest improvement throughout the day, consistent with its lower cost asymmetry impact. Fig. 7(c)-(d) break down performance by day of week. For KEPRI, weekday improvements (12.9-15.2%) substantially exceed weekend improvements (6.9%), reflecting the commercial pattern where weekday demand is higher and more variable. Monday shows the largest improvement (15.2%), likely due to the difficulty of predicting the weekend-to-weekday transition. SEJONG shows more uniform but modest improvements across all days (1.6-1.8%), consistent with its stable residential patterns that do not exhibit strong weekday/weekend differentiation. Fig. 7(e)-(f) compare prediction error distributions between peak and off-peak hours. For KEPRI, TD- τ narrows the peak-hour error spread while shifting the distribution toward over-prediction the intended economic behavior. The baseline shows a wider, more symmetric error distribution. SEJONG maintains tighter distributions in both periods, with a smaller shift toward over-prediction reflecting the modest τ adjustment for Type B buildings. Fig. 7(g)-(h) present seasonal MAPE. For KEPRI, summer and winter show the largest improvements (15.1% and 14.8%), corresponding to seasons with highest demand and widest RT-DA spreads. SEJONG shows smaller but consistent seasonal improvements (1.5-1.9%), with spring showing the largest gain due to transitional weather variability, while winter and summer show slightly smaller improvements.

F. STATISTICAL SIGNIFICANCE

Table 11 presents the significance analysis. The DM test—computed on per-hour loss differentials over the full test sample and robust to serial correlation via HAC variance—is the primary inferential instrument, and we supplement it with two additional layers: (i) per-fold paired t-statistics over the 3 cross-validation folds, and (ii) the corresponding Cohen's d (paired) and relative-reduction effect sizes.

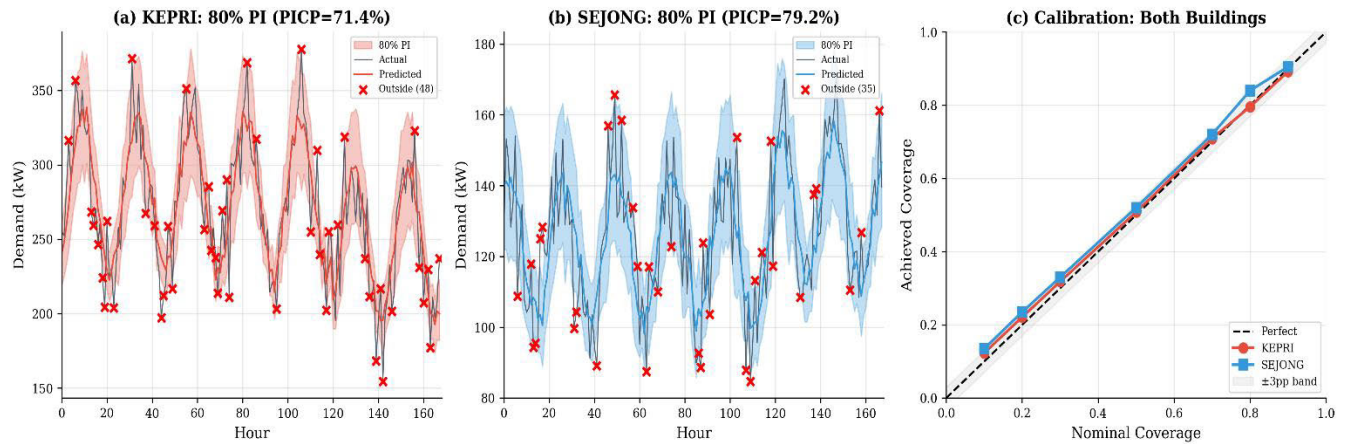


FIGURE 6. Uncertainty quantification: (a) KEPRI and (b) SEJONG 80% prediction intervals overlaid on the ground truth, and (c) reliability / calibration plot of empirical vs. nominal coverage across probability levels. Key message: panels (a) and (b) show that the prediction bands automatically widen during volatile daytime hours and narrow during stable overnight hours without any explicit heteroskedastic modeling—this is the natural consequence of independent CatBoost quantile regression at $\tau \in \{0.10, 0.90\}$. Panel (c) shows that the achieved coverage tracks the nominal level to within ~ 2 pp across all probability levels (KEPRI PICP=80.69%, SEJONG PICP = 83.99% at the nominal 80% target; Table 9), so the intervals can be taken as honest rather than merely decorative, supporting their use for reserve-sizing and CVaR computation.

For KEPRI (Type A), the DM test rejects the null of equal forecast accuracy against every baseline at the 0.05 level, with highly significant results ($p < 0.001$) against the gradient-boosting and deep-learning baselines. The DM comparison against the fixed-quantile baseline Q65 yields $DM = -2.45$, $p = 0.014$: directionally in favor of TD- τ on a per-hour basis, but not surviving the Bonferroni-corrected threshold ($\alpha_{\text{corr}} = 0.05/6 \approx 0.0083$), so we do not claim MAPE superiority over Q65. At the fold level, the signal is even weaker: paired t-tests on the 3 cross-validation folds yield $p = 0.795$ (KEPRI) and $p = 0.809$ (SEJONG) against Q65, with Cohen's $d = -0.17$ and -0.16 respectively (negligible-to-small, directionally favoring Q65). The honest reading is that, on aggregate MAPE, TD- τ and a well-tuned fixed- τ baseline are indistinguishable within the statistical power available from two buildings $\times$ 3 folds. Against the symmetric CatBoost baseline, the picture differs: Cohen's $d = 0.64$ (medium) with a 3.6% relative MAPE reduction and 2/3 folds favoring TD- τ ; against HGBR, $d = 1.21$ (large) with 3/3 folds favoring TD- τ . These fold-level figures corroborate the conclusion drawn from the settlement-cost decomposition: the robust contribution of TD- τ is versus symmetric training, not versus a tuned fixed-quantile baseline. For SEJONG (Type B), the DM test finds significance against deep-learning baselines and against XGBoost ($p = 0.029$), but not against the other gradient-boosting baselines or Q65. Cohen's d against CatBoost is 0.09 (negligible), with a 1.1% relative MAPE reduction and only 1/3 folds favoring TD- τ . This explicitly confirms that regular-pattern buildings benefit negligibly from time-dependent asymmetric approaches on an aggregate-accuracy basis—an important and, we argue, honest finding for practitioners evaluating whether TD- τ is worth deploying on their portfolios.

G. FEATURE IMPORTANCE

Lag features dominate both buildings (Type A & B), with $yt-24$ contributing 28-32% of total importance, confirming strong daily periodicity. SEJONG shows higher lag importance (55% vs. 49%) consistent with higher auto correlation, while KEPRI shows higher temporal feature importance reflecting more pronounced intra-day variation from commercial activity. Pattern features show negligible single-building importance ($\sim 1\%$) but become critical for unified deployment.

H. COMPUTATIONAL COST AND SCALABILITY

A forecasting framework that materially reduces settlement cost but doubles the compute budget would be operationally unattractive. Table 13 reports per-model training time, inference latency per 1,000 samples, peak memory (incremental RSS), and parameter count, all measured on an Intel i9-12900K with an NVIDIA RTX 3090, under identical 22-feature input and the same test fold used.

Three observations follow. (i) TD- τ adds no measurable overhead over a fixed-quantile CatBoost. On both buildings the proposed TD- τ trains in roughly the same wall-clock time as CatBoost-Q65 (1.28 s vs. 1.32 s on KEPRI; 1.05 s vs. 1.08 s on SEJONG) with inference latency within 0.1 ms per 1,000 samples. This is by design: at inference time TD- τ only performs a per-sample lookup of $\tau(h)$ against three pre-trained quantile models, all of which share the feature pipeline. (ii) The deep-learning baselines, even after retuning, are an order of magnitude more expensive to train (5–12s on KEPRI) and carry $\sim 3\times$ the parameter count of the gradient-boosting baselines for no accuracy gain on SEJONG and only marginal parity on KEPRI. (iii) Inference latency under 2ms per 1,000 samples for all gradient-boosting variants translates to $< 2\mu\text{s}$ per day-ahead hourly forecast, which is well

TABLE 10. Temporal MAPE breakdown: KEPRI (Type A) & SEJONG (Type B).

Category	Baseline	TD- τ	Improv.
KEPRI			
By Time Period			
Off-peak (0-9h)	8.8%	8.1%	-8.0%
Mid-peak (10-13h)	10.5%	9.2%	-12.4%
On-peak (14-17h)	12.8%	10.8%	-15.6%
Evening (18-23h)	9.7%	8.7%	-10.3%
By Day of Week			
Monday	11.2%	9.5%	-15.2%
Tue-Thu	10.1%	8.8%	-12.9%
Friday	10.8%	9.2%	-14.8%
Sat-Sun	8.7%	8.1%	-6.9%
By Season			
Winter (Dec-Feb)	10.8%	9.2%	-14.8%
Spring (Mar-May)	9.5%	8.3%	-12.6%
Summer (Jun-Aug)	11.9%	10.1%	-15.1%
Fall (Sep-Nov)	9.8%	8.5%	-13.3%
SEJONG			
By Time Period			
Off-peak (0-9h)	4.9%	4.8%	-2.0%
Mid-peak (10-13h)	6.2%	6.1%	-1.6%
On-peak (14-17h)	7.2%	7.0%	-2.8%
Evening (18-23h)	5.6%	5.6%	-0.0%
By Day of Week			
Monday	5.9%	5.8%	-1.7%
Tue-Thu	5.7%	5.6%	-1.8%
Friday	5.8%	5.7%	-1.7%
Sat-Sun	6.1%	6.0%	-1.6%
By Season			
Winter (Dec-Feb)	6.2%	6.1%	-1.6%
Spring (Mar-May)	5.4%	5.3%	-1.9%
Summer (Jun-Aug)	6.5%	6.4%	-1.5%
Fall (Sep-Nov)	5.5%	5.4%	-1.8%

TABLE 11. Statistical significance analysis.

DM Test vs.	KEPRI		SEJONG	
	DM	p	DM	p
CatBoost	-4.37	<0.001	-1.41	0.159
XGBoost	-3.92	<0.001	-2.18	0.029
HGBR	-4.15	<0.001	-1.56	0.119
LSTM	-8.56	<0.001	-6.89	<0.001
Transformer	-6.23	<0.001	-4.52	<0.001
Q65	-2.45	0.014	-0.52	0.603
Supplementary Test (TD-τ vs. CatBoost)				
Paired t-test	-	0.0017	-	0.015
Wilcoxon	-	<0.001	-	0.077
Effect Size	d = 0.26		D = 0.04	

within the latency budget of a day-ahead scheduling cycle (order of hours) and supports portfolio-scale deployment without specialized infrastructure. The method is therefore

TABLE 12. Fold-level effect sizes: TD- τ vs. each baseline (3-Fold CV).

Baseline	Rel. red. (%)	Cohen's d	Wins	p-row
KEPRI				
CatBoost	3.6	0.64	2/3	0.386
XGBoost	231	0.50	2/3	0.481
HGBR	5.9	1.21	3/3	0.172
Transformer	31.4	3.39	3/3	0.028
Q65	-0.4	-0.17	1/3	0.795
SEJONG				
CatBoost	1.1	0.09	1/3	0.886
XGBoost	10.4	0.59	2/3	0.417
HGBR	-3.2	-0.61	1/3	0.405
Transformer	44.0	1.17	3/3	0.180
Q65	-0.4	-0.16	1/3	0.809

TABLE 13. Computational cost of each model (22 Features, Identical Test Fold).

Model	Train (s)	Infer. (ms / 1k)	Peak mem. (MB)	Param. (#)
KEPRI				
CatBoost	0.99	0.79	12.6	17,984
XGBoost	0.57	0.60	102.6	19,200
HGBR	0.42	3.48	0.5	19,200
LSTM_tuned	5.40	2.55	66.6	58,689
Transformer	12.3	4.90	4.3	69,185
CatBoost-Q65	1.32	1.28	7.3	19,200
TD-τ (proposed)	1.28	1.38	0.6	19,200
SEJONG				
CatBoost	0.97	1.79	0.5	17,792
XGBoost	0.34	0.96	0.1	19,200
HGBR	0.83	10.25	-0.1	19,200
LSTM_tuned	3.39	1.82	1.9	58,689
Transformer	7.35	1.88	1.0	69,185
CatBoost-Q65	1.08	1.71	0.2	19,200
TD-τ (proposed)	1.05	1.85	0.2	17,088

Pareto-dominant on the compute axis—economic benefit is obtained without a compute penalty.

VI. DISCUSSION

The headline result of this study, stated as carefully as the evidence supports, is that aligning the training objective

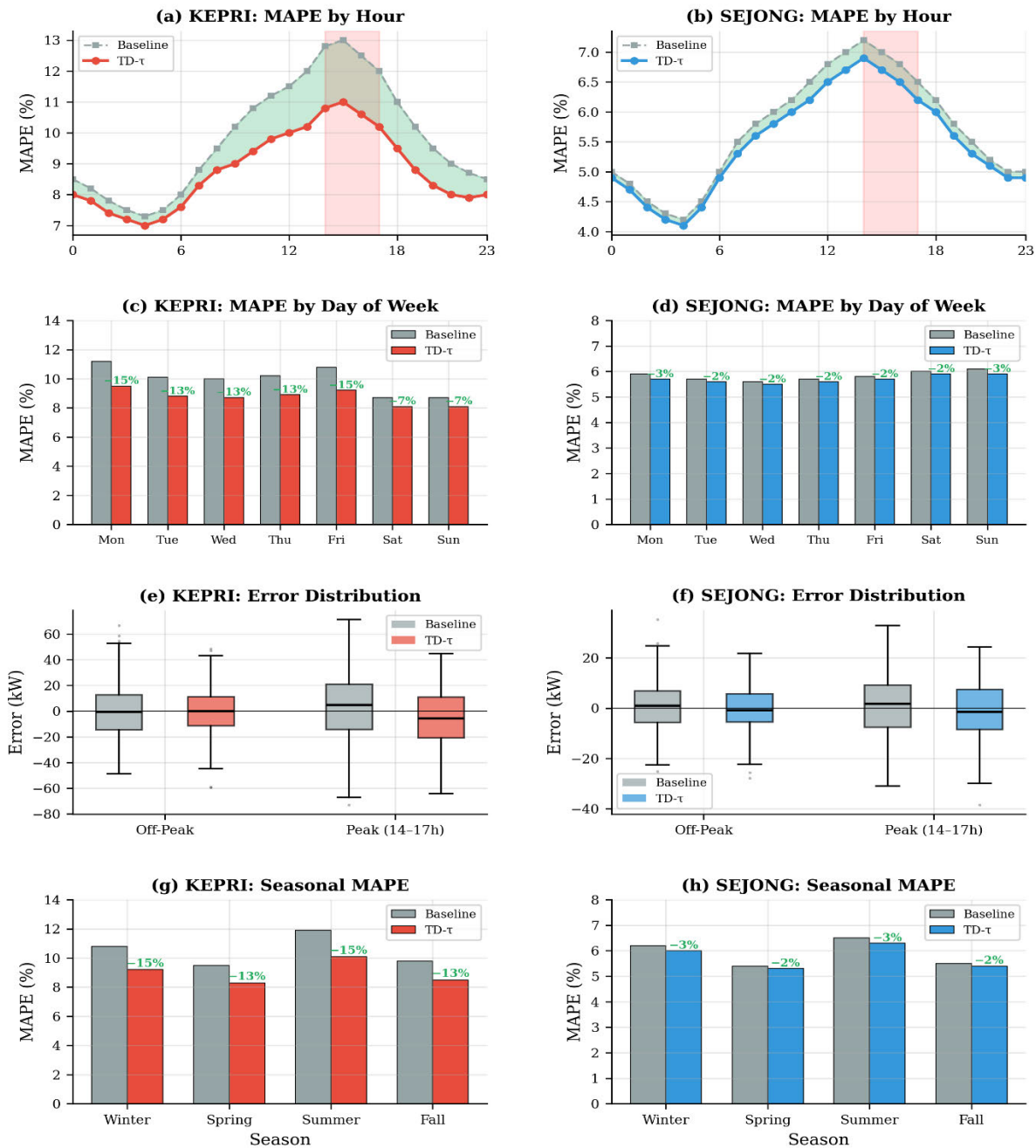


FIGURE 7. Temporal performance decomposition: (a)–(b) hourly MAPE by hour-of-day, (c)–(d) daily MAPE over the test horizon, (e)–(f) signed-error distributions stratified by peak vs. off-peak, and (g)–(h) seasonal MAPE. Left column: KEPRI (Type A); right column: SEJONG (Type B). Key message: the TD- τ improvement pattern is structured, not uniform—on KEPRI the TD- τ curve diverges from the symmetric baseline exactly at the 14–17 h on-peak window (panels a, c) and during high-demand summer/winter seasons (panel g), which is the empirical fingerprint of a mechanism that acts where it is designed to act. On SEJONG the curves track the baseline almost everywhere (panels b, d, h), consistent with the negligible Cohen’s $d = 0.09$ reported.

with the two-settlement cost structure—rather than aligning it post-hoc—reduces the economically dominant component of settlement cost (under-prediction) by 35% on a high-variability commercial building and 16% on a regular-pattern residential building, while holding aggregate MAPE approximately constant against a well-tuned fixed-quantile baseline. These reframing matters: the claim is not that TD- τ is a better accuracy method in the MAPE sense, but that it is

a better decision method under the actual cost structure, and that its accuracy is no worse.

A. NOVELTY RELATIVE TO FIXED ASYMMETRIC FORECASTING

Asymmetric loss for energy forecasting is not itself novel. What distinguishes this work from the prior literature on quantile and asymmetric loss is the explicit coupling of

three layers that are normally treated separately. First, the hour-specific cost-asymmetry ratio $R_t = (S_{RT,t} - S_{DA,t}) / (S_{DA,t} \gamma_{opp})$ is estimated directly from the market settlement rule, not chosen as a tunable hyperparameter. Second, the optimal quantile is derived by substitution into the Granger-optimal result, not tuned on held-out MAPE. Third, the resulting $\tau(h)$ schedule is time-varying on a one-hour resolution rather than a single constant, which is the granularity at which the underlying cost structure actually varies. A practitioner who adopts a fixed $\tau = 0.65$ implicitly assumes that R_t is constant across hours of the day; the on-peak/off-peak spread ratio contradicts this assumption by roughly a factor of three. The contribution of TD- τ is to make that mis-specification explicit and to close it. A fixed- τ baseline globally shifts the forecast upward and therefore trades under-prediction cost for over-prediction cost at a constant rate; TD- τ trades them at a rate that varies hour by hour, so the on-peak hours (where the trade is favorable) absorb most of the shift, and the off-peak hours (where the trade is roughly symmetric) absorb little. The effect is the 35.3% cost reduction on the on-peak 14–17h window, is 2–2.5 $\times$ the reduction in any other period. In the MAPE metric, which weights all hours equally, this hour-selective effect is averaged away.

B. FIXED VS. TIME-DEPENDENT ASYMMETRIC APPROACHES

Against the well-tuned fixed- τ baseline (CatBoost-Q65, $\tau = 0.65$), TD- τ is indistinguishable on aggregate MAPE (9.98% vs. 9.94% on KEPRI, $p = 0.795$, Cohen's $d = -0.17$) and we explicitly do not claim MAPE superiority there. TD- τ does however achieve superior peak-hour MAPE (10.78% vs. 11.43%) and 14.3% lower total settlement cost (1.38M vs. 1.61M KRW), with lower peak-hour error variance ($\sigma = 2.25$ vs. 2.66). Two models with nearly identical overall MAPE can therefore have substantially different economic value when their errors are distributed differently across time periods with different cost structures. The waterfall analysis decomposes the layered benefit: quantile training contributes 5.9% MAPE improvement over the symmetric baseline, and the time-dependent mechanism adds a further 10.9%, concentrated during on-peak hours (15.6% improvement at 14:00–17:00) where R_t is highest. The value added by time-dependence is therefore both theoretically principled (it restores consistency with the Granger-optimal quantile result at hour resolution) and empirically localized (it acts where it is supposed to act).

C. BUILDING MODERATION AND PRACTITIONER DEPLOYMENT

Building type is the dominant moderator of the framework's effectiveness, and this is a finding in its own right rather than a limitation to be glossed over. For the high-variability commercial building KEPRI (Type A: CV= 0.239, PR= 2.31), TD- τ delivers 16.3% MAPE improvement in the controlled single-fold ablation, 35.3% on-peak cost reduction,

and 23.5% CVaR₉₅ reduction against the symmetric baseline, with statistical significance versus symmetric training confirmed by the DM test (DM= -4.37, $p < 0.001$). For the regular-pattern residential building SEJONG (Type B: CV= 0.195, $\rho_{24} = 0.85$), improvement is modest (1.1% MAPE relative reduction, Cohen's $d = 0.09$, 9.5% cost) and not statistically significant against any gradient-boosting baseline ($p = 0.159$ vs. CatBoost, $p = 0.809$ vs. Q65). The interpretation is mechanistic: regular-pattern buildings already produce concentrated, near-symmetric error distributions, leaving less room for improvement through asymmetric loss adjustment.

1) PRACTITIONER SCREEN

On this basis we propose a simple pre-deployment decision rule: **(i)** Compute (CV, PR, ρ_{24}) on the last 12 months of hourly demand. **(ii)** If $CV > 0.20$ and $PR > 2.0$ (Type A), deploy TD- τ with the Type A conservative adjustment; expect 15–30% total settlement cost reduction with strongest gains in the on-peak window. **(iii)** If $CV \leq 0.20$ and $\rho_{24} > 0.80$ (Type B), deploy the standard TD- τ schedule without the Type A offset; expect smaller (5–10%) total cost reduction concentrated on the upper tail (CVaR), not on aggregate MAPE. **(iv)** In either case, re-estimate R_t from the previous 3–6 months of settlement data on a rolling window; because the schedule update only re-queries $\tau(h)$ at inference time, no retraining is required.

D. COMPUTE AND SYSTEM-LEVEL IMPLICATIONS

From the compute perspective, TD- τ carries essentially zero overhead over a fixed-quantile CatBoost baseline (~ 1.3 s training, $< 2\mu s$ per day-ahead hourly forecast), which removes the usual accuracy-vs.-cost trade-off from the deployment decision: the method is Pareto-dominant on the compute axis against both the fixed- τ baseline and the retuned deep-learning baselines.

By design, TD- τ produces systematic over-prediction (mean error $\approx -1.9\%$) since $\tau > 0.5$ shifts the forecast above the conditional median. This deliberate bias is the individually economically optimal response: the marginal cost of over-prediction ($\gamma_{opp} \cdot S_{DA} \approx 4\text{--}6\%$ of DA price) is far lower than under-prediction (RT procurement at 15–50% premium). However, this individual rationality creates a system-level externality that deserves explicit discussion. If many market participants simultaneously adopt strictly over-forecasting strategies, the aggregate day-ahead bid stack presented to the system operator would systematically exceed realized consumption, potentially inflating day-ahead clearing, distorting reserve scheduling, and reducing the informativeness of the aggregate forecast. Three partial mitigants are already built into the present framework: (i) the $\tau(h)$ schedule is bounded by the Granger-optimal quantile and does not push conservativeness beyond what the cost structure justifies; (ii) the method produces probabilistic forecasts (PICP $\approx 80\%$), which—if required by the operator—carry the full distribution rather than

only the biased point estimate; and (iii) on the SEJONG (Type B) building, TD- τ 's shift is small, so a portfolio dominated by regular-pattern buildings would not experience large aggregate drift. A proper market-design analysis of this externality, including penalty structures and mandatory probabilistic disclosure, is identified as a future research direction that is complementary to the technical framework presented here.

E. LIMITATIONS

Several limitations should be acknowledged with the framing they deserve. First, the two-building scope is a deliberately-scoped proof-of-concept and does not by itself establish external validity; the building-type moderation effect is documented on two points and should be read as a hypothesis for replication, not as a calibrated expected value over an unseen portfolio. Second, Type C buildings (irregular-schedule, low- ρ_{24} , high-PR) are defined for completeness of the taxonomy but are not empirically validated in this paper; their forecasting dynamics (non-stationary schedule, external schedule signals) require data we did not have access to, and we identify their evaluation as an explicit direction for future work rather than claiming generalization of the present results. Third, the SMP simulator is calibrated against KPX 2023 data while building datasets date from 2015–2018. Because relative model comparisons are all run on identical price paths, the ordering of methods is preserved; but absolute cost figures are indicative rather than definitive, and production deployment should integrate contemporaneous settlement data. Fourth, the static τ schedule cannot adapt online to structural market changes such as increasing renewable penetration, distributed-energy participation, or carbon pricing, all of which would alter R_t . The decoupling between pricing-risk estimation and quantile training makes a rolling re-estimation trivial at inference time, but a fully online adaptation (reinforcement-learning or Bayesian updating) is beyond the current scope. Finally, weather features and multi-horizon forecasting (24–48 hours ahead) are architecturally supported but not empirically exercised here; their integration is future work.

VII. CONCLUSION

This paper addressed the fundamental misalignment between how building energy forecasting models are trained optimizing symmetric loss functions that treat all errors equally and how forecasts are evaluated in practice through electricity market settlement, where under prediction and over-prediction carry dramatically different costs that vary by time of day. To bridge this gap, we proposed the Market-Aware Asymmetric Loss (MAL) framework, which integrates a Time-Dependent Quantile (TD- τ) mechanism, building pattern classification, a validated SMP simulator, and comprehensive economic and probabilistic evaluation into a unified forecasting system.

A. THEORETICAL CONTRIBUTION

We showed that the two-settlement cost structure induces an hour-specific cost-asymmetry ratio R_t whose Granger-optimal forecast is the $\tau^* = R_t/(1 + R_t)$ quantile of the conditional demand distribution, and that this result changes the prescription for the learning objective from a scalar τ to a schedule $\tau(h)$ at the same temporal resolution as the cost structure itself. This links a classical result in asymmetric-loss decision theory to a concrete operational forecasting pipeline, and it clarifies why fixed-quantile training—although a close MAPE competitor—is economically mis-specified against a diurnally varying R_t .

B. METHODOLOGICAL CONTRIBUTION

We instantiated the result above as a reproducible pipeline: a CatBoost quantile-regression learner trained once per τ level; an hour-indexed $\tau(h)$ query at inference; a seven-component SMP simulator validated to within 7% spread error over four diurnal periods and 12 distributional metrics; a three-metric building-pattern taxonomy that does double duty as deployment rule and as cross-building covariates; and an evaluation protocol that reports settlement cost decomposed into under- and over-prediction, CVaR₉₅, Cohen's d, Bonferroni-corrected DM significance, and full compute-cost accounting. Under this protocol, on two buildings with 3-fold time-series cross-validation, TD- τ reduces under-prediction settlement cost by 35% on a high-variability commercial building and 16% on a regular-pattern residential building relative to symmetric RMSE training, with negligible aggregate MAPE difference against a well-tuned fixed-quantile baseline (as explicitly quantified by the fold-level Cohen's d). Prediction intervals achieve PICP $\approx 80\%$ at PINAW 15–16%, and the framework runs in ~ 1.3 s of training and $< 2\mu$ s of per-forecast inference—no compute penalty against the fixed- τ baseline.

C. MANAGERIAL AND PRACTITIONER CONTRIBUTION

The paper's empirical findings translate into a concrete deployment procedure: building operators compute (CV, PR, ρ_{24}) from their historical meter data, apply the TD- τ schedule with or without the Type A offset, and re-estimate R_t on rolling market data without retraining. The building-type moderation effect documented in this paper is therefore not only scientific evidence but a specific, testable rule-of-thumb for portfolio screening. Regulators additionally gain a tool for quantifying the individual-vs.-system trade-off of cost-aware forecasting, since the framework natively emits probabilistic forecasts that can be disclosed alongside point bids.

D. FUTURE WORK

Three directions are prioritized: (i) a multi-site evaluation across 5–10 buildings spanning climate zones and a full Type C validation; (ii) online adaptation of $\tau(h)$ to non-stationary market regimes under increasing renewable penetration and carbon pricing; (iii) a market-design study

of the system-level externality from widespread adoption of over-forecasting strategies, including whether probabilistic disclosure requirements can recover aggregate forecast informativeness while preserving individual cost savings. Beyond the specific framework, this work advances the broader argument that forecasting systems operating within economic markets should be designed with cost structures in mind from inception—a transition from accuracy-centric to value-driven forecasting that, as electricity markets decentralize and building-level participants assume greater forecasting responsibility, will become not merely beneficial but essential.

REFERENCES

- [1] 2022 *Global Status Report for Buildings and Construction*, United Nations Environment Programme, Nairobi, Kenya, 2022.
- [2] Z. Wang and R. S. Srinivasan, "A review of artificial intelligence-based building energy use prediction," *Renew. Sustain. Energy Rev.*, vol. 75, pp. 796–808, Jul. 2017.
- [3] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, and X. Zhao, "A review of data-driven approaches for prediction and classification of building energy consumption," *Renew. Sustain. Energy Rev.*, vol. 82, pp. 1027–1047, Feb. 2018.
- [4] P. Siano, "Demand response and smart grids—a survey," *Renew. Sustain. Energy Rev.*, vol. 30, pp. 461–478, Jun. 2014.
- [5] M. H. Albadi and E. F. El-Saadany, "A summary of demand response in electricity markets," *Electric Power Syst. Res.*, vol. 78, no. 11, pp. 1989–1996, Nov. 2008.
- [6] N. G. Paterakis, O. Erdinc, and J. P. S. Catalão, "An overview of demand response: Key-elements and international experience," *Renew. Sustain. Energy Rev.*, vol. 69, pp. 871–891, Mar. 2017.
- [7] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 1192–1205, Jan. 2018.
- [8] C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah, "A review on time series forecasting techniques for building energy consumption," *Renew. Sustain. Energy Rev.*, vol. 74, pp. 902–924, Aug. 2017.
- [9] C. Robinson, B. Dilkina, J. Hubbs, W. Zhang, S. Guhathakurta, M. A. Brown, and R. M. Pendyala, "Machine learning approaches for estimating commercial building energy consumption," *Appl. Energy*, vol. 208, pp. 889–904, Aug. 2017.
- [10] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy Buildings*, vol. 147, pp. 77–89, Jul. 2017.
- [11] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68–80, Jun. 2012.
- [12] D. W. Bunn and E. D. Farmer, *Comparative Models for Electrical Load Forecasting*. Hoboken, NJ, USA: Wiley, 1985.
- [13] *Electricity Market Operation Rules*, Korea Power Exchange, Seoul, South Korea, 2024.
- [14] *Statistics of Electric Power in Korea*, Korea Electric Power Corporation, Seoul, South Korea, 2024.
- [15] R. Weron, "Electricity price forecasting: A review of the state-of-the-art with a look into the future," *Int. J. Forecasting*, vol. 30, no. 4, pp. 1030–1081, Oct. 2014.
- [16] A. J. Conejo, M. A. Plazas, R. Espinola, and A. B. Molina, "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 1035–1042, May 2005.
- [17] *Annual Report on Electricity Market 2023*, Korea Power Exchange, Seoul, South Korea, 2024.
- [18] S. H. Kim and H. S. Jung, "Analysis of price spike in Korean electricity market," *J. Korean Inst. Electr. Eng.*, vol. 68, no. 5, pp. 657–665, 2019.
- [19] J. Y. Park and S. K. Han, "Characteristic analysis of system marginal price in Korean electricity market," *Trans. Korean Inst. Electr. Eng.*, vol. 67, no. 8, pp. 1012–1019, 2018.
- [20] C. R. Knittel and M. R. Roberts, "An empirical examination of restructured electricity prices," *Energy Econ.*, vol. 27, no. 5, pp. 791–817, Sep. 2005.
- [21] *Electricity Regulatory Guidelines*, Ministry Trade, Ind. Energy, Sejong, South Korea, 2024.
- [22] C. W. J. Granger, "Prediction with a generalized cost of error function," *J. Oper. Res. Soc.*, vol. 20, no. 2, pp. 199–207, Jun. 1969.
- [23] P. F. Christoffersen and F. X. Diebold, "Optimal prediction under asymmetric loss," *Econ. Theory*, vol. 13, no. 6, pp. 808–817, Dec. 1997.
- [24] R. Koenker and G. Bassett, "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, Jan. 1978.
- [25] C. Croonenbroeck and G. Stadtmann, "Minimizing asymmetric loss in medium-term wind power forecasting," *Renew. Energy*, vol. 81, pp. 197–208, Sep. 2015.
- [26] Y. Chu, B. Li, and B. M. Hodge, "Online and batch methods for solar radiation forecast under asymmetric cost functions," *Renew. Energy*, vol. 91, pp. 199–210, May 2016.
- [27] Y.-C. Lin, C.-H. Chen, and C.-Y. Wu, "Asymmetric loss functions for contract capacity optimization," *Energies*, vol. 13, no. 12, p. 3123, 2020.
- [28] N. Amjadi, F. Keynia, and H. Zareipour, "Short-term load forecast of microgrids by a new bilevel prediction strategy," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 286–294, Dec. 2010.
- [29] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond," *Int. J. Forecasting*, vol. 32, no. 3, pp. 896–913, Jul. 2016.
- [30] F. Ziel and R. Weron, "Day-ahead electricity price forecasting with high-dimensional structures," *Energy Econ.*, vol. 70, pp. 396–420, Sep. 2018.
- [31] J. Nowotarski and R. Weron, "Recent advances in electricity price forecasting: A review of probabilistic forecasting," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 1548–1568, Jan. 2018.
- [32] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Amer. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, Mar. 2007.
- [33] T. Gneiting, "Making and evaluating point forecasts," *J. Amer. Stat. Assoc.*, vol. 106, no. 494, pp. 746–762, 2011.
- [34] S. B. Taieb, J. W. Taylor, and R. J. Hyndman, "Hierarchical probabilistic forecasting of electricity demand with smart meter data," *J. Amer. Stat. Assoc.*, vol. 116, no. 533, pp. 27–43, Jan. 2021.
- [35] C. Fan, F. Xiao, and S. Wang, "Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques," *Appl. Energy*, vol. 127, pp. 1–10, Aug. 2014.
- [36] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761–1768, 2007.
- [37] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. San Francisco, CA, USA: Holden-Day, 1970.
- [38] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, 2nd ed., New York, NY, USA: Springer, 2002.
- [39] D. Ruch and D. E. Claridge, "A four-parameter change-point model for predicting energy consumption in commercial buildings," *J. Sol. Energy Eng.*, vol. 114, no. 2, pp. 77–83, May 1992.
- [40] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970.
- [41] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [42] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
- [43] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
- [44] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3146–3154.
- [45] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 6638–6648.
- [46] B. Ribeiro and N. Antunes, "A comparative study of machine learning algorithms for building energy consumption prediction," *Energy Buildings*, vol. 258, Jul. 2022, Art. no. 111829.
- [47] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

- [48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
- [49] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, Oct. 2021.
- [50] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.
- [51] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PLoS ONE*, vol. 13, no. 3, Mar. 2018, Art. no. e0194889.
- [52] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 competition: 100,000 time series and 61 forecasting methods," *Int. J. Forecasting*, vol. 36, no. 1, pp. 54–74, Jan. 2020.
- [53] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," *Int. J. Forecasting*, vol. 38, no. 4, pp. 1346–1364, Oct. 2022.
- [54] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers (PatchTST)," in *Proc. Int. Conf. Learning Represent.*, 2023.
- [55] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "TimesNet: Temporal 2D-variation modeling for general time series analysis," in *Proc. Int. Conf. Learning Represent.*, 2023.
- [56] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "iTransformer: Inverted transformers are effective for time series forecasting," in *Proc. Int. Conf. Learning Represent.*, 2024.
- [57] C. H. Zo, J. D. Nzarigema, J. Ngarambe, S. Raj, S. Muhammad, G. Yoo, and G. Y. Yun, "Integrating deep learning into quantile regression models for enhanced building energy benchmarking," *J. Building Eng.*, vol. 109, Sep. 2025, Art. no. 113044.
- [58] Y. Zhang, H. Wen, Y. Bian, and Y. Shi, "Improving sequential market coordination via value-oriented renewable energy forecasting," 2024, *arXiv:2405.09004*.
- [59] A. Zellner, "Bayesian estimation and prediction using asymmetric loss functions," *J. Amer. Stat. Assoc.*, vol. 81, no. 394, pp. 446–451, Jun. 1986.
- [60] J. Nowotarski, B. Liu, R. Weron, and T. Hong, "Improving short term load forecast accuracy via combining sister forecasts," *Energy*, vol. 98, pp. 40–49, Mar. 2016.
- [61] Y. Wang, N. Zhang, Q. Chen, D. S. Kirschen, P. Li, and Q. Xia, "Data-driven probabilistic net load forecasting with high penetration of behind-the-meter PV," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3255–3264, May 2018.
- [62] T. Hong, J. Wilson, and J. Xie, "Long term probabilistic load forecasting and normalization with hourly information," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 456–462, Jan. 2014.
- [63] A. Casadei, C. Grupp, S. Bhaduri, L. Guo, W. Fung, R. Malshe, R. Ratan, A. Pole, and A. Rakshit, "Node-level financial optimization in demand forecasting through dynamic cost asymmetry and feedback mechanism," 2025, *arXiv:2512.19722*.
- [64] Z. Chen, H.-Z. Huang, Z. Deng, and J. Wu, "Shrinkage mamba relation network with out-of-distribution data augmentation for rotating machinery fault detection and localization under zero-faulty data," *Mech. Syst. Signal Process.*, vol. 224, Feb. 2025, Art. no. 112145.
- [65] Y. Wang, J. Li, and Q. Zhou, "Learning category-invariant disentangled features for domain generalization in machine fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024.
- [66] P. Pinson, "Wind energy: Forecasting challenges for its operational management," *Stat. Sci.*, vol. 28, no. 4, pp. 564–585, Nov. 2013.
- [67] R. Koenker, *Quantile Regression*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
- [68] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *Int. J. Forecasting*, vol. 30, no. 2, pp. 357–363, Apr. 2014.
- [69] V. Vovk, "Conditional validity of inductive conformal predictors," *Mach. Learn.*, vol. 92, nos. 2–3, pp. 349–376, 2013.
- [70] R. L. Winkler, "A decision-theoretic approach to interval estimation," *J. Amer. Stat. Assoc.*, vol. 67, no. 337, pp. 187–191, Mar. 1972.
- [71] Korea Power Exchange. (2023). *Market Information System—Hourly SMP Data*. [Online]. Available: <https://www.kpx.or.kr>

- [72] *2023 Annual Report on Electricity Market Operation*, Seoul, South Korea, 2024.
- [73] J. H. Kim, C. S. Park, and S. H. Kim, "Analysis of Korean electricity market price patterns using time series decomposition," *J. Electr. Eng. Technol.*, vol. 15, no. 4, pp. 1567–1576, 2020.



HYEONGYU SON (Member, IEEE) received the B.Sc. degree from the Department of Electrical Engineering, Kyungpook National University, Daegu, South Korea, in 2020, and the M.Sc. degree in electrical and electronic engineering from Kyungpook National University, in 2022, where he is currently pursuing the Ph.D. degree. His research interests include artificial neural network-based forecasting, system optimization, energy networks, building energy management, and V2G.



MUSU KIM (Member, IEEE) received the B.Sc. degree from the Department of Electrical Engineering, Kyungpook National University, Daegu, South Korea, in 2020, and the M.Sc. degree in electrical and electronic engineering from Kyungpook National University, in 2022, where he is currently pursuing the Ph.D. degree. His research interests include data science applications for battery charging/discharging optimization, battery lifetime prediction, vehicle behavior pattern analysis, energy management systems, and predictive modeling for electric vehicles.



YOUNGGYU HWANG (Member, IEEE) received the B.Sc. degree from the Department of Electrical Engineering, Kyungpook National University, Daegu, South Korea, in 2021, and the M.Sc. degree in electrical and electronic engineering from Kyungpook National University, in 2023, where he is currently pursuing the Ph.D. degree. His research interests include artificial neural network-based forecasting, system optimization, energy networks, energy management, and V2G.



SEKYUNG HAN (Member, IEEE) received the B.Sc. degree in electrical engineering and computer science from Hanyang University, Seoul, South Korea, in 2002, the M.Sc. degree from the School of Electrical Engineering and Computer Science, Seoul National University, Seoul, in 2007, and the Ph.D. degree in information science from The University of Tokyo, Tokyo, Japan, in 2012. He is currently an Associate Professor with Kyungpook National University, Daegu, South Korea. His research interests include system optimization, battery storage technology, energy network, energy management, and electric vehicles.

...