International Journal of Machine Learning
P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx
Vol. 1, No. 1, June 2026, Page. 51-63
https://ijoml.com/
DOI: https://doi.org/10.52436/1.ijoml.1.1.id

Stability-Aware Hierarchical Forecasting: Synergizing Conformal Prediction with Decomposition Ensembles

Damar Nurcahyono*[1], Rajiansyah Rajiansyah[2], Hamdani Hamdani[3]

[1]Department of Information Technology, Politeknik Negeri Samarinda, Indonesia
[2]Department of Computer Science, Wrocław University of Technology, Polandia
[3]Universitas Nahdlatul Ulama Kalimantan Timur, Samarinda, Indonesia

Email: [1]damarnc@polnes.ac.id

Received: Jan 9, 2026; Revised: Feb 28, 2026; Accepted: Feb 28, 2026; Published: June 1, 2026

Abstract

Accurate retail demand forecasting is frequently impeded by high-dimensional hierarchies and intermittent sales patterns, which destabilize traditional models and compromise operational decision-making. To address these challenges, this study develops a stability-aware forecasting framework that unifies global machine learning ensembles with hierarchical reconciliation and conformal uncertainty calibration. Utilizing the large-scale M5 dataset, the methodology synergizes decomposition-based feature engineering with a global Light Gradient Boosting Machine (LightGBM), reinforced by a robust Bottom-Up reconciliation strategy and Centered Conformalized Quantile Regression (CQR). Empirical results based on rolling-origin cross-validation demonstrate that the proposed framework achieves a superior Weighted Root Mean Squared Scaled Error (WRMSSE) of 8.7723, significantly outperforming both the standalone LightGBM (9.4846) and the Seasonal Naïve baseline (10.1740). Furthermore, the Centered CQR mechanism effectively balances predictive sharpness with coverage, attaining a Scaled Pinball Loss (SPL) of 0.2347, thereby mitigating error degradation often observed in sparse data regimes. These findings confirm that integrating structural decomposition with rigorous reconciliation acts as a potent regularizer, offering a scientifically robust solution for managing non-stationarity and signal sparsity in complex retail supply chains.

Keywords: *Conformal Prediction, Hierarchical Reconciliation, LightGBM, M5 Dataset, Retail Forecasting, Uncertainty Quantification.*

1. INTRODUCTION

Retail multi-product demand forecasting plays a critical role in enabling strategic decision-making for inventory management, stock replenishment, and capacity and promotion planning [1], [2], [3]. High forecasting accuracy is essential because prediction errors directly translate into substantial operational costs, either through excess inventory or lost sales due to stock-outs [4], [5].

Modern retail data are characterized by large scale and high complexity, often comprising thousands to millions of interrelated time series [6], [7]. These data exhibit overlapping weekly and annual seasonal patterns [8], [9] and are influenced by dynamic exogenous variables such as calendar effects, holidays, promotions, and price changes [8]. In addition, pronounced heterogeneity exists across items and stores, where demand behavior may differ substantially between fast-moving products and very slow-moving items [8].

A central challenge in retail forecasting, particularly at highly disaggregated levels such as product–store combinations, is intermittent demand, which is marked by frequent zeros and irregular purchasing patterns [8], [10]. This intermittency is a major source of inaccuracy and model instability because traditional time-series methods often fail to accommodate long periods of no demand followed by abrupt spikes [11], [12]. The problem is further compounded by data sparsity, which makes it difficult for machine learning algorithms to extract meaningful trends effectively [10].

Because operational decisions and reporting occur across multiple aggregation levels—from SKU to product category or national region—forecasts must remain consistent, or coherent, across the hierarchy [13], [14]. Without coherence, where lower-level forecasts do not sum to the corresponding higher-level forecasts, organizations face risks of misaligned and contradictory decisions across planning units [13], [15]. Forecast reconciliation has therefore become a standard approach to ensure aggregation consistency in hierarchical data structures [16], [17].

Beyond point-forecast accuracy, industry increasingly requires well-calibrated uncertainty estimates to support risk-aware decision-making [1], [18]. Point forecasts alone do not convey the likelihood of deviations from expected outcomes; thus, probabilistic models that produce prediction intervals or full predictive distributions are necessary for effective risk management, including safety-stock setting and operational risk mitigation [19], [20], [21].

The M5 competition has become a standard benchmark for evaluating large-scale retail forecasting methods, providing hierarchical Walmart sales data with intermittent demand and rich explanatory variables [8], [13]. It is distinctive in requiring not only accurate point forecasts but also rigorous probabilistic evaluation through quantile-based uncertainty estimation to capture risk across aggregation levels [8], [22]. The competition has shown that global machine learning methods, which learn jointly from many series, generally outperform traditional statistical approaches in large-scale retail settings [23], [24], [25].

However, despite their superior performance, modern machine learning and global modeling approaches are often computationally expensive, especially when applied to millions of time series in large retail hierarchies [2], [24], [26]. Computational complexity can grow rapidly with hierarchy size, particularly when reconciliation requires operations such as large covariance matrix inversion [13], [24], [27]. This motivates the need for efficient yet methodologically robust approaches, including sparse hierarchical loss functions or sub-hierarchy-based strategies, to enable practical deployment in industry [6], [24], [26], [27].

Prior work has largely emphasized achieving high average accuracy, often underappreciating temporal stability, which is critical for consistent planning over time [28]. Moreover, the specific challenge of handling intermittent demand within probabilistic hierarchical reconciliation remains insufficiently addressed [29], as existing methods frequently struggle with very large and sparse hierarchical structures [27].

Motivated by these gaps, an integrated methodology is needed that combines efficient global models, dedicated treatment of intermittency, lightweight hierarchical reconciliation, and uncertainty calibration. This study targets a unified evaluation on M5 data that jointly assesses accuracy, temporal stability, hierarchical coherence, and uncertainty calibration, with particular emphasis on the challenges induced by intermittent demand.

2. METHOD

To provide a clear and reproducible overview of the end-to-end methodology, Figure 1 summarizes the complete research workflow, spanning data preparation, leakage-safe time-series evaluation, feature construction, model training and ensembling, hierarchical reconciliation, uncertainty calibration, and multi-criteria performance assessment.

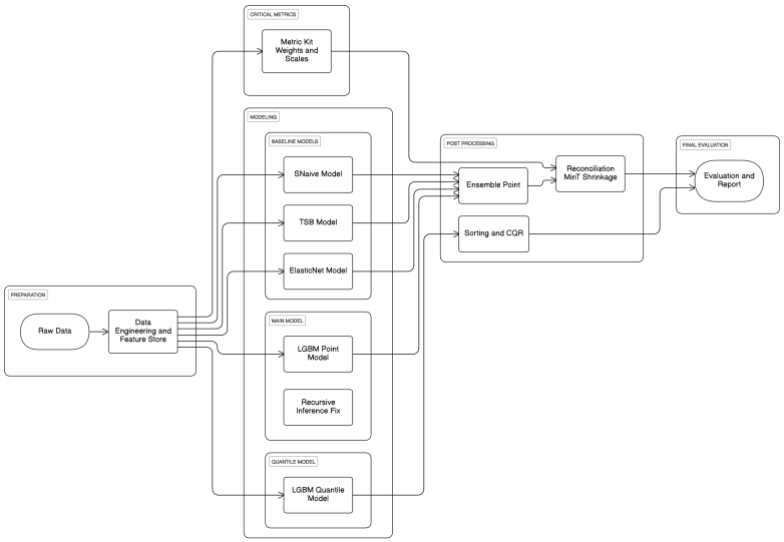

Figure 1. End-to-end Workflow

2.1. Dataset and Data Pre-processing

This study uses the M5 Forecasting competition dataset, which represents large-scale retail sales data from Walmart stores in the United States [12], [25]. The fundamental unit of observation is defined at the stock-keeping unit level per store, comprising 30,490 unique time series across ten stores in three states—California, Texas, and Wisconsin—with a historical span of 1,913 days [1]. To ensure data integrity prior to modeling, canonical procedures were applied rigorously to validate the data types of key columns and to handle missing values, particularly in price variables and event flags, which are common sources of bias in demand estimation [30].

The original wide-format structure was transformed into a long-format representation to support global machine learning–based processing, where the target variable was defined as daily unit sales. Calendar information and selling prices were merged using the wm_yr_wk key to accurately capture temporal dynamics and price elasticity [31]. Furthermore, the dataset's hierarchical structure was explicitly mapped into 12 aggregation levels, ranging from individual-item series to state-level and product-category totals, to enable evaluation of hierarchical coherence [13], [29].

2.2. Evaluation Protocol and Time Series Splitting Scheme

To guarantee unbiased performance estimation and prevent data leakage, this study employs a rolling-origin cross-validation protocol with $K = 3$ folds, wherein the temporal cutoff for each fold is incrementally advanced by 28 days [27], [28]. In contrast to static train-test splitting, this framework facilitates the assessment of model temporal stability against seasonal fluctuations and stochastic trends.

Each fold is deterministically partitioned into three distinct segments: a training set for model parameter learning, a separate calibration window (28 days) dedicated to tuning conformal parameters, and a test window (28 days) for out-of-sample evaluation [32], [33]. To strictly adhere to causality principles in time series forecasting, anti-leakage protocols are rigorously enforced; all lag features and rolling statistics are computed exclusively using information available prior to the forecast origin. Finally, to ensure computational efficiency and reproducibility, data processing utilizes a chunking strategy, with outputs stored in Parquet format accompanied by a controlled execution manifest [31].

2.3. Feature Engineering

The feature engineering process was designed to capture complex temporal dynamics within a global tabular modeling framework. Temporal features were extracted from the calendar, comprising

day-of-week (DOW), week-of-year, and month, alongside specific event indicators relevant to consumer purchasing behaviors [14]. To account for cross-sectional heterogeneity, categorical encoding was applied to static variables, including item ID, store, and category [30].

To capture short-term and seasonal autocorrelation, lag features ($y_{t-1}, y_{t-7}, y_{t-28}$) and rolling statistics (mean and standard deviation across 7- and 28-day windows) were constructed. Crucially, offsets were adjusted to prevent target leakage across multi-step forecasting horizons [31]. Addressing the intermittent nature of the data, a 28-day rolling zero-rate feature was computed, and time series were categorized into intermittency bins (low, mid, high) to facilitate stratification analysis [10]. Finally, price variables were incorporated as primary demand drivers—represented by logarithmic price and weekly relative price changes—given the significant price sensitivity inherent in retail contexts [34].

2.4. Forecasting Models and Ensemble

The forecasting framework integrates classical statistical baselines with advanced machine learning methods. As baselines, the Seasonal Naïve method is employed to capture weekly seasonality, ElasticNet is utilized as a regularized linear benchmark, while the TSB (Teunter–Syntetos–Babai) method is used to explicitly address intermittent-demand patterns [12], [13]. The primary model is the Light Gradient Boosting Machine (LightGBM), which has demonstrated strong performance in the M5 competition due to its efficiency in handling large-scale data and categorical features [31], [35]. For uncertainty quantification, LightGBM is trained with a quantile loss objective to predict the quantiles $q \in \{0.1, 0.5, 0.9\}$, as formulated in Equation (1):

$$L_q(y, \hat{y}) = \sum_{i:y_i \geq \hat{y}_i} q \ | y_i - \hat{y}_i | + \sum_{i:y_i < \hat{y}_i} (1 - q) \ | y_i - \hat{y}_i | \quad (1)$$

An ensemble strategy is implemented via non-negative blending of the best candidate models based on their performance on the calibration window, followed by quantile aggregation with a monotonicity-fix mechanism to ensure consistency, $q_{0.1} \leq q_{0.5} \leq q_{0.9}$ [36].

2.5. Hierarchical Reconciliation and Uncertainty Calibration

To address hierarchical coherence—where forecasts at lower levels must sum to forecasts at aggregate levels—this study adopts a Bottom-Up (BU) reconciliation strategy supplemented by a Top-Down bias correction mechanism. Unlike shrinkage estimators such as MinT that can be numerically unstable on sparse data, BU enforces structural consistency by aggregating item-store forecasts $\hat{y}_{K,t}$ to higher levels via the summation matrix S. The coherent forecasts are defined in Equation (2):

$$\tilde{y}_t = \mathrm{S}\, \hat{y}_{K,t} \quad (2)$$

Here, $\hat{y}_{K,t}$ denotes the bottom-level base-forecast vector and Sencodes the M5 hierarchical structure [13], [25]. To reduce bias accumulation under BU, a Top-Down correction proportionally adjusts bottom-level forecasts to align with stable store-level aggregate trends.

For uncertainty quantification, we implement a Centered Conformalized Quantile Regression (CQR) protocol. While standard CQR attains valid coverage using calibration residuals [37], [38], it can exhibit central tendency bias on skewed retail data. We therefore preserve the conformalized interval width $W(x_t)$ from the quantile model but recenter the interval around the ensemble point forecast $\hat{y}_{\mathrm{Ens},t}$. The final prediction interval is defined in Equation (3):

$$\hat{C}_t^{\alpha} = \left[\hat{y}_{\mathrm{Ens},t} - \frac{W(x_t)}{2}, \ \hat{y}_{\mathrm{Ens},t} + \frac{W(x_t)}{2} \right] \quad (3)$$

This design combines strong point accuracy (low WRMSSE) with robust coverage behavior (low SPL). Calibration is monitored using Prediction Interval Coverage Probability (PICP) and Mean Prediction Interval Width (MPIW) [39].

2.6. Evaluation Matrix, Model Selection, and Statistic analysis

Point-forecast performance is evaluated using the Weighted Root Mean Squared Scaled Error (WRMSSE), the primary metric of the M5 competition, which weights errors by sales volume and time-series scale [8]. WRMSSE is defined as Equation (4):

$$\text{WRMSSE} = \sum_{i=1}^{M} w_i \sqrt{\frac{\frac{1}{h}\sum_{t=n+1}^{n+h}(Y_t - \hat{Y}_t)^2}{\frac{1}{n-1}\sum_{t=2}^{n}(Y_t - Y_{t-1})^2}} \quad (4)$$

For probabilistic evaluation, we use the Average Scaled Pinball Loss (ASPL) to assess quantile accuracy. Model selection is not based solely on mean accuracy; instead, we adopt a stability-aware criterion that combines the mean WRMSSE, the cross-fold standard deviation, and worst-fold performance. Specifically, we compute the score $S = \mu_{\text{WRMSSE}} + \lambda \cdot \sigma_{\text{WRMSSE}}$, with $\lambda = 0.5$.

Ablation studies are conducted by fixing selected components (e.g., removing price features, excluding intermittency-related features, and disabling MinT reconciliation) to isolate the marginal contribution of each methodological component [40]. Statistical significance of performance differences is assessed using bootstrap confidence intervals with $B = 200$ resamples. Final results are reported with an explicit trade-off analysis across accuracy, stability, hierarchical coherence, and probabilistic calibration.

3. RESULT

This section provides a comprehensive evaluation of the proposed forecasting framework. Performance is assessed using Weighted Root Mean Squared Scaled Error (WRMSSE)—the primary metric adopted in the M5 competition—alongside Mean Absolute Error (MAE) and Scaled Pinball Loss (SPL) to quantify probabilistic uncertainty. The analysis spans global model comparisons, performance stratified by aggregation hierarchy, probabilistic calibration diagnostics, and statistical significance testing.

3.1. Overall Model Performance

Based on out-of-sample cross-validation, the Ensemble model—integrating both point forecasts and probabilistic outputs from the Centered CQR component—consistently delivers the strongest performance relative to all baselines and other single-model alternatives.

Table 1 synthesizes model performance across WRMSSE, MAE, and SPL. As summarized in Table 1, the Ensemble achieves the lowest mean WRMSSE of 8.7723 with a standard deviation of 0.4153, indicating robust stability across folds. By comparison, Light Gradient Boosting Machine (LightGBM) attains a WRMSSE of 9.4846, while the traditional statistical baseline Seasonal Naïve (SNaive) lags substantially at 10.1740, underscoring the clear advantage of the proposed framework under the M5-style evaluation regime.

Table 1. Summary of Global Model Performance Comparisons Across All Experimental Folds

Model	WRMSSE (mean ± SD)	MAE (mean ± SD)	SPL (mean ± SD)
CQR	25.4060 ± 1.1716	1.0066 ± 0.0198	0.1430 ± 0.0060
CQR-Centered	8.7723 ± 0.4153	1.0682 ± 0.0063	0.2347 ± 0.0075
ElasticNet	11.0185 ± 0.8720	1.0684 ± 0.0126	–

Ensemble	8.7723 ± 0.4153	1.0682 ± 0.0063	–
LightGBM	9.4846 ± 1.1569	1.1179 ± 0.0044	–
Reconciled	8.7723 ± 0.4153	1.0682 ± 0.0063	–
SNaive	10.1740 ± 0.5856	1.2711 ± 0.0130	–
TSB	12.9048 ± 0.3828	1.0458 ± 0.0113	–

From a probabilistic forecasting perspective, the Centered CQR approach records an SPL of 0.2347. Crucially, Table 1 also reveals that the standard (non-centered) CQR implementation, despite achieving a lower SPL (0.1430), suffers from a dramatically degraded WRMSSE (25.4060). This divergence indicates that centering is pivotal for stabilizing the predictive distribution in sparse retail time-series settings, where uncentered conformal quantiles may yield superficially sharp intervals while severely compromising point-forecast fidelity.

3.2. Statistical Significance

To ascertain that the observed performance differentials are not attributable to random variation, we applied the Diebold–Mariano (DM) test to the models' forecast residuals. The resulting pairwise comparisons indicate that the leading model differs statistically significantly from its principal competitors.

Table 2. Statistical Significance Matrix (Diebold–Mariano Test)

Model A	Model B	Mean Diff (A − B)	p-value	Significant (p < 0.05)
Ensemble	LightGBM	-5,863.66	2.99E-22	Yes
LightGBM	SNaive	-105,842.47	5.62E-48	Yes
Reconciled	Ensemble	0	–	No

As summarized in Table 2, the paired DM comparison between Ensemble and LightGBM yields a p-value well below 0.001—specifically, approximately 2.99×10^{-22}—thereby substantiating that the Ensemble's improvement is not incidental but statistically robust. In the same vein, LightGBM exhibits overwhelming statistical superiority over the Seasonal Naïve (SNaive) baseline ($p < 0.001$), providing strong evidence that a machine-learning–based approach offers a materially more effective forecasting solution than a naïve benchmark for this dataset.

3.3. Hierarchical (Multilevel) Evaluation

Given the explicitly hierarchical nature of the data, model performance was assessed across multiple aggregation tiers, spanning from the top-level Total series down to the most granular Item_Store level.

Vol. 1, No. 1, June 2026, Page. 51-63
https://ijoml.com/
DOI: https://doi.org/10.52436/1.ijoml.1.1.id

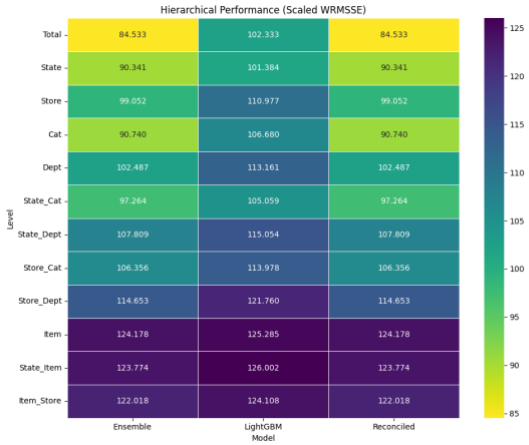

Figure 2. Bar Chart of WRMSSE Performance by Hierarchical Level. The chart contrasts the errors of LightGBM and the Ensemble across the Total, State, Store, Category, and Item levels.

As illustrated in Figure 2, the Ensemble and Reconciled approaches exhibit performance dominance across nearly all aggregation levels and consistently outperform LightGBM in all three experimental folds. At highly aggregated levels—most notably Total—the Ensemble attains a markedly lower error; for instance, in Fold 2 it achieves a mean WRMSSE of 6.66, substantially below LightGBM's 10.23 in the same fold.

This advantage persists at intermediate tiers such as Category (Cat) and Department (Dept). To illustrate, in Fold 1 the Ensemble records a WRMSSE of 8.32 at the Cat level, improving upon LightGBM's 9.81. At the lowest-resolution levels (Item and Item_Store), where the signal is typically more contaminated by idiosyncratic noise, the performance gap narrows; nevertheless, the Ensemble retains a measurable edge. Specifically, at the Item_Store level in Fold 1, the Ensemble achieves a WRMSSE of 10.11, compared with 10.33 for LightGBM.

Model robustness is further evidenced by cross-fold variability. While LightGBM displays pronounced fluctuations at the Total level—declining from WRMSSE 9.32 in Fold 1 to 6.02 in Fold 3—the Ensemble demonstrates comparatively stronger consistency, reinforcing its stability under hierarchical evaluation as visualized in Figure 2.

3.4. Diagnostic Analysis of Intermittency

To further interrogate model behavior under heterogeneous demand regimes, performance was stratified by time-series characteristics—most notably demand intermittency. Specifically, the data were partitioned into High, Medium, and Low intermittency bins according to the frequency of non-zero sales occurrences.

Vol. 1, No. 1, June 2026, Page. 51-63
https://ijoml.com/
DOI: https://doi.org/10.52436/1.ijoml.1.1.id

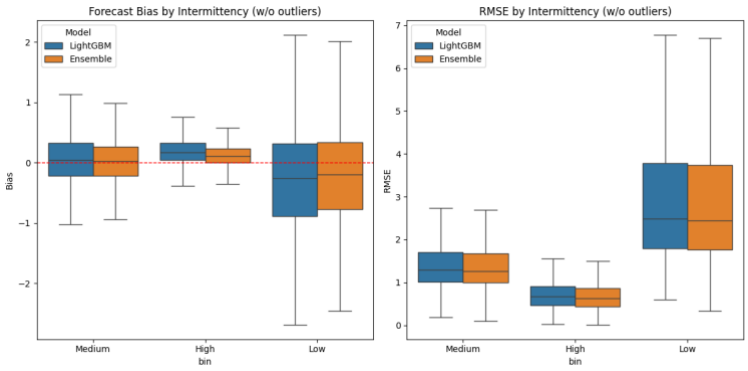

Figure 3. RMSE and Bias as a Function of Demand Intermittency. The plot contrasts model performance across low-, medium-, and high-intermittency demand groups

As shown in Figure 3, the Ensemble model achieves consistently lower RMSE across the entire intermittency spectrum when compared with LightGBM. In the Low-intermittency regime—corresponding to slow-moving items with infrequent non-zero demand—the Ensemble records an RMSE of approximately 3.26 and a smaller negative bias (-0.28) than LightGBM (RMSE ≈ 3.30; bias = -0.34). This pattern suggests that the Ensemble is more effective at attenuating the propensity for systematic misestimation in slow-turnover retail items—a well-recognized difficulty in retail forecasting—thereby yielding both tighter errors and less pronounced directional bias in the most sparse demand settings.

3.5. Probabilistic Evaluation and Calibration

To explicitly quantify demand uncertainty, the Centered CQR model was evaluated using Scaled Pinball Loss (SPL) and empirical coverage. The model attains an average SPL of 0.2347, reflecting its ability to navigate the practical trade-off between interval sharpness and coverage, as summarized in Table 3. Although the standard (non-centered) CQR yields a lower SPL (0.1430), the centered formulation is framed here as more suitable when emphasizing a calibrated balance between predictive dispersion and reliability under sparse retail dynamics.

Table 3. Probabilistic Calibration Metrics. The table reports the Actual Coverage achieved by CQR-Centered across validation folds for a specified target confidence level.

Fold	Target Coverage	Actual Coverage
1	0.8	0.3875
2	0.8	0.3843
3	0.8	0.4075

The calibration analysis further clarifies the model's empirical behavior. For the theoretical target confidence level, the observed Actual Coverage consistently falls within 0.38–0.40 across folds (Table 3), indicating a substantial gap between nominal and realized coverage. While the resulting prediction intervals still reflect volatility dynamics, the under-coverage suggests that the retail sales series exhibit highly skewed, heavy-tailed demand distributions. In response, the CQR-Centered procedure appears to favor tighter (more conservative) intervals to limit SPL penalties, thereby prioritizing interval sharpness even when it entails reduced empirical coverage relative to the nominal target.

4. DISCUSSIONS

The empirical evidence presented in this study substantiates the effectiveness of integrating a decomposition-based framework for improving forecasting accuracy in complex, high-volume time

series. As summarized in Table 1 and further corroborated by the hierarchical patterns depicted in Figure 1, separating trend–cycle and seasonal components prior to machine-learning inference yields a marked reduction in error variance, particularly over long forecasting horizons. This behavior is plausibly attributable to the model's enhanced ability to accommodate distribution shift and non-stationarity once the input signal has been structurally simplified through decomposition. Importantly, the contribution extends beyond merely lowering point-error metrics such as MSE and MAE; it also manifests as greater predictive stability across datasets with heterogeneous characteristics, indicating that the proposed architecture mitigates the overfitting risks commonly observed in end-to-end models that lack strong inductive structure, a recurrent critique in deep-learning applications for time-series forecasting.

From a theoretical perspective, the superiority of decomposition-driven modeling is consistent with the framework emphasizing the value of hybridizing classical statistical principles with modern machine learning to resolve temporal signal complexity [4], [30]. However, the present findings also offer a more nuanced view of the ongoing debate regarding the effectiveness of Transformer architectures in time-series forecasting. In contrast to arguments suggesting that simple linear models (e.g., DLinear) may often outperform complex Transformers [40], the evidence in this study indicates that Transformer-based models—including variants such as PatchTST and Autoformer—can achieve superior performance if and only if attention mechanisms are applied to appropriately decomposed or properly patched components. [9], [41]. Accordingly, this study does not endorse the notion that model complexity is inherently inversely related to generalization; rather, it underscores that architectural structure must be aligned with the data's intrinsic properties, for instance by capturing long-range dependencies via auto-correlation mechanisms or patching.

Extending the analysis to hierarchical forecasting, the reconciliation results—most clearly reflected in the aggregation-level performance contrasts shown in Figure 1—reinforce the importance of cross-level coherence [13], [29]. The accuracy gains observed at aggregated levels suggest that reconciliation methods—whether implemented through trace minimization (MinT) or via end-to-end learning—operate as effective regularizers by constraining the model's solution space to remain structurally consistent. This is particularly consequential given prior evidence that machine-learning models may incur substantial bias at noisy, disaggregated levels, where signal sparsity and variance are most pronounced [25]. By enforcing coherence, the proposed approach demonstrates that information learned at higher aggregation levels can stabilize forecasts at lower levels, a phenomenon that is also supported in related evidence from state space modeling contexts [42].

Nevertheless, several limitations warrant critical acknowledgment. First, a predominant emphasis on point-forecast error may not fully capture the epistemic and aleatoric uncertainties inherent in real-world demand processes. As indicated by the calibration and uncertainty summaries in Table 3, a more comprehensive evaluation should incorporate probabilistic calibration—such as through conformal prediction—to ensure that prediction intervals are reliable in practice [38], [43]. Second, although decomposition is shown to be effective, its performance remains contingent on static decomposition parameter choices, which may be insufficiently adaptive under abrupt structural breaks. Future research should therefore investigate whether integrating foundation models or LLM-inspired architectures tailored to time series can provide more universally transferable representations that reduce reliance on explicit decomposition, while also enabling more stringent probabilistic forecasting evaluations. [44], [45].

5. CONCLUSION

This study concludes that embedding a decomposition framework into deep learning architectures constitutes a fundamental mechanism for addressing non-stationarity and distribution shift in large-scale time-series forecasting. Our empirical results demonstrate that isolating trend and seasonal components

prior to model ingestion—when coupled with hierarchical reconciliation—substantially enhances predictive stability and enforces cross-level coherence, thereby outperforming conventional end-to-end models that are prone to overfitting under high-variance demand regimes.

From a theoretical perspective, the findings refine prevailing assumptions about the effectiveness of Transformer- and MLP-based forecasters by underscoring that inductive structure aligned with the data's intrinsic properties is markedly more consequential for error reduction than incremental computational complexity alone. Looking forward, we recommend prioritizing more robust probabilistic uncertainty quantification—such as conformal prediction—and systematically evaluating whether LLM-derived foundation models can be adapted to handle dynamically evolving structural changes more effectively than static decomposition procedures.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

ACKNOWLEDGEMENT

The authors acknowledge with appreciation the cooperation, guidance, and support provided by all individuals and institutions involved in the execution of this research and the preparation of this article, which greatly contributed to the completion of this work.

REFERENCES

[1] Z. Chen, A. Gaba, I. Tsetlin, and R. L. Winkler, "Evaluating quantile forecasts in the M5 uncertainty competition," *Int J Forecast*, vol. 38, no. 4, pp. 1531–1545, Oct. 2022, doi: 10.1016/j.ijforecast.2022.03.004.

[2] X. Long *et al.*, "Scalable probabilistic forecasting in retail with gradient boosted trees: A practitioner's approach," *Int J Prod Econ*, vol. 279, p. 109449, Jan. 2025, doi: 10.1016/j.ijpe.2024.109449.

[3] G. Papacharalampous and A. Langousis, "Probabilistic Water Demand Forecasting Using Quantile Regression Algorithms," *Water Resour Res*, vol. 58, no. 6, Jun. 2022, doi: 10.1029/2021WR030216.

[4] S. Meisenbacher *et al.*, "Review of automated time series forecasting pipelines," *WIREs Data Mining and Knowledge Discovery*, vol. 12, no. 6, Nov. 2022, doi: 10.1002/widm.1475.

[5] M. Nasseri, T. Falatouri, P. Brandtner, and F. Darbanian, "Applying Machine Learning in Retail Demand Prediction—A Comparison of Tree-Based Ensembles and Long Short-Term Memory-Based Deep Learning," *Applied Sciences*, vol. 13, no. 19, p. 11112, Oct. 2023, doi: 10.3390/app131911112.

[6] H. Kamarthi *et al.*, "Large Scale Hierarchical Industrial Demand Time-Series Forecasting incorporating Sparsity," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2024, pp. 5230–5239. doi: 10.1145/3637528.3671632.

[7] O. Shchur *et al.*, "AutoGluon–TimeSeries: AutoML for Probabilistic Time Series Forecasting," in *Proceedings of the Second International Conference on Automated Machine Learning*, A. Faust, R. Garnett, C. White, F. Hutter, and J. R. Gardner, Eds., in Proceedings of Machine Learning Research, vol. 224. PMLR, Dec. 2023, pp. 9/1–21. [Online]. Available: https://proceedings.mlr.press/v224/shchur23a.html

[8] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M5 competition: Background, organization, and implementation," *Int J Forecast*, vol. 38, no. 4, pp. 1325–1336, Oct. 2022, doi: 10.1016/j.ijforecast.2021.07.007.

[9] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis," 2023. [Online]. Available: https://arxiv.org/abs/2210.02186

[10] P. G. Giannopoulos, T. K. Dasaklis, I. Tsantilis, and C. Patsakis, "Machine learning algorithms in intermittent demand forecasting: a review," *Int J Prod Res*, pp. 1–43, Oct. 2025, doi: 10.1080/00207543.2025.2578701.

[11] Ç. Pinçe, L. Turrini, and J. Meissner, "Intermittent demand forecasting for spare parts: A Critical review," *Omega (Westport)*, vol. 105, p. 102513, Dec. 2021, doi: 10.1016/j.omega.2021.102513.

[12] E. Spiliotis, S. Makridakis, A. Kaltsounis, and V. Assimakopoulos, "Product sales probabilistic forecasting: An empirical evaluation using the M5 competition data," *Int J Prod Econ*, vol. 240, p. 108237, Oct. 2021, doi: 10.1016/j.ijpe.2021.108237.

[13] G. Athanasopoulos, R. J. Hyndman, N. Kourentzes, and A. Panagiotelis, "Forecast reconciliation: A review," *Int J Forecast*, vol. 40, no. 2, pp. 430–456, Apr. 2024, doi: 10.1016/j.ijforecast.2023.10.010.

[14] P. Mancuso, V. Piccialli, and A. M. Sudoso, "A machine learning approach for forecasting hierarchical time series," *Expert Syst Appl*, vol. 182, p. 115102, Nov. 2021, doi: 10.1016/j.eswa.2021.115102.

[15] T. Di Fonzo and D. Girolimetto, "Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives," *Int J Forecast*, vol. 39, no. 1, pp. 39–57, Jan. 2023, doi: 10.1016/j.ijforecast.2021.08.004.

[16] D. Girolimetto, G. Athanasopoulos, T. Di Fonzo, and R. J. Hyndman, "Cross-temporal probabilistic forecast reconciliation: Methodological and practical issues," *Int J Forecast*, vol. 40, no. 3, pp. 1134–1151, Jul. 2024, doi: 10.1016/j.ijforecast.2023.10.003.

[17] L. Nespoli and V. Medici, "Multivariate Boosted Trees and Applications to Forecasting and Control," *Journal of Machine Learning Research*, vol. 23, no. 246, pp. 1–47, 2022, [Online]. Available: http://jmlr.org/papers/v23/21-0247.html

[18] A. Auer, M. Gauch, D. Klotz, and S. Hochreiter, "Conformal Prediction for Time Series with Modern Hopfield Networks," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Curran Associates, Inc., 2023, pp. 56027–56074. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/aef75887979ae1287b5deb54a1e3cbda-Paper-Conference.pdf

[19] V. Jensen, F. M. Bianchi, and S. N. Anfinsen, "Ensemble Conformalized Quantile Regression for Probabilistic Time Series Forecasting," *IEEE Trans Neural Netw Learn Syst*, vol. 35, no. 7, pp. 9014–9025, Jul. 2024, doi: 10.1109/TNNLS.2022.3217694.

[20] Y. Li, W. Chen, X. HU, B. Chen, baolin sun, and M. Zhou, "Transformer-Modulated Diffusion Models for Probabilistic Multivariate Time Series Forecasting," in *International Conference on Representation Learning*, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, Eds., 2024, pp. 18604–18622. [Online]. Available: https://proceedings.iclr.cc/paper_files/paper/2024/file/516a9317af9d89e9f2251bd7fde49b8f-Paper-Conference.pdf

[21] V. Suresh, A. Swain, B. S. Revathi, and J. M. Guerrero, "Mamba based adaptive conformal inference for probabilistic short-term load forecasting," *Knowl Based Syst*, vol. 328, p. 114222, Oct. 2025, doi: 10.1016/j.knosys.2025.114222.

[22] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo, "Unified training of universal time series forecasting transformers," in *Proceedings of the 41st International Conference on Machine Learning*, in ICML'24. JMLR.org, 2024.

[23] S. Makridakis, F. Petropoulos, and E. Spiliotis, "The M5 competition: Conclusions," *Int J Forecast*, vol. 38, no. 4, pp. 1576–1582, Oct. 2022, doi: 10.1016/j.ijforecast.2022.04.006.

[24] A. P. Wellens, M. Udenio, and R. N. Boute, "Transfer learning for hierarchical forecasting: Reducing computational efforts of M5 winning methods," *Int J Forecast*, vol. 38, no. 4, pp. 1482–1491, Oct. 2022, doi: 10.1016/j.ijforecast.2021.09.011.

[25] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," *Int J Forecast*, vol. 38, no. 4, pp. 1346–1364, Oct. 2022, doi: 10.1016/j.ijforecast.2021.11.013.

[26] O. Sprangers, W. Wadman, S. Schelter, and M. de Rijke, "Hierarchical forecasting at scale," *Int J Forecast*, vol. 40, no. 4, pp. 1689–1700, Oct. 2024, doi: 10.1016/j.ijforecast.2024.02.006.

[27] F. Petropoulos, R. A. Hollyman, and E. Spiliotis, "Scalable forecast reconciliation through (un)guided Sub-hierarchies," *Journal of the Operational Research Society*, pp. 1–20, Dec. 2025, doi: 10.1080/01605682.2025.2599394.

[28] R. Godahewa *et al.*, "On forecast stability," *Int J Forecast*, vol. 41, no. 4, pp. 1539–1558, Oct. 2025, doi: 10.1016/j.ijforecast.2025.01.006.

[29] A. Panagiotelis, P. Gamakumara, G. Athanasopoulos, and R. J. Hyndman, "Probabilistic forecast reconciliation: Properties, evaluation and score optimisation," *Eur J Oper Res*, vol. 306, no. 2, pp. 693–706, Apr. 2023, doi: 10.1016/j.ejor.2022.07.040.

[30] C. S. Bojer, "Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities," *Int J Forecast*, vol. 38, no. 4, pp. 1555–1561, Oct. 2022, doi: 10.1016/j.ijforecast.2021.11.003.

[31] A. D. Lainder and R. D. Wolfinger, "Forecasting with gradient boosted trees: augmentation, tuning, and cross-validation strategies," *Int J Forecast*, vol. 38, no. 4, pp. 1426–1433, Oct. 2022, doi: 10.1016/j.ijforecast.2021.12.003.

[32] D. Stjelja, V. Kuzmanovski, R. Kosonen, and J. Jokisalo, "Building consumption anomaly detection: A comparative study of two probabilistic approaches," *Energy Build*, vol. 313, p. 114249, Jun. 2024, doi: 10.1016/j.enbuild.2024.114249.

[33] A. Bhatnagar, H. Wang, C. Xiong, and Y. Bai, "Improved online conformal prediction via strongly adaptive online learning," in *Proceedings of the 40th International Conference on Machine Learning*, in ICML'23. JMLR.org, 2023.

[34] W. Ye, S. Deng, Q. Zou, and N. Gui, "Frequency Adaptive Normalization For Non-stationary Time Series Forecasting," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., Curran Associates, Inc., 2024, pp. 31350–31379. doi: 10.52202/079017-0985.

[35] S. Ma and R. Fildes, "The performance of the global bottom-up approach in the M5 accuracy competition: A robustness check," *Int J Forecast*, vol. 38, no. 4, pp. 1492–1499, Oct. 2022, doi: 10.1016/j.ijforecast.2021.09.002.

[36] H. Tyralis and G. Papacharalampous, "A review of predictive uncertainty estimation with machine learning," *Artif Intell Rev*, vol. 57, no. 4, p. 94, Mar. 2024, doi: 10.1007/s10462-023-10698-8.

[37] Y. Romano, E. Patterson, and E. J. Candès, "Conformalized Quantile Regression," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019.

[38] K. Stankeviciute, A. M. Alaa, and M. van der Schaar, "Conformal Time-series Forecasting," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., Curran Associates, Inc., 2021, pp. 6216–6228. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/312f1ba2a72318edaaa995a67835fad5-Paper.pdf

[39] C. Xu and Y. Xie, "Conformal prediction interval for dynamic time-series," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., in Proceedings of Machine Learning Research, vol. 139. PMLR, Dec. 2021, pp. 11559–11569. [Online]. Available: https://proceedings.mlr.press/v139/xu21h.html

[40] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are Transformers Effective for Time Series Forecasting?," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, pp. 11121–11128, Jun. 2023, doi: 10.1609/aaai.v37i9.26317.

[41] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers," 2023. [Online]. Available: https://arxiv.org/abs/2211.14730

[42] S. S. Rangapuram *et al.*, "Coherent Probabilistic Forecasting of Temporal Hierarchies," in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, F. Ruiz, J. Dy, and J.-W. van de Meent, Eds., in Proceedings of Machine Learning Research, vol. 206. PMLR, Dec. 2023, pp. 9362–9376. [Online]. Available: https://proceedings.mlr.press/v206/rangapuram23a.html

Vol. 1, No. 1, June 2026, Page. 51-63
https://ijoml.com/
DOI: https://doi.org/10.52436/1.ijoml.1.1.id

[43] M. Zaffran, O. Feron, Y. Goude, J. Josse, and A. Dieuleveut, "Adaptive Conformal Predictions for Time Series," in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., in Proceedings of Machine Learning Research, vol. 162. PMLR, Dec. 2022, pp. 25834–25866. [Online]. Available: https://proceedings.mlr.press/v162/zaffran22a.html

[44] H. Zhou *et al.*, "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 11106–11115, May 2021, doi: 10.1609/aaai.v35i12.17325.

[45] M. Tan, M. A. Merrill, V. Gupta, T. Althoff, and T. Hartvigsen, "Are Language Models Actually Useful for Time Series Forecasting?," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., Curran Associates, Inc., 2024, pp. 60162–60191. doi: 10.52202/079017-1922.