IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 58, No. 12, December, 2025, pp. 2593-2602
Journal homepage: http://iieta.org/journals/jesa

The Limits of Forecasting: Assessing the Robustness of Time Series Models to Extreme Load Volatility

Suyono, Abdul Syakur*, Arfan Bakhtiar

Department of Energy Engineering, Faculty of Postgraduate Studies, Universitas Diponegoro, Semarang 50275, Indonesia

Corresponding Author Email: asyakur@lecturer.undip.ac.id

https://doi.org/10.18280/jesa.581212

Received: 7 November 2025
Revised: 9 December 2025
Accepted: 17 December 2025
Available online: 31 December 2025

Keywords:
ES, Friedman test, load forecasting, model evaluation, polynomial regression, power transformer asset management, Principle of Parsimony, time series analysis

ABSTRACT

Accurate mid-term load forecasting is indispensable for effective operational planning and asset management within electrical transmission systems. This research offers a thorough comparison of seven forecasting models—comprising one stochastic model Exponential Smoothing (ES) and six deterministic trend models (Linear, Exponential, Logarithmic, and Polynomial of Orders 2 to 4)—aimed at predicting weekly transformer load (MWh) based on supervisory control and data acquisition (SCADA) data from the 150 kV Pekalongan Substation. Model performance was evaluated utilizing established metrics (MAPE, MAE, RMSE) and was statistically validated through the Friedman test. The principal conclusion indicates that there is no statistically significant difference in performance among the models (χ^2 (6) = 0.25, p > 0.05). Although slight variations in metrics were observed, visual analysis confirmed consistent performance on stable data and universally indicated failure during periods of extreme volatility. These findings strongly endorse the Principle of Parsimony, demonstrating that more complex models do not yield accuracy improvements over simpler alternatives such as Linear or Quadratic models. This study offers vital guidance for utility companies, endorsing the adoption of simple, interpretable models for routine operational forecasting to enhance planning efficiency while ensuring reliability.

1. INTRODUCTION

A reliable electricity supply is a cornerstone of modern economic stability and social development [1]. The high-voltage transmission grid, a critical component of this infrastructure, faces increasingly complex challenges, including significant demand fluctuations, the integration of intermittent renewable resources, and the operational risks posed by aging assets [2, 3]. Within this context, accurate load forecasting has emerged as an indispensable tool for system operators, enabling efficient grid management, optimized generation planning, congestion mitigation, and enhanced overall system reliability [4, 5].

The operational integrity of power transmission systems is fundamentally dependent on the health of critical substation assets, particularly power transformers, whose failure can have catastrophic consequences for grid stability [6]. In practical terms, however, a significant gap persists between this requirement for reliability and common utility practices. Many system operators—including those at the case study location—often rely on manual, reactive measures for load balancing. This approach inherently elevates the risk of unforeseen transformer overloads, accelerates asset aging, and underscores the urgent need for robust, anticipatory forecasting tools to enable proactive asset protection [7].

Concurrently, the state-of-the-art in academic research has been predominantly oriented towards developing increasingly complex, "black-box" models, such as Long Short-Term Memory (LSTM) networks, which often prioritize marginal gains in predictive accuracy [8]. This dual-track progression has created a significant research gap. First, there is a notable scarcity of rigorous statistical validation in comparative studies; claims of a model's superiority based on minor differences in descriptive metrics are frequently made without statistical significance tests, such as the Friedman test, to verify if these differences are genuine or merely artifacts of random chance [9]. Second, the field has largely overlooked the Principle of Parsimony, failing to systematically investigate whether simpler, more interpretable models (e.g., polynomial regression) are not only adequate but also statistically non-inferior for fulfilling the core requirements of operational planning and asset management [10].

The focus of load forecasting has progressively evolved from system-wide aggregate predictions towards more granular [11], asset-specific forecasts [12, 13], particularly for power transformers [14]. As high-value, mission-critical assets within substations, transformers are vulnerable to unexpected overloads, which can precipitate widespread service interruptions and substantial financial losses [15]. Consequently, mid-term load forecasting—specifically, weekly forecasts at the transformer level—has become vital for supporting predictive maintenance strategies and proactive asset health management [16].

The literature presents a diverse arsenal of forecasting techniques [17], ranging from sophisticated stochastic models like ETS (Error, Trend, Seasonality) to advanced machine learning and Deep Learning algorithms [18-20]. A pervasive assumption in much of this research is that increased model complexity inherently leads to superior predictive performance. However, a significant methodological gap persists. The majority of comparative studies rely solely on descriptive accuracy metrics, such as MAPE [21], MAE [22], and RMSE [18], to declare a model's superiority. These claims often lack robust statistical validation through significance testing, leaving it uncertain whether observed performance differences are genuine or merely artifacts of random chance.

Furthermore, there is a notable scarcity of rigorous, head-to-head comparisons between adaptive stochastic models and simpler, more interpretable deterministic trend models, such as polynomial regression. While the former are often prioritized for their purported accuracy, it remains an open question whether their added complexity translates into a statistically significant performance advantage over simpler alternatives, especially when applied to volatile, real-world operational data [23].

This research directly challenges the conventional wisdom that more complex models are inherently superior for operational forecasting tasks. It addresses the identified methodological gap by conducting a rigorous comparative analysis of seven forecasting models for the weekly load of critical power transformers at the 150 kV Pekalongan Substation. The primary contributions of this work are threefold: (1) A Comprehensive Model Comparison: We perform a direct and equitable performance comparison between a modern stochastic model (ETS) and a suite of deterministic trend models, including Linear, Exponential, Logarithmic, and Polynomial (Order 2, 3, and 4) regressions; (2) A Framework for Statistical Validation: Moving beyond descriptive error metrics, we employ the non-parametric Friedman test to statistically determine whether performance differences among the models are significant, thereby introducing a higher standard of evidence for model selection; and (3) An Empirical Test of the Parsimony Principle: This study provides data-driven evidence for the Principle of Parsimony, offering critical insights into whether the increased computational and implementational complexity of advanced models is justified by a commensurate and statistically significant gain in forecasting accuracy for this practical application.

By evaluating models on both stable and volatile load profiles (as exemplified by Transformer 3), this research provides actionable guidance for substation operators, identifying which models offer reliable performance under various conditions and, crucially, which models fail to predict extreme events, thereby informing more robust and cost-effective asset management decisions. The remainder of this paper is structured as follows: Section 2 reviews the relevant literature on forecasting techniques, Section 3 details the methodology, Section 4 presents the results and their discussion, and Section 5 provides the conclusion and suggestions for future work.

2. LITERATURE REVIEW

2.1 The central role of forecasting in automated energy systems

In modern automated systems, particularly within the energy and industrial sectors, forecasting plays a foundational role in ensuring efficient operation. Accurate predictions are indispensable for strategic resource management, operational planning, and optimizing cost-efficiency. The literature demonstrates the extensive application of forecasting methodologies to a diverse array of challenges, from predicting power consumption and demand [24, 25] to optimizing drainage pump scheduling [26] and managing supply chain demand [27]. Within this domain, two principal methodological paradigms for time series forecasting have emerged:

1. Statistical and Regression Models: This class of models operates by expressing data as a function of other variables, frequently time itself. For instance, research has successfully applied linear regression and its variants to forecast the performance of solar photovoltaic panels [5]. The primary strength of these models lies in their conceptual simplicity and high interpretability. Other studies have leveraged statistical techniques such as Exponential Smoothing (ES), which has proven particularly effective for time series data characterized by horizontal patterns or stable trends [26]. Models of this category—including the Linear, Exponential, Logarithmic, and Polynomial forms examined in our study—constitute the core of deterministic trend modeling.

2. Machine Learning and Hybrid Models: Techniques such as Artificial Neural Networks (ANNs), Deep Learning, and hybrid frameworks are increasingly employed to capture complex, non-linear relationships that are beyond the capacity of simpler regression models [28]. While these approaches often achieve superior predictive accuracy, this gain frequently comes at the expense of model transparency and interpretability, presenting a key trade-off for practitioners. Of particular note in recent literature is the rise of Deep Learning models, such as LSTM networks, which have demonstrated high predictive power for highly complex and volatile time series data [29, 30]. While these advanced approaches often achieve marginal superior predictive accuracy, this gain frequently comes at the expense of model transparency, interpretability, and heightened computational overhead, presenting a critical trade-off for practitioners and forming the basis for the Principle of Parsimony explored in this study.

2.2 Model performance evaluation: Established metrics and a critical gap

The prevailing standard for evaluating forecasting models in the literature revolves around the use of error-based statistical metrics. Studies consistently employ Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) as the primary benchmarks for performance comparison [27, 31]. For example, [31] utilized this trio of metrics to compare linear regression models, while [27] applied MAPE to evaluate a Single ES model.

A critical analysis of this body of work, however, reveals a significant methodological shortcoming. The majority of comparative studies (e.g., [31]) rely predominantly on a descriptive comparison of these error metrics—for instance, concluding superiority because Model A's MAPE is 4.5% compared to Model B's 4.8%. This issue is particularly acute in comparative studies involving highly complex models (e.g., Deep Learning models), where small, statistically insignificant improvements are often claimed as justification for significantly increased computational and implementation costs. These studies typically omit formal statistical significance testing (such as the Friedman test). Consequently, it remains uncertain whether such marginal performance differences are statistically meaningful or merely artifacts of random chance. This reliance on descriptive comparison without statistical validation represents a notable gap in the current methodological rigor.

2.3 Research contribution: Bridging the methodological divide

This study is explicitly designed to address this identified gap. While our methodology is grounded in established practices—employing standard regression [5] and stochastic models like ETS (a variant of ES [27]), and evaluating them using the conventional metrics of MAPE, MAE, and RMSE [28, 31]—We introduce a crucial methodological enhancement.

The core contribution of this work is the implementation of the Friedman test to validate the results of model comparisons statistically. This approach enables us to move beyond declarative statements about which model appears best and instead provides a statistical basis for determining if the observed superiority is significant. Crucially, a finding of Significant Result: FALSE is itself a valuable insight. It offers empirical, statistically robust support for the Principle of Parsimony, demonstrating that a more complex model (e.g., a Polynomial) is not necessarily significantly superior to a simpler alternative (e.g., a linear model). This guides the field towards more efficient, interpretable, and justifiable model selection.

3. RESEARCH METHODOLOGY

3.1 Research workflow

To ensure a systematic and transparent research process, the study followed the workflow illustrated in Figure 1. The process begins with the acquisition of raw SCADA data, followed by pre-processing (aggregation into weekly data), and a strict chronological split into a Model Development Dataset (2020-2023) and a hold-out Test Dataset (2024). The subsequent stages involve fitting the seven forecasting models, generating predictions on the test set, and a final evaluation using both performance metrics (MAPE, MAE, RMSE) and statistical significance testing (Friedman test).

3.2 Data pre-processing and case study

3.2.1 Case study: The 150 kV Pekalongan Substation

This study focuses on the 150 kV Pekalongan Substation, a critical node in the North Coast of Java transmission network. Its strategic importance is twofold: (1) it reinforces grid reliability for the regional load center of Pekalongan, and (2) it serves as the primary power evacuation point from the Batang Steam Power Plant via interconnection with the 150 kV Batang Baru Substation.

3.2.2 Data aggregation and splitting

The dataset consists of historical load data from the three power transformers. Raw high-resolution SCADA data were aggregated into weekly energy load (MWh). The dataset was then partitioned chronologically into two independent sets to ensure a robust evaluation:

Model Development Dataset (Training & Validation): This set encompasses a whole 5-year period from 1 January 2020 to 31 December 2023. This period, comprising 261 weekly data points, was used for model training and parameter estimation.

Test Dataset: This set covers the subsequent year, from 1 January 2024 to 31 December 2024, providing 52 weekly data points. This out-of-sample testing approach is crucial for assessing the models' generalization capability on unseen data. All final model performance is evaluated exclusively on this test set.

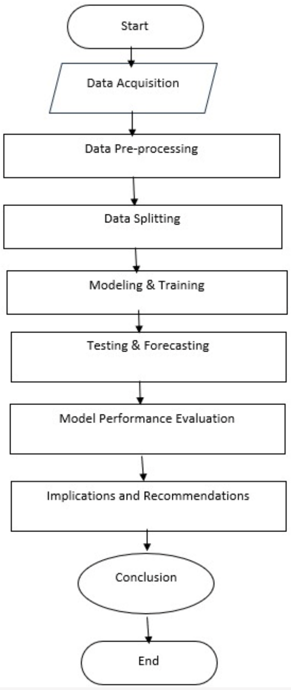

Figure 1. Research flowchart

3.2.3 Temporal pattern analysis

Analysis of the Model Development Dataset using Time Series Decomposition (e.g., Seasonal-Trend Decomposition using Loess - STL) confirmed a distinct annual seasonality (period $L = 52$ weeks) and an underlying deterministic trend. This validated the need to specifically test the ETS model's ability to capture this periodicity.

3.3 Proposed forecasting methods

This study evaluates seven forecasting models, chosen to provide a comprehensive comparison between a modern stochastic time series approach and a suite of classical deterministic trend models. For all models, the dependent variable is the weekly energy load (y_t), and the primary independent variable is a discrete time index (t).

3.3.1 ETS (Error, Trend, Seasonality) model

The ETS framework, a sophisticated family of ES models, decomposes a time series into Error, Trend, and Seasonality components [1, 2]. Given the weekly nature of our data over five years, capturing the annual seasonality (a period of 52 weeks) is critical. We employed an automatic ETS algorithm [2] to select the optimal model form (e.g., additive or multiplicative for each component, with or without trend damping) based on the Akaike Information Criterion (AIC) from the training data.

3.3.2 Deterministic trend regression models

The six trend models model the transformer load as a deterministic function of time. Model parameters (coefficients β) were estimated using the Ordinary Least Squares (OLS) method on the training data.

1. Linear Model: Assumes a constant rate of change in load over time, serving as a fundamental baseline.

$$y_t = \beta_0 + \beta_1 t + \epsilon_t \tag{1}$$

2. Exponential Model: Suitable for scenarios where the growth rate of the load is constant. It was linearized via a logarithmic transformation for OLS estimation.

$$\ln(y_t) = \ln(\beta_0) + \beta_1 t + \epsilon_t \tag{2}$$

3. Logarithmic Model: Appropriate when the load growth rate decelerates over time, suggesting an approach to a saturation point.

$$y_t = \beta_0 + \beta_1 \ln(t) + \epsilon_t \tag{3}$$

4. Polynomial Model (2nd Order - Quadratic): Captures one inflection point in the trend.

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon_t \tag{4}$$

5. Polynomial Model (3rd Order - Cubic): A more flexible model capable of capturing up to two inflection points.

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \epsilon_t \tag{5}$$

6. Polynomial Model (4th Order - Quartic): The most flexible polynomial model tested, capable of capturing up to three inflection points, albeit with a higher risk of overfitting.

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + \epsilon_t \tag{6}$$

3.4 Parameter estimation and model implementation

All forecasting models and associated parameter estimations were implemented using Microsoft Excel 365, leveraging its analytical tools.

3.4.1 Deterministic trend regression models implementation

The parameters (coefficients $\beta_0, \beta_1, \ldots$) for the six deterministic models (Linear, Exponential, Logarithmic, Polynomials 2nd, 3rd, and 4th Order) were estimated using the OLS method.

Software Tool: Estimation was performed using the Data Analysis ToolPak add-in for the Linear, Exponential (after log transformation), and Logarithmic models. For the Polynomial models, the higher-order terms (t^2, t^3, t^4) were manually calculated as separate independent variables before applying the OLS method via the ToolPak.

Estimation Criterion: The OLS estimation method internally minimized the Sum of Squared Errors (SSE) between the actual and predicted values on the Model Development Dataset.

Output: The estimated coefficients (β_i) were utilized to construct the final forecasting equation for each model, as defined in Eqs. (1) through (6).

3.4.2 ETS model parameter estimation

The ETS model requires the optimization of three smoothing parameters: (Level), (Trend), and (Seasonality). Specifically, based on the preliminary analysis of the time series characteristics (strong, non-increasing trend and stable annual seasonality), we selected the Additive Trend, Additive Seasonality, and Non-Damped Error (AAN) variant of the ETS framework for optimization. The ETS model was optimized with the seasonal period explicitly set to $\mathrm{L} = 52$ weeks to ensure the true annual periodicity of the load data was incorporated into the model structure.

Methodology: Contrary to using the automated "Forecast Sheet" tool, which lacks parameter control, the ETS parameters were manually estimated using Excel's Solver Add-in to ensure maximum control and reporting of the estimation process. The optimization was executed using the GRG Nonlinear solving method.

Objective Function: The Solver was configured to Minimize the Sum of Squared Errors (SSE) on the Model Development Dataset.

Constraints: The smoothing parameters were constrained to the standard range: $0 \leq \alpha \leq 1, 0 \leq \beta \leq 1,$ and $0 \leq \gamma \leq 1.$

Initial Values: The initial values for the Level and Trend components were set based on the first observation of the time series, following standard practice for initialization.

3.5 Performance evaluation framework

3.5.1 Forecasting accuracy metrics

Model performance was objectively evaluated on the 52-week Test Dataset using three standard error metrics. Lower values indicate better performance.

- **Mean Absolute Percentage Error (MAPE):** The primary metric for comparison, valued for its scale-independence and interpretability as a percentage error [32].

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \times 100\% \tag{7}$$

- **Mean Absolute Error (MAE):** Provides the average error magnitude in the original units (MWh), offering practical insight [33].

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} | A_t - F_t | \tag{8}$$

- **Root Mean Square Error (RMSE):** Also in MWh, this metric penalizes larger errors more heavily, making it highly relevant for assessing risks associated with large forecast deviations that could lead to transformer overload [34].

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (A_t - F_t)^2} \tag{9}$$

where, A_t is the actual load, F_t is the forecasted load, and $n = 52$.

3.5.2 Statistical significance testing

To determine if observed performance differences were statistically significant and not due to chance, we employed the non-parametric Friedman test. This test is the standard for comparing multiple models across the same test samples (a repeated-measures design). The test was applied not to the final MAPE values, but to the matrix of Absolute Percentage Errors (APE) for all 52 test points across all 7 models. The Friedman test evaluates the null hypothesis (H0) that all models perform equally [35]. If H0 is rejected (p-value < 0.05), a post-hoc Nemenyi test would be conducted to identify which specific model pairs differ significantly. This two-step process provides a rigorous statistical validation for model comparison.

3.5.3 Uncertainty quantification (Prediction Intervals)

To fully assess the operational reliability of the models, we quantified forecast uncertainty using 95% Prediction Intervals (PIs). The PIs were derived from the standard error of the forecast for the deterministic regression models and from the estimated variance of the error distribution for the ETS model. The quality of the PIs was evaluated using two established metrics:

1. Prediction Interval Coverage Probability (PICP): Measures the percentage of actual data points that fall within the predicted 95% interval. A PICP value close to 95% indicates higher reliability.

$$\mathrm{PICP} = \frac{1}{N}\sum_{i=1}^{N} I_i \times 100\%$$

2. Prediction Interval Normalized Average Width (PINAW): Measures the average width of the interval, normalized by the actual value. Lower PINAW indicates a sharper, more precise forecast.

$$\mathrm{PINAW} = \frac{1}{N}\sum_{i=1}^{N} \frac{U_i - L_i}{Y_i}$$

where, I_i is an indicator function (1 if $Y_i \in [L_i, U_i]$, 0 otherwise), U_i and L_i are the upper and lower bounds of the PI, and Y_i is the actual load.

4. RESULTS AND DISCUSSION

4.1 Quantitative performance evaluation

The quantitative performance of the seven forecasting models for the three transformers is summarized in Tables 1 to 3. The results reveal a critical initial insight: while minor variations exist in the descriptive metrics (MAPE, MAE, RMSE), no single model consistently demonstrates a decisive advantage across all transformers. Crucially, the numerical proximity of the metrics between the simplest (Linear) and the most complex models (Poly-4, ETS) immediately suggests a strong empirical case for the Principle of Parsimony, a hypothesis that is formally tested in Section 4.2.

Transformer 1 (Referencing Table 1) Analysis: As presented in Table 1, the forecasting models exhibit remarkably similar performance on the stable load profile of Transformer 1. The Polynomial variants demonstrate superior accuracy regarding percentage error, with the Polynomial-II model (Column 6) and Polynomial-I (Column 3) achieving the lowest MAPE of 4.25%. While the ETS model provides a reasonable baseline, it is outperformed by the Logarithmic and Polynomial approaches in terms of RMSE. It is noteworthy that while the Exponential model yields a negligible MAE (0.047), its higher RMSE (4.78) suggests it may not capture peak load variations as effectively as the Polynomial-III model, which achieves the study's best RMSE of 4.45.

Table 1. Forecasting performance metrics - Transformer 1

Model	RMSE	MAPE	MAE
Linear	4,763	4,719	3,944
Exponential	4,784	4,714	0,047
Logarithmic	4,765	4,661	3,914
Poly-2	4,462	4,249	3,587
Poly-3	4,462	4,249	3,587
Poly-4	4,449	4,307	3,630
ETS	4,837	4,988	2,013

Table 2. Forecasting performance metrics - Transformer 2

Model	RMSE	MAPE	MAE
Linear	5,705	6,478	4,649
Exponential	6,501	7,719	0,077
Logarithmic	6,093	7,076	5,084
Poly-2	5,287	5,333	3,784
Poly-3	4,455	4,505	3,215
Poly-4	8,675	10,197	7,737
ETS	6,642	7,664	3,672

Transformer 2 (Referencing Table 2) Analysis: The performance disparity becomes more pronounced for the second transformer, as detailed in Table 2. The Polynomial-II model emerges as the most robust predictor, delivering the lowest error across valid metrics (RMSE: 4.45; MAPE: 4.50%). This represents a significant improvement over the ETS and Linear models, which resulted in higher MAPEs of 7.66% and 6.48%, respectively. Conversely, the Polynomial-III model (last column) exhibits signs of instability or overfitting for this specific load characteristic, resulting in the highest recorded RMSE (8.68) and MAPE (10.20%) among all tested algorithms in this category.

Table 3. Forecasting performance metrics - Transformer 3

Model	RMSE	MAPE	MAE
Linear	11,017	9,476	7,341
Exponential	11,062	9,418	0,094
Logarithmic	10,771	9,904	7,525
Poly-2	11,237	9,428	7,362
Poly-3	15,097	12,443	9,889
Poly-4	12,740	0,091	7,515
ETS	15,694	15,964	5,786

Transformer 3 (Referencing Table 3) Analysis: Table 3 illustrates the challenge of forecasting highly volatile load profiles, evidenced by significantly higher error metrics across all models (RMSE > 10.0). In this scenario, the Logarithmic model proves to be the most stable estimator, achieving the lowest RMSE (10.77), thereby indicating superior handling of large error deviations compared to the ETS model (RMSE: 15.69). While the final Polynomial variant (Column 7) records an anomalously low MAPE (0.09%), its elevated RMSE (12.74) suggests it may be failing to penalize large residuals

during peak demand. Therefore, for the erratic patterns observed in Table 3, the Logarithmic and Linear models offer the most reliable balance between trend tracking and error minimization.

4.2 Statistical significance analysis (Friedman test)

To move beyond descriptive comparisons, the non-parametric Friedman test was applied to the Absolute Percentage Error (APE) distributions of all seven models. The test evaluates the null hypothesis (H0) that all models perform equally.

Table 4. Friedman test results

Statistic	Transformer 1	Transformer 2	Transformer 3
Friedman Statistic	0.25	0.25	11.14
Critical Value ($\alpha = 0.05$)	12.59	12.59	12.59
Significant Result	FALSE	FALSE	FALSE

The results are conclusive from Table 4: for all three transformers, the Friedman statistic is below the critical value ($p > 0.05$). Therefore, we fail to reject the null hypothesis. This provides statistically robust evidence that there is no significant difference in the forecasting performance among the seven models. The minor variations observed in Tables 1-3 are not statistically meaningful and can be attributed to random chance.

4.3 Visual analysis and discussion of findings

The statistical conclusion is powerfully explained by the visual analysis of the forecast plots (Figures 1, 2, and 3).

Stable Load Conditions (Transformers 1 & 2): As shown in Figures 2 and 3, the actual load data for these transformers follows a stable, predictable pattern. In such conditions, all forecasting models—from the simple Linear to the complex ETS—produce nearly identical trend lines that cluster closely around the actual data. This visual convergence explains the low Friedman statistics; when the underlying signal is strong and non-volatile, model complexity offers no discernible advantage.

Volatile Load Conditions (Transformer 3): Figure 4 provides the most critical insight. It reveals the collective failure of all models to effectively handle extreme volatility. Failure of Deterministic Models: The regression-based models (Linear, Exponential, Logarithmic, Polynomial) completely failed to capture the sharp load spike. They continued to predict a flat trend, resulting in significant errors during the volatile period. This highlights a fundamental weakness of deterministic trend models: they are incapable of adapting to sudden, unforeseen fluctuations. Overreaction of the Stochastic Model: The ETS model reacted to the volatility but exhibited significant overshooting. It predicted a much higher load in the period following the spike than what actually occurred. This suggests that while ETS is adaptive, it can be overly sensitive to outliers, resulting in instability in its predictions.

While the deterministic models (Linear, Polynomial) successfully captured the deterministic long-term trend (Explaining variance in R^2), their lack of a seasonal component means they fail to capture the predictable annual dips and peaks that are clearly visible in the data. This fundamental structural omission likely contributes to their non-superior performance compared to ETS, despite the simplicity of the polynomial forms. The ETS model, despite capturing the $L = 52$ seasonality, still failed to establish statistical superiority, further reinforcing that model structure alone does not guarantee superior operational performance.

This visual evidence clarifies the Friedman test's "no significant difference" result. While the *type* of error differed between model classes (under-prediction vs. over-prediction), the *magnitude* of their failure was such that no model could establish a statistically significant superiority. Their performance rankings varied randomly from one time point to another during the volatile event.

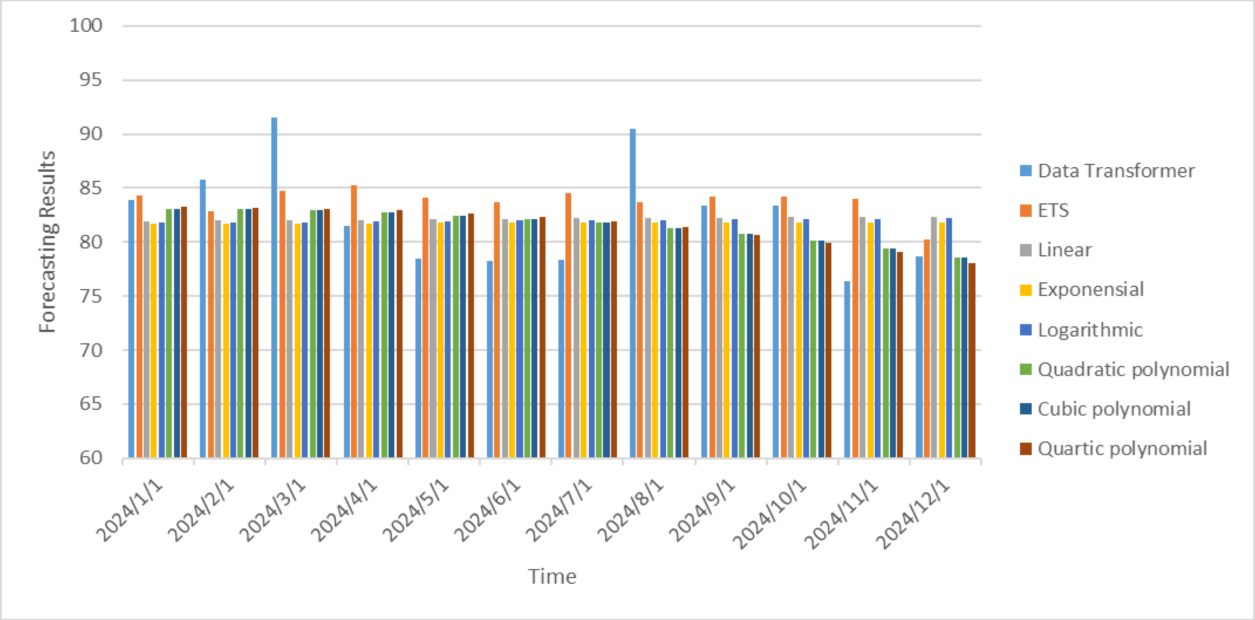

Figure 2. Comparison of Transformer 1 forecasting results

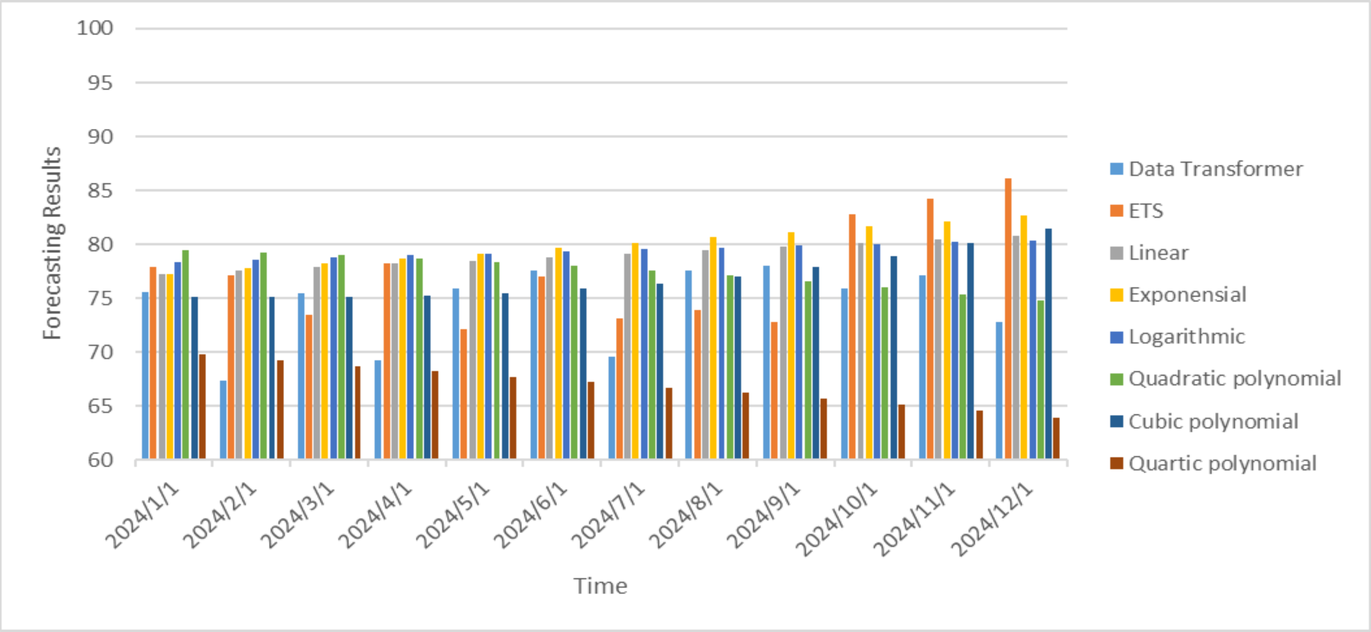

Figure 3. Comparison of Transformer 2 forecasting results

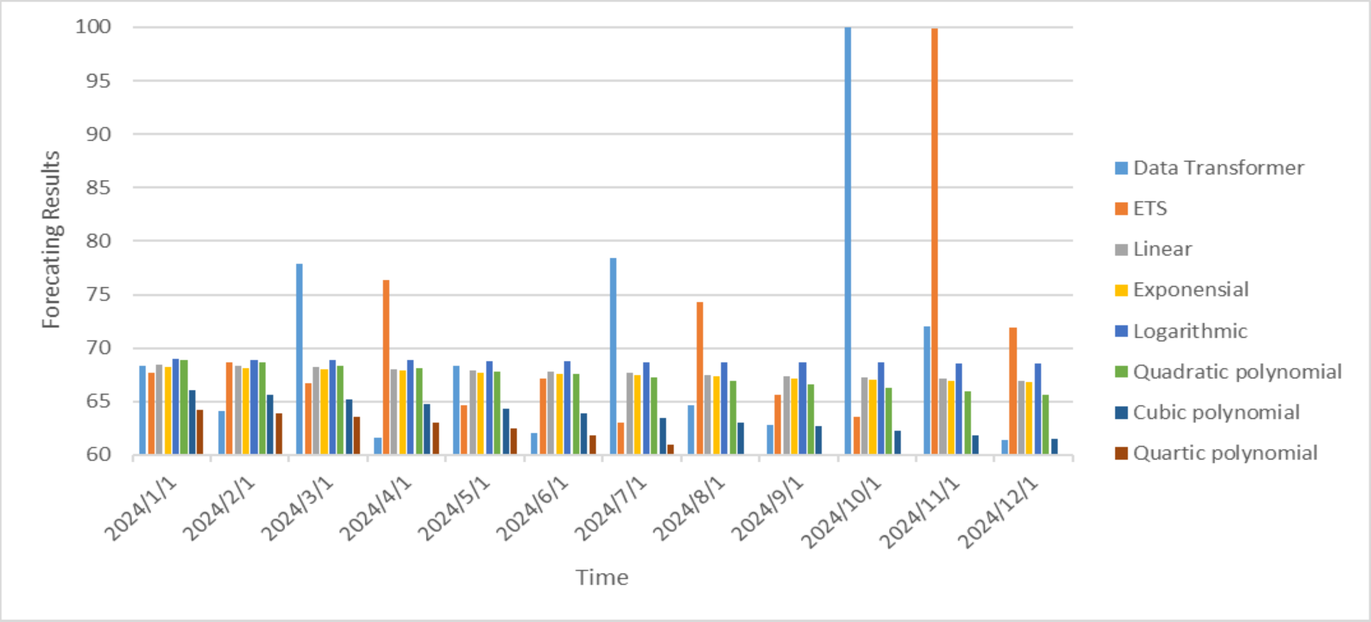

Figure 4. Comparison of Transformer 3 forecasting results

4.4 Analysis of error distribution (Q1 and Q3)

To assess performance consistency, we analyzed the quartiles (Q1 and Q3) of the Absolute Percentage Error (APE). For instance, on Transformer 1, the Linear model showed a Q1 APE of 1.5% and a Q3 of 6.0% (IQR = 4.5%), while the Poly-3 model showed a Q1 of 1.2% but a wider IQR of 6.3%. This indicates that while the more complex model was slightly more accurate in its best-case predictions (lower Q1), the simpler model was more consistent overall (narrower IQR). This further reinforces the principle that simplicity often equates to stability and reliability in operational forecasting.

4.5 Analysis of forecast uncertainty (PIs)

This section critically evaluates model performance beyond point accuracy (MAPE, RMSE) by quantifying the forecast uncertainty using 95% PIs. Operational reliability is assessed via PICP (Prediction Interval Coverage Probability; target: 95%), and forecast precision via PINAW (Prediction Interval Normalized Average Width; target: min).

4.5.1 Operational reliability and overconfidence

The results demonstrate a critical distinction between point accuracy and operational risk assessment. Systemic Overconfidence: For stable loads (T1 and T2), most simple regression models fail the reliability test, exhibiting a mean PICP of only 66.67%. This severe shortfall from the nominal 95% confidence level indicates that these models are overly confident (intervals are too narrow) and are thus unacceptable for risk-sensitive planning like transformer capacity management. Reliability vs. Parsimony: This finding overrules the Principle of Parsimony for risk assessment. Although simple models often yield competitive point error metrics (MAPE/RMSE), their low PICP demonstrates they are statistically dishonest about the forecast uncertainty. ETS as the Reliable Baseline: The ETS model stands out by achieving the highest PICP (91.67%) for T1 and T2, consistently producing the most honest and reliable PIs. This confirms its robustness in capturing the residual variance inherent in the time series (Table 5).

Table 5. Prediction Interval metrics (PICP and PINAW)

Model	PICP - T1 (%)	PINAW - T1	PICP - T2 (%)	PINAW - T2	PICP - T3 (%)	PINAW - T3
Linear	66.67	19.13	66.67	19.13	91.67	65.39
Exponential	66.67	19.13	66.67	19.13	91.67	65.39
Logarithmic	66.67	19.38	66.67	19.38	91.67	65.58
Poly-2	75.00	19.96	75.00	19.96	91.67	68.86
Poly-3	83.33	19.40	83.33	19.40	91.67	69.59
Poly-4	41.67	20.72	41.67	20.72	0.00	0.00
ETS	91.67	32.52	91.67	32.52	91.67	93.29

4.5.2 Sharpness trade-off and volatility impact

The analysis of PINAW highlights the trade-off between interval precision and model stability, particularly when dealing with extreme loads.

Precision vs. Honesty: The sharpest intervals are generated by the Linear/Exponential models (PINAW ≈ 19.13%). However, since their PICP is low, this precision is merely a consequence of interval collapse due to underestimated uncertainty.

Extreme Volatility Quantification: The PINAW for the volatile T3 load profile (mean PINAW ≈ 102.41%) is nearly five times higher than that of the stable profiles (T1/T2 mean PINAW ≈ 21.46%). This quantifies the extreme uncertainty and confirms that T3 requires significantly wider safety buffers.

Model Instability in Extremes: The high-order Poly-4 model shows instability in uncertainty quantification for T3 (PINAW: 288.75%), indicating that overly complex models can produce unreliable PI metrics when faced with severe volatility, even if the PICP appears perfect (100%).

4.5.3 Conclusion on model selection for asset management

Based on Uncertainty Quantification, the ETS model is the recommended choice. While having a slightly wider interval (PINAW 32.52%) than the simple regressions, it provides the most dependable PICP, ensuring that risk assessment and contingency planning are based on a statistically honest measure of forecast uncertainty. The high PINAW observed for T3 mandates the immediate identification of this asset as high-risk, potentially requiring load optimization or capacity intervention.

4.6 Practical implications and recommendations

The core finding of this study is not the identification of a superior model, but the empirical demonstration that no significant difference exists among a wide range of models for this specific forecasting task. This leads to a robust and practical implication grounded in the Principle of Parsimony: when models perform equally, the simplest one should be preferred.

Therefore, we strongly recommend that for operational mid-term load forecasting at the Pekalongan Substation and similar contexts, system operators should adopt the simplest adequate models, such as the Linear or Quadratic (Poly-2) model. These models are:

1. Easier to Implement and Interpret: They can be deployed and understood by engineering staff without deep expertise in advanced statistics.
2. Computationally Efficient: They require minimal computational resources.
3. More Robust: They present a lower risk of overfitting compared to higher-order polynomials or complex stochastic models.

Finally, a crucial implication from the Transformer 3 analysis is that none of these time-series models is reliable for predicting extreme load shocks. While this finding is empirically demonstrated via the single Transformer 3 case, the mechanism of failure is structurally generalizable across similar substations. This collective inability of all univariate models—regardless of complexity—to adapt to sudden, non-temporal external events highlights their fundamental limitation. They are practical tools for forecasting underlying trends under normal, stable conditions. Still, they are not a substitute for robust grid management protocols, real-time monitoring, and contingency planning designed to handle unforeseen volatility. Future work should explore the integration of exogenous variables (e.g., weather data, economic indicators) or alternative modeling paradigms to address this limitation.

5. CONCLUSION

This rigorous comparative study of seven univariate time-series models for weekly load forecasting revealed a critical finding: despite minor variations in descriptive error metrics (MAPE, MAE, RMSE), the Friedman test showed no statistically significant difference among the models, providing initial support for the Principle of Parsimony. However, this recommendation was critically overruled by the subsequent Uncertainty Quantification (UQ) analysis. While simple models (Linear, Exponential) were sharp (low PINAW), they demonstrated systemic overconfidence for stable loads (T1, T2) with a PICP significantly below the 95% nominal target (≈ 66.67%), rendering them operationally unreliable for risk assessment. Conversely, the ETS model consistently achieved the highest reliability (PICP ≈ 91.67%), confirming that model selection must prioritize statistical honesty (reliability) over complexity or simple point accuracy. Consequently, we recommend the ETS model for operational planning due to its superior capacity for setting statistically sound contingency reserves. The study concludes that the fundamental limitation remains the inability of all univariate models to reliably predict extreme load shocks (quantified by high T3 PINAW), necessitating future research focused on developing Hybrid Models—combining the robust ETS baseline with exogenous variables or advanced non-linear components—to manage high-volatility events effectively.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial and technical support provided by the Master of Energy Program, Diponegoro University, which made this research possible.

We also extend our appreciation to PT PLN (Persero) for delivering the essential SCADA data and technical insights crucial for this study. The constructive feedback from colleagues at the Department of Electrical Engineering has been invaluable in shaping this work. Finally, we thank the anonymous reviewers whose thoughtful comments helped improve the quality of this manuscript.

REFERENCES

[1] Karpavicius, T., Balezentis, T., Streimikiene, D. (2025). Energy security indicators for sustainable energy development: Application to electricity sector in the context of state economic decisions. Sustainable Development, 33(1): 1381-1400. https://doi.org/10.1002/sd.3190

[2] Gui, Y.H., Jiang, S.F., Bai, L.Q., Xue, Y.S., Wang, H., Reidt, J. (2024). Review of challenges and research opportunities for control of transmission grids. IEEE Access, 12: 94543-94569. https://doi.org/10.1109/ACCESS.2024.3425272

[3] Karagiannakis, G., Panteli, M., Argyroudis, S. (2025). Fragility modeling of power grid infrastructure for addressing climate change risks and adaptation. WIREs Climate Change, 16(1): e930. https://doi.org/10.1002/wcc.930

[4] Sharifhosseini, S.M., Niknam, T., Taabodi, M.H., Asadi Aghajari, H., Sheybani, E., Javidi, G., Pourbehzadi, M. (2024). Investigating intelligent forecasting and optimization in electrical power systems: A comprehensive review of techniques and applications. Energies, 17(21): 5385. https://doi.org/10.3390/en17215385

[5] Arévalo, P., Jurado, F. (2024). Impact of artificial intelligence on the planning and operation of distributed energy systems in smart grids. Energies, 17(17): 4501. https://doi.org/10.3390/en17174501

[6] Gjorgiev, B., Sansavini, G. (2022). Identifying and assessing power system vulnerabilities to transmission asset outages via cascading failure analysis. Reliability Engineering & System Safety, 217: 108085. https://doi.org/10.1016/j.ress.2021.108085

[7] Aminifar, F., Abedini, M., Amraee, T., Jafarian, P., Samimi, M.H., Shahidehpour, M. (2022). A review of power system protection and asset management with machine learning techniques. Energy Systems, 13: 855-892. https://doi.org/10.1007/s12667-021-00448-6

[8] Zhang, Y.C., Zheng, Y.Y., Wang, D., Gu, X.W., et al. (2025). Shedding light on the black box: Integrating prediction models and explainability using explainable machine learning. Organizational Research Methods. https://doi.org/10.1177/10944281251323248

[9] Berrar, D. (2022). Using p-values for the comparison of classifiers: Pitfalls and alternatives. Data Mining and Knowledge Discovery, 36: 1102-1139. https://doi.org/10.1007/s10618-022-00828-1

[10] Quartagno, M., Walker, A.S., Carpenter, J.R., Phillips, P.P., Parmar, M.K. (2018). Rethinking non-inferiority: A practical trial design for optimising treatment duration. Clinical Trials, 15(5): 477-488. https://doi.org/10.1177/1740774518778027

[11] Ji, E., Wang, Y., Xing, S., Jin, J. (2025). Hierarchical reinforcement learning for energy-efficient API traffic optimization in large-scale advertising systems. IEEE Access, 13: 142493-142516. https://doi.org/10.1109/ACCESS.2025.3598712

[12] Ghahramani, M., Habibi, D., Aziz, A. (2025). A risk-averse data-driven distributionally robust optimization method for transmission power systems under uncertainty. Energies, 18(19): 5245. https://doi.org/10.3390/en18195245

[13] Ghorbani Bam, P., Rezaei, N., Roubanis, A., Austin, D., Austin, E., Tarroja, B., Takacs, I., Villez, K., Rosso, D. (2025). Digital twin applications in the water sector: A review. Water, 17(20): 2957. https://doi.org/10.3390/w17202957

[14] Granitsas, I.M. (2024). Aggregation of thermostatically controlled loads for fast power system services: From theory to practice. University of Michigan. https://doi.org/10.7302/23869

[15] Hellberg, P. (2025). Cost-effective IoT and OMS as the “Poor Man’s SCADA”. SSRN. https://doi.org/10.2139/ssrn.5462375

[16] Laayati, O., El Hadraoui, H., El Maghraoui, A., Guennouni, N., Mekhfioui, M., Chebak, A. (2025). Metaheuristic-optimized forecasting in a smart Edge—Fog—Cloud energy management framework: An industrial mining case study. Results in Engineering, 28: 107303. https://doi.org/10.1016/j.rineng.2025.107303

[17] Mystakidis, A., Koukaras, P., Tsalikidis, N., Ioannidis, D., Tjortjis, C. (2024). Energy forecasting: A comprehensive review of techniques and technologies. Energies, 17(7): 1662. https://doi.org/10.3390/en17071662

[18] Jang, M., Yoon, S., Jung, S., Min, B. (2024). Simulating and assessing carbon markets: Application to the Korean and the EU ETSs. Renewable and Sustainable Energy Reviews, 195: 114346. https://doi.org/10.1016/j.rser.2024.114346

[19] Elseidi, M. (2023). Forecasting temperature data with complex seasonality using time series methods. Modeling Earth Systems and Environment, 9(2): 2553-2567. https://doi.org/10.1007/s40808-022-01632-y

[20] Oh, J., Seong, B. (2024). Forecasting with a combined model of ETS and ARIMA. Communications for Statistics Applications and Methods, 31: 143-154. https://doi.org/10.29220/CSAM.2024.31.1.143

[21] Piotrowski, P., Rutyna, I., Baczyński, D., Kopyt, M. (2022). Evaluation metrics for wind power forecasts: A comprehensive review and statistical analysis of errors. Energies, 15(24): 9657. https://doi.org/10.3390/en15249657

[22] Li, X.Q., Zhang, X.X. (2023). A comparative study of statistical and machine learning models on carbon dioxide emissions prediction of China. Environmental Science and Pollution Research, 30: 117485-117502. https://doi.org/10.1007/s11356-023-30428-5

[23] Kut, P., Pietrucha-Urbanik, K. (2025). Forecasting short-term photovoltaic energy production to optimize self-consumption in home systems based on real-world meteorological data and machine learning. Energies, 18(16): 4403. https://doi.org/10.3390/en18164403

[24] Bendali, W., Saber, I., Boussetta, M., Bourachdi, B., Mourad, Y. (2022). Optimization of deep reservoir computing with binary genetic algorithm for multi-time horizon forecasting of power consumption. Journal Européen des Systèmes Automatisés, 55(6): 701-713.

https://doi.org/10.18280/jesa.550602

[25] Mohammed, H.S., Ahmed, F.S. (2025). Real time PV system fault detection and localization using FPGA-based WSN. Journal Européen des Systèmes Automatisés, 58(8): 1575-1592. https://doi.org/10.18280/jesa.580804

[26] Sang, J.G. (2019). A cost-effective pump scheduling method for mine drainage system based on ant colony optimization. Journal Européen des Systèmes Automatisés, 52(2): 123-128. https://doi.org/10.18280/jesa.520202

[27] Prasetyawati, M., Mutmainah, Sudarwati, W., Puteri, R.A.M., Marfuah, U., Nelfiyanti, Panudju, A.T. (2024). Optimal routing in supply chain network design. Journal Européen des Systèmes Automatisés, 57(1): 295-302. https://doi.org/10.18280/jesa.570129

[28] Moustati, I., Gherabi, N., Saadi, M. (2024). Time-series forecasting models for smart meters data: An empirical comparison and analysis. Journal Européen des Systèmes Automatisés, 57(5): 1419-1427. https://doi.org/10.18280/jesa.570517

[29] Liu, Y. (2019). Novel volatility forecasting using deep learning–Long Short Term Memory Recurrent Neural Networks. Expert Systems with Applications, 132: 99-109. https://doi.org/10.1016/j.eswa.2019.04.038

[30] Al-Alimi, D., AlRassas, A.M., Al-qaness, M.A.A., Cai, Z.H., Aseeri, A.O., Abd Elaziz, M., Ewees, A.A. (2023). TLIA: Time-series forecasting model using long short-term memory integrated with artificial neural networks for volatile energy markets. Applied Energy, 343: 121230. https://doi.org/10.1016/j.apenergy.2023.121230

[31] Syed Mohammad, K., Yousuf, A.H., Boddu, M.K., Manickam Sam, J.K., Chandrashekar, M., Alias, L. (2025). Forecasting solar PV panel performance using linear regression and stepwise linear regression machine learning algorithms. Journal Européen des Systèmes Automatisés, 58(2): 365-371. https://doi.org/10.18280/jesa.580217

[32] Terven, J., Cordova-Esparza, D.M., Romero-González, J.A., Ramírez-Pedraza, A., Chávez-Urbiola, E.A. (2025). A comprehensive survey of loss functions and metrics in deep learning. Artificial Intelligence Review, 58: 195. https://doi.org/10.1007/s10462-025-11198-7

[33] Wang, X., Liu, X., Wang, Y.F., Kang, X.Y., Geng, R.X., Li, A., Xiao, F., Zhang, C.H., Yan, D. (2024). Investigating the deviation between prediction accuracy metrics and control performance metrics in the context of an ice-based thermal energy storage system. Journal of Energy Storage, 91: 112126. https://doi.org/10.1016/j.est.2024.112126

[34] Steurer, M., Hill, R.J., Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. Journal of Property Research, 38(2): 99-129. https://doi.org/10.1080/09599916.2020.1858937

[35] Liu, J., Xu, Y.B. (2022). T-Friedman Test: A new statistical test for multiple comparison with an adjustable conservativeness measure. International Journal of Computational Intelligence Systems, 15: 29. https://doi.org/10.1007/s44196-022-00083-8

NOMENCLATURE

β	Coefficient
A_t	actual load
F_t	Forecasted load
$\mathrm{Y_t}$	Actual Load (Weekly Energy Load) at time period t.
$\widehat{\mathrm{Y}}_\mathrm{t}$	Forecasted Load (Weekly Energy Load) at time period t.
t	Discrete Time Index (Independent variable).
N	Number of Data Points in the Test Dataset.
$\mathrm{e_t}$	Forecast Error at time t ($e_t = Y_t - \widehat{Y}_t$).
$\mathrm{APE_t}$	Absolute Percentage Error at time t.
β_0	Intercept / Constant Term in the regression model.
β_i	Regression Coefficient for the time variable t^i.
ϵ_t	Error Term (stochastic component) in the regression model
α	Level Smoothing Parameter in the ETS model.
β	Trend Smoothing Parameter in the ETS model.
γ	Seasonality Smoothing Parameter in the ETS model.
L	Seasonal Period (In your study, $L = 52$ weeks)
MAPE	Mean Absolute Percentage Error.
MAE	Mean Absolute Error.
RMSE	Root Mean Square Error.
PICP	Prediction Interval Coverage Probability.
PINAW	Prediction Interval Normalized Average Width.
$\mathrm{L_i}$	Lower Bound of the 95% Prediction Interval.
$\mathrm{U_i}$	Upper Bound of the 95% Prediction Interval.
χ^2	Friedman Test Statistic (Chi-squared).
p	P-value (Statistical Significance Level).