

Impact of Input Sequence Length on Transformer-Based Time Series Forecasting: Extreme Weather Sensitivity

Mostafa Aliyari¹, Enrique Lopez Droguett², Javad Barabady³, Yonas Zewdu Ayele⁴

¹ Faculty of Computer Science, Engineering and Economics, Østfold University of Applied Sciences, Norway

² Department of Civil and Environmental Engineering, University of California, Los Angeles, USA

³ Department of Technology and Safety, UiT the Arctic University of Norway, Tromsø, Norway

⁴ Department of Built Environment, Oslo Metropolitan University (OsloMet), Norway

ABSTRACT

This study investigates how input sequence length affects Transformer-based (deep learning) short-term electricity load forecasting (2- and 6-hour horizons), comparing univariate and weather-informed multivariate configurations using 138,561 hourly records from Norway (2009–2024). A total of 400 input sequence lengths, spanning 2–800 hourly timesteps, are evaluated. Results show a consistent U-shaped relationship between sequence length and forecasting error (MAE and MSE), where performance improves up to an optimal range and degrades for excessively long sequences. Optimal lengths vary across configurations, but clear clustering is observed between approximately 100–502 timesteps, particularly for the inverted Transformer (iTransformer). Optimal sequence lengths yield up to 84% MAE reduction relative to the minimum sequence length ($L=2$) in each configuration (iTransformer, multivariate, 6-hour prediction horizon). Across all settings, iTransformer consistently outperforms the standard Transformer, achieving 31–38% lower forecasting errors (MAE and MSE) with tighter error distributions. Incorporating weather and seasonal features reduces errors by 30–71%, with the largest gains observed for iTransformer in multivariate settings (41% MAE reduction vs. 30% for Transformer). For MSE, multivariate inputs reduce error by up to 71% for Transformer (6-hour horizon) and 47% for iTransformer (2-hour horizon). All improvements are statistically significant ($p < 10^{-12}$) with large effect sizes (Cohen's d up to -6.41). Computational analysis shows that Transformer training time exhibits approximately square-root-like growth with sequence length ($T(L) \approx 563(L^{0.549})$, $R^2 = 0.774$), while iTransformer exhibits near flat or weakly dependent scaling ($b \approx 0$, $|b| < 0.039$), with low explanatory power ($R^2 < 0.094$ in all cases). At optimal configurations, iTransformer reduces training time by approximately $17\times$ ($\approx 1,721$ s vs. $\approx 29,493$ s) while achieving lower error (MAE: 0.027 vs. 0.042). For a tenfold increase in sequence length (100 to 1,000 timesteps), Transformer training time grows by a factor of $3.3\times$, whereas iTransformer remains essentially unchanged. These findings suggest a clear advantage of iTransformer for climate-adaptive energy forecasting, balancing accuracy, robustness, and computational efficiency, with optimal lookback windows concentrated between 100–502 timesteps for short-horizon prediction.

Keywords: Transformer, iTransformer, input sequence length, time series forecasting, electricity load forecasting, multivariate forecasting, extreme weather

¹ Faculty of Computer Science, Engineering and Economics, Østfold University of Applied Sciences, Norway

² Department of Civil and Environmental Engineering, University of California, Los Angeles, USA

³ Department of Technology and Safety, UiT the Arctic University of Norway, Tromsø, Norway

⁴ Department of Built Environment, Oslo Metropolitan University (OsloMet), Norway

1. INTRODUCTION

Time series forecasting plays a pivotal role in various domains, including energy management, finance, weather forecasting and healthcare, where accurate predictions enable informed decision-making and resource optimization [1]. In the energy sector, short-term electricity load forecasting is particularly critical for grid stability, demand-side management, and renewable integration, as it helps mitigate imbalances between supply and demand [2]. Moreover, significant operational cost reduction can be achieved by accurate power load forecasting, where studies have shown a 1% improvement in prediction error results in up to 10 million GBP reduction in electricity operational costs [3, 4]. Traditional statistical models, such as Autoregressive Integrated Moving Average (ARIMA) [5], have long been used due to their interpretability, but their performance often degrades with non-stationary data and is highly sensitive to input sequence length, with shorter sequences introducing significant bias [6]. Deep learning models have emerged as powerful alternatives for capturing complex temporal dependencies in time series data. Transformer architectures, originally introduced for natural language processing [7], have shown remarkable proficiency in time series forecasting by leveraging self-attention mechanisms to model long-range dependencies [8, 9]. Variants like the Informer [8] and Autoformer [9] have extended Transformers to handle extended sequences efficiently, while the inverted Transformer (iTransformer) further optimizes multivariate forecasting by inverting dimensions to better capture inter-series correlations [10]. These advancements have led to state-of-the-art results in electricity load prediction, where Transformers outperform traditional methods like ARIMA, support vector machine (SVM), random forest (RF) and other baselines [11, 12].

Despite these successes, a critical aspect of Transformer-based forecasting, the impact of input sequence length, remains underexplored, particularly in short-horizon predictions ahead. While longer sequences intuitively allow models to capture more historical context, excessive lengths can introduce noise, increase computational overhead, and degrade performance due to vanishing gradients or irrelevant dependencies [13-15]. Analogous challenges have been observed in other sequence-based machine learning tasks; for instance, in bioinformatics, Carvalho et al. (2015) demonstrated that shorter DNA sequences often yield superior promoter prediction accuracy and reduced processing time compared to longer ones when using classifiers like SVM and RF, highlighting how sequence length directly influences model complexity and generalization [16]. Similarly, in time series contexts, Qin et al. (2019) found that ARIMA forecasting accuracy peaks at intermediate lengths, with shorter series producing the highest biases and longer ones not always guaranteeing improvements due to increased autocorrelation and residual errors [6]. However, these insights from traditional or non-Transformer models do not directly translate to attention-based architectures, where sequence length interacts uniquely with self-attention layers [13, 14]. A systematic evaluation of sequence length in Transformers for time series is lacking, especially in energy applications where multivariate inputs could amplify its effects [17]. This gap is significant, as suboptimal sequence lengths may lead to inefficient training and inaccurate forecasts, undermining the practical deployment of these models.

Recent advances in extreme weather load forecasting have explored various deep learning architectures and weather integration strategies. Debnath et al. [18] demonstrated that Transformer models achieve 98.7% accuracy during demand spikes by leveraging self-attention to capture weather-load relationships, outperforming Long Short-Term Memory (LSTM) models that achieve 95.7% accuracy. Similarly, Unlu and Peña [19] evaluated hybrid deep learning architectures and reported that a hybrid Convolutional Neural Network--Gated Recurrent Unit (CNN--GRU)-based model achieved the best forecasting performance during hurricane events, with an RMSE of 9.112 MW compared to the other tested architectures. Complementary studies have also compared different AI models under extreme weather conditions, showing that LSTM can outperform traditional machine learning approaches such as Extreme Gradient Boosting (XGBoost) and SVM, although with unstable errors during extreme events [20]. In parallel, several works have explored alternative machine learning and hybrid strategies. Deng et al. [21] proposed a Bagging-XGBoost framework with mutual information analysis, reducing peak load forecasting errors by 3-10%. Other approaches focus on incorporating additional information sources, such as behind-the-meter solar integration [22], and meteorological variables for urban

load forecasting using Artificial Neural Networks (ANN) and Bagged Regression Trees [23]. Research has also investigated advanced data-driven strategies including Conditional Generative Adversarial Network (CGAN)-based data augmentation to address imbalanced extreme-weather samples [24], probabilistic forecasting frameworks considering weather uncertainty [25], and clustering-enhanced hybrid models for extreme-weather load analysis [26].

Despite these advances, two aspects remain largely unexplored for Transformer-based architectures: the interaction between input sequence length and forecasting performance, and the impact of feature selection (weather-informed multivariate versus single-variable electricity consumption). This study addresses this gap by systematically investigating the influence of input sequence length on the forecasting accuracy of Transformer and iTransformer models in univariate and multivariate electricity consumption prediction. The multivariate setup is augmented with seasonal indicators derived from sine and cosine transformations of temporal variables. The analysis is based on a dataset of 138,561 hourly electricity consumption records from Norway, covering the period from January 1, 2009, to October 27, 2024. A total of 400 sequence lengths, ranging from 2 to 800 timesteps, were evaluated.

The remainder of this paper is organized as follows. Section 2 describes the methodology, including the dataset, selected models, performance metrics, and experimental setup. Section 3 presents the results and findings. Section 4 concludes the paper and suggests future research directions.

2. METHODOLOGY

This section outlines the models used, the dataset, and the experimental configuration for evaluating the impact of input sequence length on Transformer-based forecasting.

2.1. DATA COLLECTION

The dataset comprises 138,561 hourly electricity consumption records from Norway spanning January 1, 2009, to October 27, 2024. It includes normalized electricity consumption, outdoor temperature, and cyclical temporal features encoded using sine and cosine transformations of day-of-year and hour-of-day to capture seasonal and diurnal patterns (Figure 1). Preprocessing involved min-max normalization, outlier removal, smoothing, anomaly correction, forward-filling of minor missing values, and duplicate elimination. The cleaned data were divided into training (70%), validation (10%), and testing (20%) subsets. Historical input sequences ranging from 2 to 800 hourly timesteps were evaluated for forecasting 2-hour and 6-hour horizons. In the univariate setup, only electricity consumption was used, whereas the multivariate configuration additionally incorporated temperature and cyclical temporal features. Further details about the dataset are provided in [13, 14].

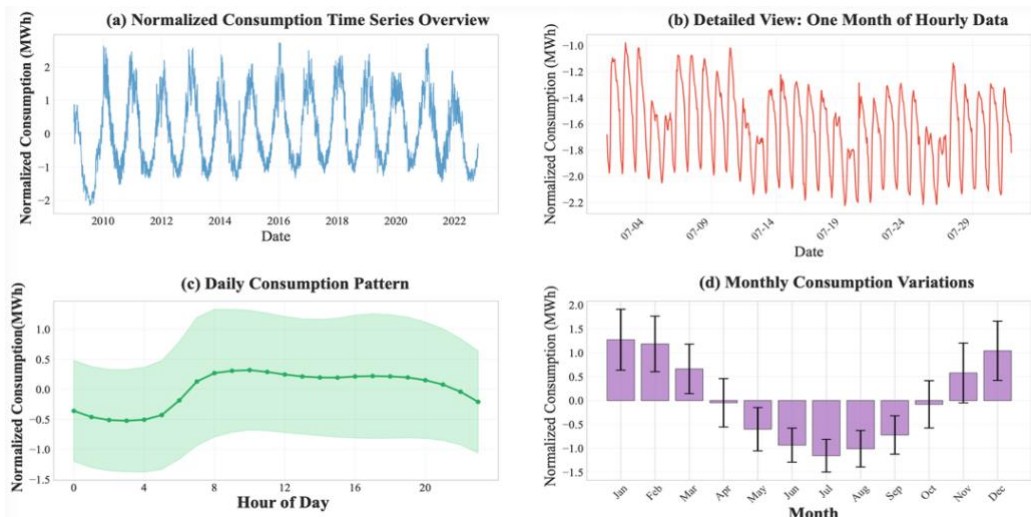


Figure 1. Overview of the hourly electricity consumption dataset used in this study.

2.2. SELECTED MODELS

This study focuses on Transformer-based architectures, specifically the original Transformer and the Inverted Transformer (iTransformer), to investigate the effect of input sequence length on their performance. The vanilla Transformer, referring to the original Transformer architecture introduced for sequence-to-sequence learning tasks, is adapted here for time series forecasting using an encoder-decoder structure [7]. Time steps are treated as tokens and processed through multi-head self-attention to capture long-range temporal dependencies, followed by position-wise feed-forward networks. Positional (time) embeddings preserve sequential order.

The iTransformer restructures the input representation without modifying the core Transformer blocks [10]. Instead of modeling time steps as tokens, it treats each variable's full historical sequence as a token. Self-attention is applied across these variate tokens to capture inter-variable dependencies, while feed-forward layers learn nonlinear representations per variable. This variate-centric design improves multivariate dependency modeling, enhances generalization across varying lookback windows, and achieves strong performance on real-world datasets.

2.3. Experimental Setups

To evaluate the effect of input sequence length, controlled experiments are conducted using Transformer and iTransformer models in both univariate (S) and multivariate (M) settings. Sequence length is systematically varied from 2 to 800 (step 2), while forecasting horizons are fixed at 2 and 6 hours. All other factors, including features and hyperparameters, are held constant to ensure fairness and reproducibility. The dataset is split into training, validation, and test sets. The validation set is used during training for early stopping based on validation loss, while the final performance metrics are reported on the test dataset. Models are trained and evaluated with MSE loss. Experiments are executed on an NVIDIA RTX A5000 (24 GB) with dual Intel Xeon Gold 6248R CPUs. Detailed settings are provided in Table 1.

Table 1. *Experimental Setup*

Forecasting Task	
Sequence lengths	2, 4, 6, ..., 800 (400 input sequences)
Prediction horizons	2 and 6 hours
Features	Multivariate (consumption + temperature + cyclical sine/cosine) & Univariate (consumption only)
Model Parameters (Shared)	
Encoder layers	2
Decoder layers	1
Number of heads	8
Model dimension (d_{model})	512
Feed-forward dimension	2048
Activation	GELU
Dropout	0.1
Embed time features	TRUE
Output attention	FALSE
Early stopping	TRUE
Run Parameters	
Train epochs	10
Batch size	32
Learning rate	0.0001
Optimizer	Adam
Loss	MSE
Patience (early stopping)	3
Num workers	10
Use AMP	FALSE
Hardware	
GPU	NVIDIA RTX A5000 (24 GB GDDR6)
CPU	Dual Intel Xeon Gold 6248R (24 cores / 48 threads, 3.00–4.00 GHz)

2.4. Evaluation Metrics

Normality of forecasting error distributions is assessed using the Shapiro–Wilk test [27], which evaluates whether a sample originates from a normally distributed population. To quantitatively assess forecasting performance, Mean Absolute Error (MAE) and Mean Squared Error (MSE) are computed on the unseen test dataset as primary accuracy metrics due to their widespread use and robustness in time series forecasting [28]. MAE and MSE are computed on the test dataset for final performance assessment, while MSE is used as the training loss function.

Mean Absolute Error (MAE) quantifies the average absolute deviation between predicted and actual values, offering robustness to outliers due to its linear treatment of errors [28]:

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i| \quad (1)$$

Mean Squared Error (MSE) measures the average squared deviation, placing greater emphasis on larger errors and penalizing significant deviations more heavily [28]:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (2)$$

where Y_i is the actual value, $\hat{Y}_i$ is the predicted value, and n is the number of observations.

To characterize the relationship between input sequence length (h) and forecasting error, following complementary metrics are used.

Kendall's Tau (τ) measures the ordinal association between input sequence length h and forecasting error (MAE) [29].

$$\tau = \frac{2}{n(n-1)} \sum_{i < j} \text{sgn}(h_j - h_i) \cdot \text{sgn}(MAE_j - MAE_i) \quad (3)$$

Values of τ near zero suggest weak monotonic dependence, which is consistent with a potential nonlinear (U-shaped) relationship. The function $\text{sgn}(\cdot)$ denotes the sign function.

Spearman's rank correlation coefficient (ρ) is used to assess monotonic associations between input sequence length and forecasting error [30]. Unlike Pearson correlation, Spearman's ρ is non-parametric and operates on ranked observations:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (4)$$

where d_i is the difference between paired ranks and n is the number of observations. Values of ρ close to -1 or $+1$ indicate strong monotonic relationships, while values near zero indicate weak or absent monotonic association.

Lind–Mehlum Test is used to assess whether the relationship between input sequence length h and forecasting error $MAE(h)$ follows a U-shaped pattern [31]. The quadratic specification is given as:

$$MAE(h) = \beta_0 + \beta_1 h + \beta_2 h^2 + \varepsilon \quad (5)$$

The slope is defined as:

$$\frac{d(MAE)}{dh} = \beta_1 + 2\beta_2 h$$

For a U-shaped relationship on the interval $[h_l, h_h]$, the slope must be negative at h_l and positive at h_h , such that:

$$\beta_1 + 2\beta_2 h_l < 0 \text{ and } \beta_1 + 2\beta_2 h_h > 0$$

The null hypothesis is that a U-shaped relationship is not supported. Rejection occurs when both one-sided t-tests are statistically significant. In addition, the overall significance of the quadratic regression is evaluated using the F-test and corresponding p-values.

The two-lines test proposed by Simonsohn is used as a robustness check against misspecification of the quadratic functional form [32]. For a candidate break point h_c , the model is specified as:

$$MAE(h) = \beta_0 + \beta_L(h - h_c) \cdot 1[h < h_c] + \beta_H(h - h_c) \cdot 1[h \geq h_c] + d \cdot 1[h \geq h_c] + \varepsilon \quad (6)$$

where $1[\cdot]$ denotes the indicator function. A U-shaped relationship is supported if $\beta_L < 0$ and $\beta_H > 0$, with both coefficients statistically significant at the $\alpha = 0.05$ level. The break point h_c is determined using a data-driven procedure described in [32]. It identifies a candidate region for the extremum of the relationship and selects a split that avoids arbitrary breakpoint selection while maintaining appropriate statistical inference.

Mann-Whitney U test is used as a non-parametric test for distributional differences to compare multivariate and univariate configurations [33]. The test statistic is defined as:

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbb{1}[y_i > y_j] \quad (7)$$

where $\mathbb{1}[\cdot]$ denotes the indicator function.

Rank-biserial correlation (RBC) is also used as a non-parametric effect size measure associated with the Mann-Whitney U test [34]. RBC quantifies the degree of separation between two distributions and ranges from -1 to 1, where larger absolute values indicate stronger effects.

$$RBC = 1 - \frac{2U}{n_1 n_2} \quad (8)$$

where U is the Mann-Whitney test statistic and n_1, n_2 are the group sample sizes.

Cohen's d test is used to quantify effect size [35].

$$d = \frac{\hat{Y}_1 - \hat{Y}_2}{s_{pooled}} \quad (9)$$

where the pooled standard deviation is:

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Finally, training time (in seconds) is recorded as an indicator of computational efficiency, capturing the practical trade-off between forecasting accuracy and computational cost across sequence lengths. Together, these metrics provide a rigorous evaluation framework for assessing model sensitivity to input sequence length and the benefits of multivariate inputs.

3. RESULTS

This section summarizes the performance outcomes of the conducted forecasting experiments under different model architectures, feature settings, and prediction horizons.

3.1. Descriptive Performance Overview

Figure 2 presents descriptive statistics for forecasting accuracy and computational cost across all experimental settings. The iTransformer consistently outperforms Transformer across both prediction horizons. In the multivariate H=2 configuration, iTransformer achieves 31% lower median MAE (0.0316 vs. 0.0459) and 38% lower median MSE (0.0034 vs. 0.0055) compared to Transformer. Crucially, iTransformer reduces training time by approximately 85% relative to Transformer (from $\approx 15,089$ s to $\approx 2,346$ seconds) in the same setting. These advantages persist at H=6, where iTransformer maintains lower prediction errors (MAE: 0.0437 vs. 0.0621) and narrower distributions. Extending the prediction horizon from 2 to 6 hours increases median MAE by 35% for Transformer and 38% for iTransformer in multivariate settings, yet iTransformer retains its accuracy advantage across both prediction horizons. Univariate configurations show similar patterns, with iTransformer achieving MAE reductions of 18% (H=2) and 17% (H=6) compared to Transformer. These results indicate that iTransformer delivers more accurate, stable, and computationally efficient forecasting performance across all evaluated conditions.

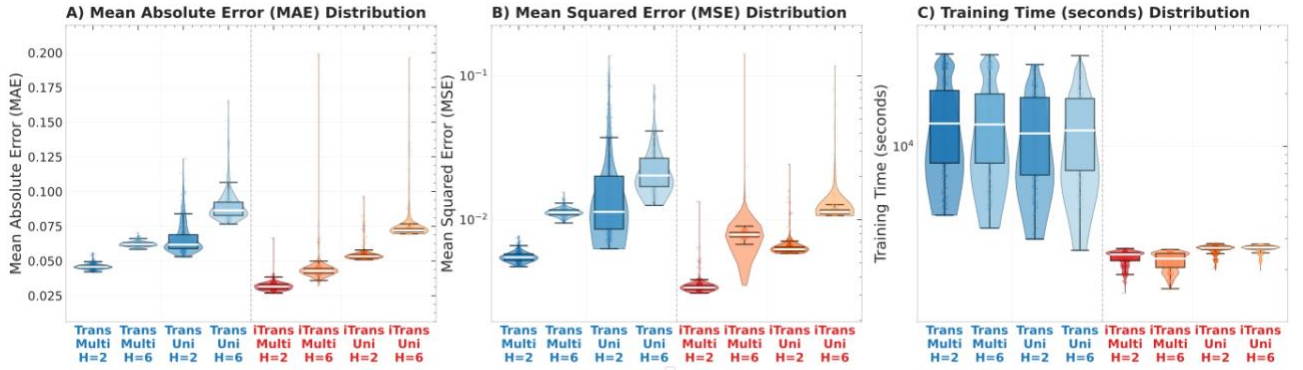


Figure 2. Distribution of Forecasting Error (MAE/MSE) and Training Time Across Models, Feature Inputs, and Prediction Horizon

3.2. Optimal Input Sequence Lengths

Table 2 reports the optimal sequence lengths (L^*) identified separately for MAE and MSE minimization across all model configurations. A striking finding emerges: in six of eight configurations, the MAE-optimal and MSE-optimal sequence lengths are exactly same. These include both univariate settings ($L^* = 174$ – 222), the multivariate H=6 condition for both models ($L^* = 192$ for Transformer, 502 for iTransformer), and importantly, the Transformer univariate configurations across both horizons. This convergence suggests that, for most practical forecasting scenarios, optimizing for MAE also yields near-optimal MSE performance, simplifying model selection.

However, two notable exceptions occur in the multivariate H=2 setting. For Transformer, the MAE-optimal sequence length ($L^* = 782$) differs markedly from its MSE-optimal length ($L^* = 88$)—a gap of 694 timesteps. Choosing the MAE-optimal sequence therefore increases MSE by only 4% (from 0.0047 to 0.0049) but increases training time nearly fivefold (from $\approx 5,517$ to $\approx 29,493$ seconds). Conversely, for iTransformer in the same setting, the MAE-optimal length ($L^* = 100$) and MSE-optimal length ($L^* = 230$) differ by 130 timesteps, but the trade-offs are reversed: the MSE-optimal sequence increases MAE by less than 1% ($0.0268 \rightarrow 0.0270$) while adding only ≈ 320 s of training time. This reflects metric-dependent sensitivity to error structure at short horizons.

Table 2. Optimal sequence length (L^*) based on forecasting performance (MAE and MSE) across models, prediction horizons, and feature inputs

Feature	h	Model	MAE- Optimal Seq Length (L^* MAE)				MSE-Optimal Seq Length (L^* MSE)				Same L^*
			L^* MAE	MAE	MSE	Time L^* MAE (s)	L^* MSE	MSE	MAE	Time L^* MSE(s)	
Multi	2	Trans	782	0.0416	0.0049	29492.9	88	0.0047	0.043	5517.3	694 X
Multi	2	iTrans	100	0.0268	0.0032	1720.85	230	0.0031	0.027	2040.6	130 X
Multi	6	Trans	192	0.0584	0.0094	7916.37	192	0.0094	0.058	7916.4	0 ✓
Multi	6	iTrans	502	0.0318	0.0035	1996.68	502	0.0035	0.032	1996.7	0 ✓
Uni	2	Trans	218	0.0531	0.0063	8435.23	218	0.0063	0.053	8435.2	0 ✓
Uni	2	iTrans	174	0.0510	0.0058	2809.20	174	0.0058	0.051	2809.2	0 ✓
Uni	6	Trans	222	0.0765	0.0125	6935.03	222	0.0125	0.077	6935	0 ✓
Uni	6	iTrans	206	0.0695	0.0107	2708.21	206	0.0107	0.07	2708.2	0 ✓

Selecting an input sequence within this range is crucial: sequences shorter than L^* lack sufficient temporal context, degrading forecast accuracy, while sequences substantially longer than L^* introduce noise and overfitting, increasing error and, for Transformer, substantially increased computational cost. The U-shaped relationship between sequence length and forecasting error, explored in the following section, provides formal statistical evidence for this behavior.

3.3. U-Shaped Relationship Between Input Sequence Length and Model Performance

Figure 3 illustrates the relationship between input sequence length and forecasting error (MAE). A U-shaped pattern is evident across most configurations: error initially decreases as longer sequences provide sufficient temporal context, reaches an optimum (L^*), then increases as excessively long sequences introduce noise and overfitting. The pattern is particularly pronounced for iTransformer models, while Transformer exhibits greater variability, especially in multivariate inputs and 6 hours prediction horizon.

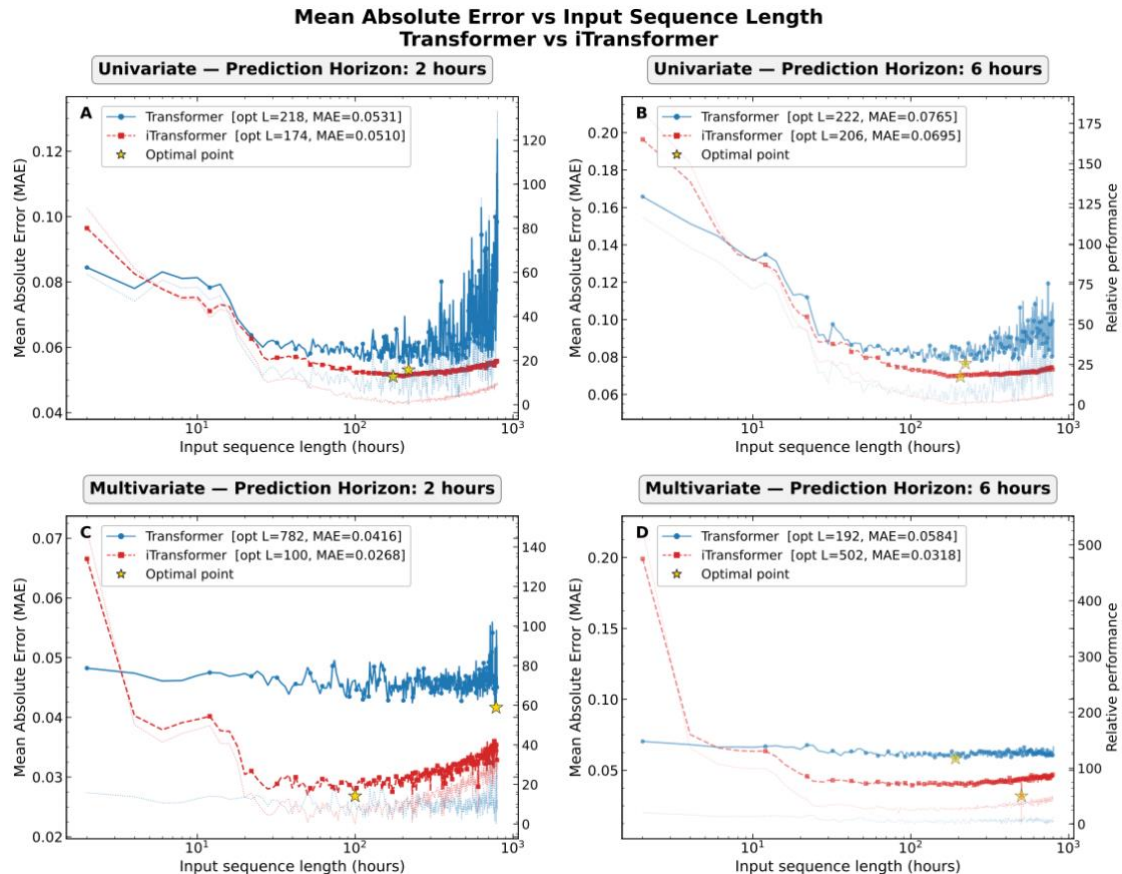


Figure 3. MAE vs. input sequence length Across Models, Input Types, and Prediction Horizons.

Table 3 and **Table 4** present comprehensive statistical tests characterizing the relationship between input sequence length and forecasting error. The U-shaped pattern, where error initially decreases, reaches an optimum, and then increases, is formally evaluated using complementary approaches: Kendall’s τ for monotonic association, quadratic regression (R^2) for curvature, the Lind–Mehlum test for formal U-shape slope conditions, the Two-Lines test for piecewise linear breakpoint identification (where P_L and P_R denote the p-values of the left and right segments, respectively, with $P_L < 0.05$ and $P_R < 0.05$ confirming a U-shaped relationship), and phase-specific Spearman correlations (ρ_{Dec} / ρ_{Inc}) to verify that forecasting error decreases before L^* (negative ρ) and increases after L^* (positive ρ). Configurations marked Δ indicate partial or asymmetric U-shaped behavior where only some statistical criteria are satisfied. Reported percentage changes should be interpreted as relative fold-changes, since very small optimal error baselines at L can inflate percentage values beyond 100%.

As **Table 3** shows, all eight configurations exhibited evidence consistent with a U-shaped relationship under at least one statistical criterion, although only four configurations satisfied both the Lind–Mehlum and Two-Lines tests simultaneously. The Lind–Mehlum test is significant for all configurations ($p < 0.05$ in every case), indicating evidence consistent with a U-shaped relationship under its endpoint slope criteria. However, the Two-Lines test, which requires both the left descending slope and right ascending slope to be individually significant, passes in only four configurations: Transformer univariate $H=6$ ($P_L = 3.27e-06$, $P_R = 1.76e-06$), iTransformer multivariate $H=6$ ($P_L = 0.004$, $P_R = 6.87e-28$), and both iTransformer univariate cases ($P_L = 5.64e-13$ to $8.29e-15$; $P_R = 2.49e-102$ to $1.70e-89$). The four configurations where Two-Lines fails all show non-significant left slopes ($P_L = 0.068$ – 0.458) despite strongly significant right slopes ($P_R = 0.00016$ – $4.95e-33$), indicating that while error clearly increases beyond L^* , the descending phase is statistically flat. The Spearman phase correlations (ρ_{Dec} / ρ_{Inc}) in Table 3 confirm this asymmetry: configurations with confirmed U-shape ($\checkmark$) transition from negative to positive correlation (e.g., iTransformer univariate $H=2$: -0.98 to $+0.98$), while weak U-shape (Δ) show flat or wrong-signed phase relationships (e.g., Transformer multivariate $H=2$: $+0.07$ to -0.45). Kendall’s τ indicates that, across the full sequence-length range, iTransformer configurations tend to exhibit stronger overall monotonic increase in error, suggesting that the post-optimal error growth dominates the global rank association. As iTransformer shows strong positive association overall ($\tau = 0.260$ – 0.663 , all $p < 1e-14$), reflecting consistent error increase beyond L^* . While Transformer shows weak association with $\tau = 0.044$ – 0.347 , and $p > 0.05$ for multivariate cases. Quadratic R^2 values range from 0.011 – 0.377 , with the weakest fit for Transformer multivariate $H=6$ ($R^2 = 0.011$, p value = 0.041), indicating that while directional slope conditions are met, the quadratic model explains little variance.

Table 3. U-shape statistical analysis based on MAE

Model	Feature	h	L^*	Min MAE	$\downarrow$ to L^* (%)	$\uparrow$ after L^* (%)	Kendall's τ	R^2	Spearman Phase (ρ_{Dec}/ρ_{Inc})	Lind Mehlum Test	Two Lines Test	U Shape
Trans	Multi	2	782	0.0416	13.7	8.2	0.045	0.071	0.0656 / -0.45	TRUE	FALSE	Δ
Trans	Multi	6	192	0.0584	17.0	13.9	0.044	0.011	-0.6115 / 0.0395	TRUE	FALSE	Δ
Trans	Uni	2	218	0.0531	37.2	51.1	0.347	0.377	-0.4415 / 0.5838	TRUE	FALSE	Δ
Trans	Uni	6	222	0.0765	53.8	5.0	0.320	0.165	-0.6779 / 0.548	TRUE	TRUE	$\checkmark$
iTrans	Multi	2	100	0.0268	59.7	27.3	0.663	0.298	-0.6001 / 0.9064	TRUE	FALSE	Δ
iTrans	Multi	6	502	0.0318	84.0	45.7	0.625	0.068	0.4884 / 0.6829	TRUE	TRUE	$\checkmark$
iTrans	Uni	2	174	0.0510	47.1	10.3	0.464	0.284	-0.9754 / 0.9785	TRUE	TRUE	$\checkmark$
iTrans	Uni	6	206	0.0695	64.6	6.4	0.260	0.313	-0.989 / 0.9528	TRUE	TRUE	$\checkmark$

As **Table 4** shows, two Transformer configurations fail to demonstrate a U-shaped pattern entirely when evaluated on MSE. Transformer exhibits partial evidence of U-shaped behavior in two MSE configurations according to the Lind–Mehlum test, although none satisfy the stricter Two-Lines criterion (multivariate $H=6$, univariate $H=2$). The Lind–Mehlum test failing for

multivariate $H=2$ ($p = 0.442$) and univariate $H=6$ ($p = 0.537$), and critically, the Two-Lines test fails in all four Transformer cases ($\times$). This indicates that Transformer's MSE behavior is largely non-U-shaped or flat. In contrast, iTransformer maintains statistically significant U-shaped patterns across all four configurations: the Lind–Mehlum test returns TRUE ($p < 0.001$ for three configurations, $p = 0.0003$ for multivariate $H=6$), and the Two-Lines test confirms significance in all cases (P_R : $6.91e-59$ to $2.04e-76$; P_L : 0.0003 to $3.57e-10$). Despite iTransformer's lower Quad_ R^2 values (0.039 – 0.256) compared to Transformer (0.086 – 0.267), the formal slope conditions (Lind–Mehlum) and segment-specific significance (Two-Lines) provide stronger evidence for a genuine U-shaped relationship. The phase-specific Spearman correlations in Table 4 reinforce the U-shaped evidence: for iTransformer, all configurations show strong negative-to-positive transitions (e.g., univariate $H=6$: $-0.989 \rightarrow +0.941$, both $p < 1e-84$), whereas Transformer's phase correlations are inconsistent or non-significant, consistent with its weaker U-shaped support. Kendall's τ supports this: iTransformer shows moderate-to-strong positive association ($\tau = 0.215$ – 0.604 , all $p < 1e-10$), reflecting consistent error increase beyond L^* , while Transformer shows weaker association ($\tau = 0.194$ – 0.298).

Table 4. U-shape statistical analysis based on MSE

Model	Feature	h	L^*	Min MSE	↓ to L^* (%)	↑ after L^* (%)	Kendall's τ	R^2	Spearman Phase (ρ_{Dec}/ρ_{Inc})	Lind Mehlum Test	Two Lines Test	U Shape
Trans	Multi	2	88	0.005	22.4	23.8	0.243	0.151	-0.7088 / 0.2881	FALSE	FALSE	$\times$
Trans	Multi	6	192	0.009	39.6	37.3	0.194	0.086	-0.7152 / 0.3463	TRUE	FALSE	Δ
Trans	Uni	2	218	0.006	59	5.59 $\times$	0.298	0.267	-0.1382 / 0.4498	TRUE	FALSE	Δ
Trans	Uni	6	222	0.013	78.1	8.67	0.224	0.112	-0.3207 / 0.3272	FALSE	FALSE	$\times$
iTrans	Multi	2	230	0.003	76.9	15.9	0.604	0.092	-0.7301 / 0.9299	TRUE	TRUE	$\checkmark$
iTrans	Multi	6	502	0.004	97.5	1.44 $\times$	0.507	0.039	0.0638 / 0.7576	TRUE	TRUE	$\checkmark$
iTrans	Uni	2	174	0.006	76.1	16.8	0.348	0.256	-0.9830 / 0.9645	TRUE	TRUE	$\checkmark$
iTrans	Uni	6	206	0.011	90.9	11.5	0.215	0.223	-0.9893 / 0.9409	TRUE	TRUE	$\checkmark$

For iTransformer multivariate $H=6$, selecting the optimal sequence length ($L^* = 502$) reduces MAE by 84.0% relative to the shortest sequence, while error rises by 45.7% beyond L^* . Similarly, for MSE, the same configuration shows a 97.5% drop to optimum and a 144% rise after, the largest effect observed. Drop percentages for iTransformer range from 47.1–84% (MAE) and 76.9–97.5% (MSE), substantially larger than Transformer's 14–54% (MAE) and 22–78% (MSE). Kendall's τ values further support this: iTransformer shows strong positive monotonic association overall ($\tau = 0.260$ – 0.663 , all $p < 1e-14$), reflecting consistent error increase beyond L^* , while Transformer shows weak or non-significant association ($\tau = 0.044$ – 0.347 , p often > 0.05).

Table 5. Optimal sequence length ranges and statistical confidence across model configurations

Configuration	Optimal Range (L^*)	Statistical Confidence
Trans – Multi	88–782 (divergent)	Weak or inconsistent U-shape evidence
Trans – Uni	218–222 (tight)	Moderate but mixed U-shape evidence
iTrans– Multi	100–502	Strong U-shape evidence
iTrans–Uni	174–206	Strong U-shape evidence

Integrating the MAE and MSE analyses as demonstrated in **Table 5**, optimal sequence lengths vary substantially depending on model type, feature configuration, and forecasting horizon, ranging from approximately 88 to 782 hourly timesteps for this dataset and short 2- and 6-hour prediction horizons. This wide range reflects differences in architecture and feature sensitivity rather than a single universal optimum. However, a clear clustering emerges: iTransformer configurations consistently achieve optimal performance at shorter and more stable lookback windows (approximately 100–502 timesteps depending on horizon and feature set), while

Transformer models exhibit more dispersed and less stable optima, particularly under multivariate settings.

3.4. Computational Cost: Derived Power-Law Relationships

Training time exhibits clear architecture-dependent scaling behavior with respect to input sequence length (**Figure 4**).

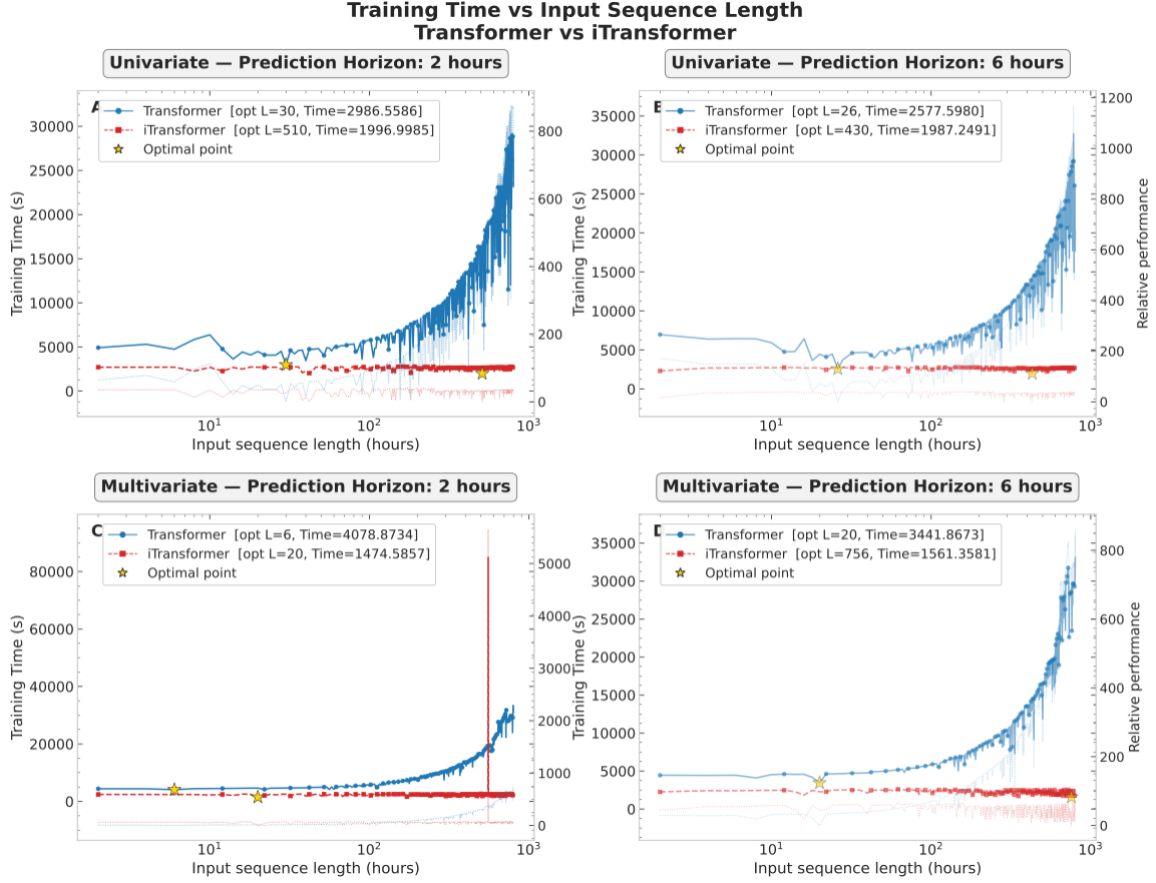


Figure 4. Training Time Across Models, Input Types, and Prediction Horizons

To move beyond the theoretical $O(L^2)$ complexity of self-attention and provide empirically grounded formulas, a power-law model $T(L) = a \cdot L^b$. Each configuration by ordinary least-squares regression on the log-log scale is fitted. The exponent b characterises the scaling regime, where $O(L^b)$ denotes the computational complexity class: $b = 2$ corresponds to the theoretical quadratic self-attention complexity, $b = 1$ to linear scaling, $b = 0.5$ to square-root growth, and $b \approx 0$ to constant time $O(1)$.

For the Transformer, training time increases monotonically with sequence length and is well described by a square-root power law. Across all four configurations the pooled exponent is $b \approx 0.52$ (range 0.488–0.549), with $R^2 = 0.71$ –0.77 on the original scale, confirming empirical consistency with square-root-like scaling. The per-configuration formulas are:

Multivariate, H=2:	$T(L) = 563 \cdot L^{0.549}$	(95%, CI: [0.522, 0.575]; $R^2 = 0.774$)
Multivariate, H=6:	$T(L) = 582 \cdot L^{0.542}$	(95%, CI: [0.515, 0.569]; $R^2 = 0.758$)
Univariate, H=2:	$T(L) = 657 \cdot L^{0.501}$	(95%, CI: [0.470, 0.531]; $R^2 = 0.712$)
Univariate, H=6:	$T(L) = 721 \cdot L^{0.488}$	(95%, CI: [0.459, 0.518]; $R^2 = 0.722$)

In all cases the 95% confidence interval for b lies entirely below 2, indicating that the empirically observed runtime scaling in this experimental setup is substantially weaker than the theoretical quadratic self-attention complexity. The observed empirical runtime scaling is approximately proportional to $L^{0.52}$ within the evaluated sequence-length range.

The iTransformer presents a fundamentally different scaling profile and is approximately sequence-length invariant within the evaluated range, as quantified by the very low fitted exponents:

$$\begin{aligned}
\text{Multivariate, } H=2: & \quad T(L) = 2205 \cdot L^{0.010} \quad (95\%, CI: [0.000, 0.020]; R^2 = 0.008) \\
\text{Multivariate, } H=6: & \quad T(L) = 2788 \cdot L^{-0.039} \quad (95\%, CI: [-0.050, -0.027]; R^2 = 0.094) \\
\text{Univariate, } H=2: & \quad T(L) = 2532 \cdot L^{0.006} \quad (95\%, CI: [-0.000, 0.011]; R^2 = 0.006) \\
\text{Univariate, } H=6: & \quad T(L) = 2731 \cdot L^{-0.006} \quad (95\%, CI: [-0.011, -0.001]; R^2 = 0.015)
\end{aligned}$$

In every configuration the confidence interval for b contains or straddles zero, and the near-zero R^2 values (0.006–0.094) confirm that sequence length explains negligible variance in iTransformer training time. This behaviour likely arises from the inverted attention mechanism, in which pairwise interactions are computed across feature channels rather than across time steps, resulting in weak empirical dependence on temporal horizon within the evaluated range.

The exponent ratio b (Transformer ≈ 0.52 versus iTransformer ≈ 0) has direct practical consequences. For a tenfold increase in sequence length ($L = 100 \rightarrow 1,000$), Transformer training time grows by a factor of $10^{0.52} \approx 3.3\times$, whereas iTransformer time remains essentially unchanged. At the MAE-optimal sequence lengths (**Table 2**), the Transformer required up to 29,493 s (multivariate, $H=2$), while the iTransformer required only 1,721 s, a 17-fold reduction, for a configuration where iTransformer also achieved lower absolute error (MAE = 0.0268 vs 0.0416). These results establish that L^* represents a Pareto-efficient operating point on both the accuracy and cost dimensions, and that iTransformer dominates the Transformer across the vast majority of tested configurations.

3.5. Extreme Weather Adaptivity (Impact of Input Dimensionality)

Beyond computational efficiency and sequence-length scaling, an important objective is evaluating model performance under extreme weather conditions, where abnormal temperature fluctuations can produce atypical electricity demand patterns. To assess whether richer contextual information improves robustness, univariate and multivariate input configurations were compared for both Transformer and iTransformer architectures.

Table 6 summarizes the descriptive statistics across all configurations. Incorporating additional meteorological and periodicity-related variables consistently improved forecasting accuracy for both architectures and horizons. For Transformer, multivariate inputs reduced MAE by approximately 30.0% ($h=2$) and 30.4% ($h=6$), while MSE decreased by 70.7% and 52.3%, respectively. The improvement was even larger for iTransformer, where multivariate inputs reduced MAE by 41.4% ($h=2$) and 41.2% ($h=6$), with corresponding MSE reductions of 46.8% and 34.0%. These reductions are reflected in substantially lower mean errors for multivariate configurations. For example, Transformer MAE decreased from 0.06558 to 0.04592 at $h=2$, while iTransformer MAE decreased from 0.05397 to 0.03161.

Table 6. Descriptive Statistics of MAE and MSE Across Models and Input Types

Feature	h	Model	Metric	Mean	Median	Std	Min	Max	IQR
Multi	2	Trans	MAE	0.04592	0.04574	0.00176	0.04164	0.05593	0.00182
Multi	2	Trans	MSE	0.00551	0.00547	0.00044	0.00471	0.00768	0.00055
Multi	2	iTrans	MAE	0.03161	0.03148	0.00289	0.02682	0.06655	0.00349
Multi	2	iTrans	MSE	0.00343	0.00337	0.00057	0.00308	0.01335	0.00023
Multi	6	Trans	MAE	0.06211	0.06196	0.00165	0.05840	0.07036	0.00206
Multi	6	Trans	MSE	0.01129	0.01122	0.00076	0.00941	0.01558	0.00089
Multi	6	iTrans	MAE	0.04365	0.04280	0.00856	0.03184	0.19908	0.00348
Multi	6	iTrans	MSE	0.00834	0.00786	0.00679	0.00350	0.14165	0.00056
Uni	2	Trans	MAE	0.06558	0.06170	0.01050	0.05306	0.12367	0.01003
Uni	2	Trans	MSE	0.01884	0.01132	0.01856	0.00627	0.13822	0.01145
Uni	2	iTrans	MAE	0.05397	0.05329	0.00400	0.05101	0.09649	0.00226
Uni	2	iTrans	MSE	0.00645	0.00621	0.00133	0.00582	0.02431	0.00042
Uni	6	Trans	MAE	0.08919	0.08639	0.01058	0.07653	0.16581	0.00953
Uni	6	Trans	MSE	0.02369	0.02024	0.01058	0.01254	0.08674	0.00973

Uni	6	iTrans	MAE	0.07421	0.07188	0.01145	0.06951	0.19642	0.00232
Uni	6	iTrans	MSE	0.01264	0.01128	0.00766	0.01066	0.11750	0.00069

presents the statistical comparisons, which confirm that these improvements are highly significant across all experiments. All MWU tests produced extremely small p-values ($p < 10^{-12}$), strongly rejecting the null hypothesis of equal distributions between univariate and multivariate configurations. Effect sizes were consistently large, particularly for iTransformer MAE ($d = -6.41$, $RBC = 0.995$ at $h=2$) and Transformer MAE ($d = -3.58$, $RBC = 1.0$ at $h=6$). Even the smallest observed effect, iTransformer MSE at $h=6$, remained statistically significant ($p = 1.26 \times 10^{-123}$) with a moderate-to-large practical effect ($d = -0.594$, $RBC = 0.966$). Collectively, these findings demonstrate that integrating richer temporal and weather-related features substantially enhances forecasting robustness and improves model adaptability under extreme weather conditions, with iTransformer benefiting most from multivariate contextual information.

Table 7 presents the statistical comparisons, which confirm that these improvements are highly significant across all experiments. All MWU tests produced extremely small p-values ($p < 10^{-12}$), strongly rejecting the null hypothesis of equal distributions between univariate and multivariate configurations. Effect sizes were consistently large, particularly for iTransformer MAE ($d = -6.41$, $RBC = 0.995$ at $h=2$) and Transformer MAE ($d = -3.58$, $RBC = 1.0$ at $h=6$). Even the smallest observed effect, iTransformer MSE at $h=6$, remained statistically significant ($p = 1.26 \times 10^{-123}$) with a moderate-to-large practical effect ($d = -0.594$, $RBC = 0.966$). Collectively, these findings demonstrate that integrating richer temporal and weather-related features substantially enhances forecasting robustness and improves model adaptability under extreme weather conditions, with iTransformer benefiting most from multivariate contextual information.

Table 7. Statistical Comparison Across Models and Input Types

Model	h	Metric	MWU p	Cohen's d	RBC	% Diff	Multi Mean	Uni Mean	Spearman (Multi)	Spearman (Uni)
Trans	2	MAE	$<10^{-131}$	-2.612	0.999	-29.98	0.0459	0.0656	0.063 ($p=0.207$)	0.508 ($p < 10^{-27}$)
Trans	2	MSE	$<10^{-131}$	-1.015	0.998	-70.74	0.0055	0.0188	0.350 ($p < 10^{-12}$)	0.449 ($p < 10^{-20}$)
Trans	6	MAE	$<10^{-131}$	-3.576	1.000	-30.36	0.0621	0.0892	0.059 ($p=0.239$)	0.448 ($p < 10^{-20}$)
Trans	6	MSE	$<10^{-130}$	-1.653	0.997	-52.32	0.0113	0.0237	0.279 ($p < 10^{-7}$)	0.335 ($p < 10^{-11}$)
iTrans	2	MAE	$<10^{-130}$	-6.407	0.995	-41.43	0.0316	0.0540	0.784 ($p < 10^{-83}$)	0.463 ($p < 10^{-21}$)
iTrans	2	MSE	$<10^{-129}$	-2.956	0.992	-46.80	0.0034	0.0065	0.722 ($p < 10^{-64}$)	0.318 ($p < 10^{-10}$)
iTrans	6	MAE	$<10^{-129}$	-3.023	0.991	-41.18	0.0437	0.0742	0.738 ($p < 10^{-69}$)	0.183 ($p < 10^{-3}$)
iTrans	6	MSE	$<10^{-122}$	-0.594	0.966	-34.00	0.0083	0.0126	0.627 ($p < 10^{-44}$)	0.129 ($p=0.0099$)

Spearman correlation analysis is used to evaluate how forecasting error changes with input sequence length within each configuration separately, rather than comparing multivariate and univariate models directly. Transformer multivariate MAE configurations exhibit weak and statistically non-significant correlations with sequence length ($\rho=0.063$, $p=0.207$ for $h=2$; $\rho=0.059$, $p=0.239$ for $h=6$), indicating relatively stable performance across varying temporal contexts. In contrast, Transformer univariate configurations show stronger positive correlations ($\rho=0.508$, $p < 10^{-27}$ for $h=2$), suggesting that forecasting errors become more sensitive to sequence length when fewer features are available. For iTransformer, both multivariate MAE and MSE configurations exhibit strong positive and highly significant sequence-length dependencies, particularly for MAE ($\rho=0.784$, $p < 10^{-83}$ at $h=2$) and MSE ($\rho=0.722$, $p < 10^{-64}$). Compared with the corresponding univariate configurations ($\rho=0.463$ and $\rho=0.318$, respectively), this indicates that sequence length plays a more influential role when richer contextual information (features) is included. The Spearman analysis suggests that multivariate inputs improve accuracy, but may introduce stronger interaction effects with sequence length in some architectures.

Overall, these findings demonstrate that incorporating meteorological and periodicity-related variables substantially improves forecasting robustness and enhances adaptability under extreme weather conditions, with iTransformer benefiting most from multivariate contextual information.

4. CONCLUSIONS AND FUTURE DIRECTIONS

This study systematically investigated the role of input sequence length in Transformer-based short-term electricity load forecasting using 400 evaluated sequence lengths spanning 2–800 hourly timesteps, across univariate and weather-informed multivariate settings, and for both Transformer and iTransformer architectures. The results demonstrate that input sequence length is a critical and highly nonlinear hyperparameter that significantly affects forecasting accuracy, computational cost, and model robustness.

Across all configurations, a consistent U-shaped relationship between sequence length and forecasting error (MAE and MSE) was observed, supported by multiple statistical tests including quadratic regression, Lind–Mehlum slope conditions, and two-lines validation. Forecasting accuracy improves as sequence length increases from very short histories due to richer temporal context, but deteriorates beyond an optimal range as irrelevant or noisy historical information dominates model learning. This confirms that neither short nor excessively long input windows are optimal, but instead that performance is governed by a clear bias–variance trade-off specific to architecture and input structure.

Optimal sequence lengths were found to vary substantially depending on model type, feature configuration, and forecasting horizon, ranging from approximately 88 to 782 timesteps. However, a clear clustering emerges: iTransformer configurations consistently achieve optimal performance at shorter and more stable lookback windows (approximately 100–502 timesteps depending on horizon and feature set), while Transformer models exhibit more dispersed and less stable optima, particularly under multivariate settings. Importantly, deviations from these optimal ranges lead to substantial performance degradation, while correct tuning yields improvements between 13.7–84% in MAE and 22.4–97.5% in MSE.

In terms of predictive performance, iTransformer consistently and robustly outperforms the standard Transformer across nearly all configurations, achieving 31–38% lower forecasting errors and substantially tighter error distributions. This superiority is particularly pronounced in multivariate settings, where iTransformer benefits more effectively from weather and seasonal covariates. Statistical testing confirms that multivariate inputs significantly improve forecasting accuracy for both architectures ($p < 10^{-12}$ across all comparisons), with large effect sizes, highlighting the importance of incorporating exogenous meteorological and cyclical variables for robust load prediction under volatile conditions.

From a computational perspective, Transformer training time exhibits approximately square-root-like scaling with sequence length ($T(L) \propto L^{0.52}$), indicating a consistent sub-linear empirical growth pattern within the evaluated range. Performance degradation beyond the optimal sequence length can be substantial, with error increasing by up to approximately 5–6× depending on configuration. This results in substantial efficiency gains, with iTransformer reducing training time by up to an order of magnitude while simultaneously improving accuracy. These findings establish a clear Pareto advantage for iTransformer in both predictive performance and computational efficiency.

Overall, the study demonstrates that optimal forecasting performance in Transformer-based load prediction cannot be achieved through fixed heuristic sequence lengths. Instead, sequence length must be treated as a dataset- and architecture-dependent optimization variable. For short-term electricity load forecasting (2- and 6-hour horizons), the most reliable performance is achieved through properly tuned intermediate sequence lengths combined with multivariate meteorological inputs, particularly when using iTransformer architectures.

Future research should extend this analysis to longer forecasting horizons, additional Transformer variants, and cross-country electricity systems to validate generalizability. Further investigation into adaptive or learnable sequence length mechanisms may eliminate the need for manual tuning altogether. In addition, integrating probabilistic forecasting frameworks and extreme-weather regime modeling could further enhance robustness under rare but critical load conditions.

5. REFERENCES

1. Wen, Q., et al., *Transformers in time series: A survey*. arXiv preprint arXiv:2202.07125, 2022.
2. Wang, Y., et al., *Short-term load forecasting for industrial customers based on TCN-LightGBM*. IEEE transactions on power systems, 2020. **36**(3): p. 1984-1997.
3. Černe, G., D. Dovžan, and I. Škrjanc, *Short-term load forecasting by separating daily profiles and using a single fuzzy model across the entire domain*. IEEE Transactions on Industrial Electronics, 2018. **65**(9): p. 7406-7415.
4. Bunn, D. and E. Farmer, *Economic and operational context of electric load prediction. Comparative models for electrical load forecasting*, 1985: p. 3-11.
5. Valipour, M., M.E. Banihabib, and S.M.R. Behbahani, *Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir*. Journal of hydrology, 2013. **476**: p. 433-441.
6. Qin, L., et al., *The impact of lengths of time series on the accuracy of the ARIMA forecasting*. International Research in Higher Education, 2019. **4**(3): p. 58-68.
7. Vaswani, A., et al., *Attention is all you need*. Advances in neural information processing systems, 2017. **30**.
8. Zhou, H., et al. *Informer: Beyond efficient transformer for long sequence time-series forecasting*. in *Proceedings of the AAAI conference on artificial intelligence*. 2021.
9. Wu, H., et al., *Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting*. Advances in neural information processing systems, 2021. **34**: p. 22419-22430.
10. Liu, Y., et al., *itransformer: Inverted transformers are effective for time series forecasting*. arXiv preprint arXiv:2310.06625, 2023.
11. Tarmanini, C., et al., *Short term load forecasting based on ARIMA and ANN approaches*. Energy Reports, 2023. **9**: p. 550-557.
12. Amarasinghe, K., D.L. Marino, and M. Manic. *Deep neural networks for energy load forecasting*. in *2017 IEEE 26th international symposium on industrial electronics (ISIE)*. 2017. IEEE.
13. Aliyari, M. and Y.Z. Ayele, *Application of artificial neural networks for power load prediction in critical infrastructure: A comparative case study*. Applied System Innovation, 2023. **6**(6): p. 115.
14. Aliyari, M., et al., *Ant colony optimization of hyper-parameters in multi-head attention layer for time series forecasting*. Applied Soft Computing, 2025: p. 113562.
15. Zeng, A., et al. *Are transformers effective for time series forecasting?* in *Proceedings of the AAAI conference on artificial intelligence*. 2023.
16. Carvalho, S.G., R. Guerra-Sá, and L.H. de C Merschmann, *The impact of sequence length and number of sequences on promoter prediction performance*. BMC bioinformatics, 2015. **16**(Suppl 19): p. S5.
17. Zhao, Z., et al., *Short-term load forecasting based on the transformer model*. information, 2021. **12**(12): p. 516.
18. Debnath, S., et al. *Extreme Weather Grid Load Forecasting using Weather-Informed LSTM and Transformer Machine Learning Models*. in *2025 57th North American Power Symposium (NAPS)*. 2025. IEEE.
19. Unlu, A. and M. Peña, *Comparative analysis of hybrid deep learning models for electricity load forecasting during extreme weather*. Energies, 2025. **18**(12): p. 3068.
20. Guo, F., J.-Z. Wu, and L. Pan. *An empirical study of AI model's performance for electricity load forecasting with extreme weather conditions*. in *International Conference on Science of Cyber Security*. 2023. Springer.
21. Deng, X., et al., *Bagging-XGBoost algorithm based extreme weather identification and short-term load forecasting model*. Energy Reports, 2022. **8**: p. 8661-8674.
22. Sun, M., C. Feng, and J. Zhang. *Factoring behind-the-meter solar into load forecasting: Case studies under extreme weather*. in *2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. 2020. IEEE.
23. Dehalwar, V., et al. *Electricity load forecasting for Urban area using weather forecast information*. in *2016 IEEE International Conference on Power and Renewable Energy (ICPRE)*. 2016. IEEE.

24. Wang, J., et al., *Power system source-load forecasting based on scene generation in extreme weather*. Energy, 2025. **330**: p. 136991.
25. Xu, L., S. Wang, and R. Tang, *Probabilistic load forecasting for buildings considering weather forecasting uncertainty and uncertain peak load*. Applied energy, 2019. **237**: p. 180-195.
26. Sun, M., et al., *Load Characteristic Analysis and Load Forecasting Method Considering Extreme Weather Conditions*. Electronics, 2025. **14**(20): p. 3978.
27. Shapiro, S.S. and M.B. Wilk, *An analysis of variance test for normality (complete samples)*. Biometrika, 1965. **52**(3-4): p. 591-611.
28. Hyndman, R.J. and A.B. Koehler, *Another look at measures of forecast accuracy*. International journal of forecasting, 2006. **22**(4): p. 679-688.
29. Sen, P.K., *Estimates of the regression coefficient based on Kendall's tau*. Journal of the American statistical association, 1968. **63**(324): p. 1379-1389.
30. Spearman, C., *The proof and measurement of association between two things*. 1961.
31. Lind, J.T. and H. Mehlum, *With or without U? The appropriate test for a U-shaped relationship*. Oxford bulletin of economics and statistics, 2010. **72**(1): p. 109-118.
32. Simonsohn, U., *Two lines: A valid alternative to the invalid testing of U-shaped relationships with quadratic regressions*. Advances in Methods and Practices in Psychological Science, 2018. **1**(4): p. 538-555.
33. Mann, H.B. and D.R. Whitney, *On a test of whether one of two random variables is stochastically larger than the other*. The annals of mathematical statistics, 1947: p. 50-60.
34. Cureton, E.E., *Rank-biserial correlation*. Psychometrika, 1956. **21**(3): p. 287-290.
35. Cohen, J., *Statistical power analysis for the behavioral sciences*. 2013: routledge.