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

Ingénierie des Systèmes d'Information
Vol. 31, No. 7, July, 2026, pp. 2241-2254
Journal homepage: http://iieta.org/journals/isi

A Hybrid Evaluation Metric for Time Series Forecasting: Integrating Magnitude Accuracy and Temporal Dynamics Across Multiple Domains

Anass Zatri[1*], Mohammed Hatim Rziki[2], Khalid Zine-Dine[3]

[1] National School of Computer Science and Systems Analysis (ENSIAS), Mohammed V University in Rabat, Rabat 10112, Morocco
[2] Faculty of Sciences, Moulay Ismail University, Meknes 11201, Morocco
[3] Faculty of Sciences, Mohammed V University in Rabat, Rabat 10106, Morocco

Corresponding Author Email: anass_zatri@um5.ac.ma

https://doi.org/10.18280/isi.310710

Received: 16 March 2026
Revised: 3 July 2026
Accepted: 21 July 2026
Available online: 31 July 2026

Keywords:
time series forecasting, forecast evaluation, hybrid evaluation metric, multi-objective assessment, temporal dynamics, model ranking, cross-domain evaluation

ABSTRACT

Reliable evaluation of time series forecasting models remains challenging because conventional error metrics primarily quantify numerical deviations while overlooking the preservation of temporal dynamics. Metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Scaled Error (MASE) may therefore produce incomplete or conflicting assessments when forecasting models achieve different levels of magnitude accuracy and trajectory consistency. This study introduces a Hybrid Evaluation Metric (HEM), a unified and interpretable framework that integrates normalized magnitude error and temporal trajectory disagreement through a controllable trade-off parameter. Unlike single-objective metrics, HEM enables evaluation preferences to be explicitly adjusted according to operational requirements. The proposed metric is evaluated across three heterogeneous forecasting scenarios: household electricity consumption, financial stock prices, and meteorological temperature. Five representative forecasting models, including statistical, machine learning, and deep learning approaches, are compared using rolling-window evaluation, sensitivity analysis, statistical validation, and controlled perturbation experiments. The results demonstrate that model rankings vary with both the dataset characteristics and the relative importance assigned to numerical accuracy or temporal fidelity. No single forecasting model consistently dominates across all domains, highlighting the importance of objective-aware evaluation. HEM provides a transparent and flexible approach for comparing forecasting models when forecast quality cannot be adequately characterized by pointwise errors alone. The proposed framework offers a practical tool for multi-domain forecasting evaluation and supports more informed model selection in real-world applications.

1. INTRODUCTION

Time series forecasting plays a central role in modern decision-making systems, with applications in energy demand management, financial risk analysis, supply chain planning, transportation, weather prediction, and healthcare resource allocation. In these domains, forecasting models are often compared using standard error metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), symmetric MAPE (sMAPE), or Mean Absolute Scaled Error (MASE) [1-3]. These metrics are useful because they quantify the numerical distance between observed and predicted values. However, they mainly evaluate pointwise deviations and do not fully describe whether a forecast preserves the temporal dynamics of the observed series.

This limitation is important in many real forecasting problems. A model may achieve a low RMSE while failing to reproduce the shape of the observed signal. Conversely, another model may follow the trend and peaks of the series but produce biased values. In energy demand forecasting, for example, the shape of the consumption profile can be operationally important because it reflects demand dynamics, peak periods, and load variations. In financial forecasting, short-term trajectories are noisy and difficult to predict, so evaluating both magnitude error and directional behavior becomes important. In weather forecasting, models must reproduce both the level and the temporal evolution of variables such as temperature.

In this work, we propose a hybrid error metric (HEM) for time series forecast evaluation. HEM combines normalized RMSE, which measures magnitude accuracy, with a correlation-based disagreement term, which measures temporal trajectory fidelity. The two components are combined through a trade-off parameter α. Low values of α give more importance to trajectory preservation, while high values give more importance to numerical accuracy. Therefore, HEM does not impose a single universal notion of forecasting quality. Instead, it makes the evaluation preference explicit and adaptable to the operational objective.

The motivation for using a hybrid metric is that forecasting performance is often multi-dimensional. Classical metrics may lead to conflicting conclusions when one model is better in magnitude accuracy, and another model is better in trajectory alignment. HEM provides a unified and interpretable framework for analyzing this trade-off. The metric is especially useful when the preferred model depends on whether the user prioritizes accurate values, temporal shape preservation, or a compromise between both.

We evaluate HEM across three forecasting domains with different temporal characteristics. The first dataset is the UCI Household Power Consumption dataset, used for short-term energy demand forecasting. The second dataset contains Netflix stock closing prices, used as a financial forecasting case study. The third dataset is the Jena Climate dataset, used for weather temperature forecasting. These datasets allow us to analyze HEM under different conditions: daily energy consumption cycles, noisy financial dynamics, and smoother meteorological seasonality.

We compare five forecasting models: Seasonal Autoregressive Integrated Moving Average (SARIMA) [4], Prophet [5], XGBoost [6], PatchTST [7], and TimesNet [8]. SARIMA and Prophet represent classical statistical and additive forecasting approaches. XGBoost represents a machine learning model based on lagged, rolling, and calendar features. PatchTST and TimesNet are included as recent deep learning architectures for time series forecasting. This model set allows HEM to be evaluated across traditional, machine learning, and modern deep learning paradigms.

The experimental protocol uses rolling forecasting windows to obtain repeated evaluations over time. We compute traditional error metrics, Pearson correlation, and HEM over the full grid $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$. We also perform bootstrap confidence interval estimation, Wilcoxon signed-rank tests, ranking stability analysis, and controlled perturbation experiments. These analyses are designed to evaluate not only average performance but also statistical reliability, α sensitivity, and the behavior of HEM under known forecast distortions such as bias, scaling, smoothing, temporal shifts, outliers, noise, and trajectory inversion.

The main contribution of this work is an interpretable evaluation framework that links magnitude accuracy and temporal trajectory fidelity within a single metric. We also provide a systematic interpretation of α, showing how it can be used as a decision parameter rather than as an arbitrary weight. By applying HEM across energy, finance, and weather forecasting, we show that the preferred model depends on both the dataset and the evaluation objective. This makes HEM useful as a decision-oriented metric for comparing forecasting models in contexts where forecast quality cannot be reduced to pointwise error alone.

2. MATERIALS AND METHODS

2.1 Forecast evaluation metrics

Forecast evaluation relies on metrics that quantify different aspects of the discrepancy between observed and predicted values. Let $y = (y_1, y_2, ..., y_n)$ denote the observed time series and $\hat{y} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$ the corresponding forecast.

2.1.1 Root Mean Squared Error

The RMSE measures the quadratic average deviation between observed and predicted values:

$$RMSE(y, \hat{y}) = \sqrt{\{((1/n)\Sigma_i{}^{=1n}(y_i - \hat{y}_i)^2\}} \quad (1)$$

RMSE is widely used because it is expressed in the same unit as the target variable and strongly penalizes large errors. However, this sensitivity to large deviations also makes RMSE vulnerable to outliers and does not indicate whether the forecast preserves the temporal shape of the series [1, 2].

2.1.2 Mean Absolute Error

The MAE computes the average absolute deviation:

$$MAE(y, \hat{y}) = (1/n) \times \Sigma_i{}^{=1n}|y_i - \hat{y}_i| \quad (2)$$

MAE is easier to interpret than RMSE because it treats all errors linearly. It is less sensitive to extreme values, but like RMSE, it remains a pointwise error metric and does not capture temporal alignment or trajectory similarity [2].

2.1.3 Mean Absolute Percentage Error

The MAPE expresses the error as a percentage of the observed value:

$$MAPE(y, \hat{y}) = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (3)$$

MAPE is useful when relative errors are required, but it becomes unstable when observed values are close to zero. This limitation is particularly important in domains where the target variable may contain small or near-zero values [2, 3].

2.1.4 Symmetric Mean Absolute Percentage Error

The sMAPE reduces some of the asymmetry of MAPE by normalizing the absolute error using both observed and predicted values:

$$sMAPE(y, \hat{y}) = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \quad (4)$$

sMAPE is often used in forecasting competitions and comparative studies because it provides a bounded relative error formulation. Nevertheless, it can still be difficult to interpret when both observed and predicted values are small [3, 9].

2.1.5 Mean Absolute Scaled Error

The MASE compares the model error with the error of a naïve seasonal benchmark:

$$MASE = \frac{MAE}{\frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|} \quad (5)$$

where, m is the seasonal period, and T is the length of the in-sample training series.

MASE is scale-independent and allows comparisons across series with different units. A MASE value below one indicates that the model outperforms the seasonal naïve benchmark. However, MASE remains primarily an error-magnitude metric and does not directly evaluate whether the temporal trajectory is preserved [3, 10].

2.1.6 Pearson correlation coefficient

The Pearson correlation coefficient measures the linear association between the observed and predicted series:

$$\rho(y,\hat{y}) = Cov(y,\hat{y})/\left(\sigma_y \times \sigma_{\hat{y}}\right) \quad (6)$$

Unlike error-based metrics, correlation focuses on co-movement and temporal shape similarity. A high correlation indicates that the forecast follows the direction and dynamics of the observed series. However, correlation alone does not penalize amplitude bias: a forecast may have high correlation while systematically overestimating or underestimating the true values.

Overall, these metrics provide complementary but partial views of forecasting performance. RMSE, MAE, MAPE, sMAPE, and MASE evaluate magnitude errors from different perspectives, while correlation evaluates trajectory similarity. In practice, these metrics may lead to conflicting conclusions. A model may achieve low error but fail to reproduce the temporal shape, while another model may preserve the trajectory but produce biased magnitudes. This motivates the use of a hybrid evaluation metric that explicitly combines magnitude accuracy and temporal trajectory fidelity.

2.2 Composite, multi-objective, and shape-aware evaluation

Forecast evaluation is not limited to pointwise error metrics [11, 12]. Several studies have shown that model rankings can change substantially depending on the metric used for evaluation, especially in large forecasting benchmarks and competitions [3, 9]. This has motivated the use of composite metrics, multi-objective ranking, and shape-aware similarity measures.

Composite metrics combine several performance indicators into a single score. Their advantage is that they can summarize multiple aspects of model quality. However, they often lack a clear interpretation of how each component contributes to the final ranking. Multi-objective ranking avoids reducing performance to one criterion, but it may be difficult to use when a practical decision requires selecting a single model [13].

Shape-aware metrics, such as correlation-based measures and dynamic time warping (DTW), focus on the similarity between temporal patterns [11, 14-17]. These approaches are useful when the evolution of the signal is important. Nevertheless, they may not sufficiently penalize magnitude errors. DTW, for example, can tolerate temporal distortions by elastically aligning two sequences, whereas in many forecasting applications the timing of peaks remains operationally important.

HEM is positioned between classical error metrics and shape-aware evaluation. It does not aim to replace all existing metrics. Instead, it provides a simple and interpretable score that explicitly combines normalized magnitude error and synchronous trajectory fidelity. Its main advantage is that the trade-off between these two objectives is controlled by α, making the evaluation preference transparent.

2.3 Hybrid and composite metrics

Recent research has explored composite and multi-objective approaches to forecast evaluation. Pinson and Tastu [12] emphasized that probabilistic forecast evaluation should consider more than average error magnitude, including aspects such as calibration and sharpness. Makridakis et al. [4, 10] also showed through large-scale forecasting competitions that model rankings can change substantially depending on the selected metric. Similar observations have been reported in load forecasting and renewable energy forecasting, where different metrics may favor different models depending on whether the evaluation prioritizes magnitude accuracy, peak prediction, or temporal dynamics [18].

Composite metrics aggregate several evaluation criteria into a single score. Their advantage is that they provide a compact summary of model performance. However, the interpretation of the final score can become unclear when the components measure different aspects of forecast quality. Multi-objective ranking approaches preserve several evaluation criteria separately, but they may not provide a direct decision rule when a single model must be selected.

Shape-aware metrics, such as correlation-based measures and DTW, focus on the similarity between temporal patterns. These approaches are useful when the evolution of the signal is important. However, they may either ignore magnitude bias or tolerate temporal distortions. For example, DTW can align shifted sequences, whereas in many forecasting applications, especially short-term energy forecasting, the timing of peaks remains operationally important.

HEM is positioned between classical error metrics, composite metrics, and shape-aware evaluation. It does not aim to replace all existing metrics. Instead, it provides a simple and interpretable score that explicitly combines normalized magnitude error and synchronous trajectory fidelity. Its main contribution is to make the trade-off between these two objectives transparent through the parameter α. Table 1 summarizes the main characteristics of HEM compared with commonly used forecast evaluation approaches, highlighting its differences in terms of evaluation focus, interpretation, and decision-making capability.

Table 1. Comparison between hybrid error metric (HEM) and existing evaluation approaches

Approach	Main Focus	Strength	Limitation	Difference from HEM
RMSE / MAE	Magnitude error	Simple and interpretable	Ignore trajectory fidelity	HEM adds temporal trajectory evaluation
Pearson correlation	Shape similarity	Captures co-movement	Ignores magnitude bias	HEM adds normalized error penalty
DTW	Shape alignment	Handles shifted patterns	May tolerate timing errors	HEM keeps synchronous evaluation
Multi-objective ranking	Multiple criteria	Flexible	Harder to obtain one decision score	HEM provides one α-controlled score
Composite metrics	Aggregated performance	Compact summary	Component interpretation may be unclear	HEM uses two interpretable components
HEM	Magnitude + trajectory	Explicit trade-off through α	Requires α interpretation	Designed for decision-oriented forecast evaluation

Note: RMSE = Root Mean Squared Error, MAE = Mean Absolute Error, DTW = dynamic time warping, HEM = hybrid error metric.

3. HYBRID ERROR METRIC

3.1 Definition of the hybrid error metric

The objective of the HEM is to evaluate forecasting performance by considering both magnitude accuracy and temporal trajectory fidelity. Let $y = (y_1, y_2, ..., y_n)$ be the observed time series and $\hat{y} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$ be the corresponding predicted series.

1. The proposed HEM is defined as:

$$HEM_\alpha(y,\hat{y}) = \alpha \times \frac{RMSE(y,\hat{y})}{\sigma_y} + (1-\alpha) \times \big(1 - \rho(y,\hat{y})\big) \tag{7}$$

where, $\alpha \in [0,1]$ controls the relative importance assigned to magnitude accuracy and trajectory fidelity, σ_y is the standard deviation of the observed series, and $\rho(y, \hat{y})$ is the Pearson correlation coefficient between the observed and predicted series.

The term $RMSE(y, \hat{y}) / \sigma_y$ corresponds to the normalized RMSE, denoted as Normalized Root Mean Square Error (NRMSE). Therefore, HEM can also be written as:

$$HEM_\alpha(y,\hat{y}) = \alpha \times NRMSE(y,\hat{y}) + (1-\alpha) \times \big(1 - \rho(y,\hat{y})\big) \tag{8}$$

The first component, $NRMSE(y, \hat{y})$, measures normalized magnitude error. The second component, $1-\rho(y, \hat{y})$, measures trajectory disagreement. A lower HEM value indicates better forecasting performance.

3.2 Magnitude accuracy component

The magnitude component of HEM is based on the RMSE normalized by the standard deviation of the observed series. The normalized error component is defined as:

$$NRMSE(y,\hat{y}) = \frac{RMSE(y,\hat{y})}{\sigma_y} \tag{9}$$

where RMSE is defined as:

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

and σ_y is the standard deviation of the observed series:

$$\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{11}$$

where,

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{12}$$

The normalization by σ_y makes the magnitude component dimensionless. This is important because HEM is evaluated across datasets with different units and scales, such as electricity consumption, stock prices, and temperature. Using raw RMSE alone would make the magnitude component dependent on the unit of the target variable, while NRMSE allows a more comparable interpretation across domains.

3.3 Temporal trajectory fidelity component

The trajectory component of HEM is based on the Pearson correlation coefficient between the observed and predicted series. It is defined as:

$$\rho(y,\hat{y}) = \frac{\sum_{i=1}^{n}(y_i - \overline{y})\big(\hat{y}_i - \overline{\hat{y}}\big)}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}\big(\hat{y}_i - \overline{\hat{y}}\big)^2}} \tag{13}$$

where,

$$\overline{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i \tag{14}$$

The Pearson correlation coefficient satisfies:

$$-1 \le \rho(y,\hat{y}) \le 1 \tag{15}$$

Therefore, the trajectory disagreement term satisfies:

$$1 - \rho(y,\hat{y}) \ge 0 \tag{16}$$

When $\rho(y, \hat{y}) = 1$, the predicted series perfectly follows the temporal evolution of the observed series, and the trajectory disagreement term becomes zero. When $\rho(y, \hat{y})$ is low or negative, the forecast does not correctly reproduce the observed temporal dynamics, and the disagreement term increases.

3.4 Interpretation of hybrid error metric

HEM can be interpreted as a weighted combination of two complementary forecasting objectives: normalized magnitude accuracy and temporal trajectory fidelity. The first term penalizes numerical deviations between observed and predicted values, while the second term penalizes forecasts that fail to reproduce the temporal evolution of the observed series.

The parameter α controls the trade-off between these two objectives. When α is high, HEM gives more importance to magnitude accuracy. When α is low, HEM gives more importance to temporal trajectory fidelity.

When $\alpha = 1$, HEM becomes:

$$HEM_1(y,\hat{y}) = \frac{RMSE(y,\hat{y})}{\sigma_y} = NRMSE(y,\hat{y}) \tag{17}$$

In this case, HEM behaves as a purely magnitude-oriented metric.

When $\alpha = 0$, HEM becomes:

$$HEM_0(y,\hat{y}) = 1 - \rho(y,\hat{y}) \tag{18}$$

In this case, HEM behaves as a purely trajectory-oriented metric. For intermediate values of α, HEM provides a

compromise between magnitude accuracy and trajectory fidelity. Low values of α give more importance to temporal shape preservation, while high values of α give more importance to numerical accuracy.

3.5 Theoretical properties

HEM satisfies several useful properties for forecast evaluation.

Non-negativity: Since RMSE is non-negative and $\sigma_y > 0$ for a non-constant observed series, the normalized error component is non-negative:

$$RMSE(y, \hat{y})/\sigma_y \geq 0 \tag{19}$$

Since $-1 \leq \rho(y, \hat{y}) \leq 1$, the trajectory disagreement term is also non-negative:

$$1 - \rho(y, \hat{y}) \geq 0 \tag{20}$$

Therefore:

$$HEM_\alpha(y, \hat{y}) \geq 0, \forall \alpha \in [0,1] \tag{21}$$

Lower-bound behavior: For a perfect forecast:

$$y_i = \hat{y}_i, \forall i = 1, \dots, n \tag{22}$$

we have:

$$RMSE(y, \hat{y}) = 0, NRMSE(y, \hat{y}) = 0, \rho(y, \hat{y}) = 1 \tag{23}$$

Therefore:

$$HEM_\alpha(y, \hat{y}) = 0 \tag{24}$$

This confirms that the minimum value of HEM is obtained when the forecast is both numerically exact and perfectly aligned with the observed trajectory.

Scale normalization: The use of RMSE(y, $\hat{y}$) / σ_y ensures that the magnitude component is scale-independent. This makes HEM suitable for cross-domain evaluation, where the target variables may have different units and ranges.

Sensitivity to outliers: HEM inherits part of the sensitivity of RMSE to large errors through the normalized magnitude component. As α increases, large numerical deviations have a stronger influence on the final HEM score. Conversely, when α is low, the metric becomes less dominated by isolated amplitude errors and gives more importance to trajectory preservation.

Sensitivity to shape distortion. The correlation disagreement component makes HEM sensitive to temporal shape distortion. Forecasts that are shifted, smoothed, inverted, or unable to reproduce peaks and troughs receive a higher trajectory penalty. This property is useful when the temporal dynamics of the forecast are important for decision-making.

3.6 Selection strategy for α

The parameter α should not be interpreted as an arbitrary constant. It represents the evaluation preference between magnitude accuracy and trajectory fidelity.

In this study, HEM was computed over the discrete grid $\alpha \in \{0.0, 0.1, 0.2, ..., 1.0\}$. For clarity and compactness, the main tables report representative α values: $\alpha = 0.3$, 0.5, 0.7, and 1.0. These values correspond to shape-oriented, balanced, magnitude-oriented, and pure magnitude-error configurations, respectively.

The value $\alpha = 0.5$ is used as a balanced reference because it assigns equal importance to normalized magnitude error and trajectory disagreement. However, the full α-sensitivity analysis is reported to examine whether model rankings remain stable or depend on the evaluation preference.

For profile-oriented forecasting tasks, such as reproducing the shape of electricity consumption, low α values may be appropriate because trajectory fidelity is prioritized. For quantity-oriented tasks, such as numerical demand estimation, procurement, or capacity planning, higher α values may be more appropriate because magnitude accuracy becomes more important.

A data-driven strategy can also be used by analyzing ranking stability across α values. Let $\text{Rank}_m(\alpha)$ denote the rank of model m under HEM for a given value of α. The average rank of model m over the α grid is defined as:

$$\bar{R_{\text{m}}} = (1/|A|) \times \Sigma\alpha \in A Rank_{\text{m}}(\alpha) \tag{25}$$

A model with a consistently low average rank across α values can be considered robust to the choice of evaluation preference. If the best model changes across α values, this indicates that the preferred model depends on the operational objective.

When an explicit downstream cost or utility function is available, α can also be calibrated by selecting the value that maximizes the agreement between HEM-based rankings and application-specific rankings:

$$\alpha^* = \arg\max_{\alpha \in [0,1]} A\,g\,reement \left(Rank_{HEM}(\alpha), Rank_{Utility}\right) \tag{26}$$

In the absence of such an external utility function, reporting $\alpha = 0.5$ together with the full α-sensitivity analysis provides a transparent and reproducible selection strategy.

4. EXPERIMENTAL DESIGN

We evaluate the behavior of HEM across three forecasting domains with different temporal characteristics: household electricity consumption, financial stock prices, and meteorological temperature. The purpose of this multi-domain design is to assess whether HEM provides meaningful interpretations beyond a single benchmark dataset.

The three datasets differ in their temporal structure. Household energy consumption contains strong daily cycles and operationally important load profiles. Netflix stock prices represent a noisy financial series with weak short-term predictability. Jena Climate temperature data provide smoother meteorological dynamics with daily and seasonal patterns. These differences allow HEM to be evaluated under distinct forecasting conditions.

4.1 Dataset description

The first dataset is the UCI Household Power Consumption dataset [4]. The target variable is global active power, resampled at hourly frequency. The forecasting horizon is set

to 24 hours, corresponding to short-term day-ahead energy demand forecasting.

The second dataset consists of Netflix stock closing prices. The target variable is the daily closing price, evaluated at business-day frequency [5]. The forecasting horizon is set to five business days. This dataset is used to represent noisy financial forecasting, where short-term trajectory alignment is difficult.

The third dataset is the Jena Climate dataset. The target variable is air temperature, T (degC), resampled at hourly frequency [6]. The forecasting horizon is set to 24 hours. This dataset is used to evaluate HEM in a smoother weather forecasting setting with daily and seasonal structure.

4.2 Data preprocessing

We preprocessed each dataset according to its temporal resolution and forecasting objective. For the household energy dataset, the original measurements were aggregated to hourly frequency using the mean value, and missing values were handled by time-based interpolation. The target variable was global active power, and the forecasting horizon was set to 24 hours.

For the Netflix stock dataset, the date column was converted to a business-day time index. The closing price was used as the target variable. Missing business days were forward-filled in order to preserve a regular time index required by the forecasting models. The forecasting horizon was set to five business days.

For the Jena Climate dataset, the original high-frequency observations were resampled to hourly frequency using the mean value. Temperature, T (degC), was used as the target variable, and missing values were interpolated. The forecasting horizon was set to 24 hours.

For all datasets, the same general evaluation logic was used. Each series was sorted chronologically, transformed into a regular time index, and evaluated using rolling forecasting windows. No information from the test horizon was used during training, ensuring a temporally consistent evaluation protocol.

4.3 Forecasting models

Five forecasting models are evaluated: SARIMA, Prophet, XGBoost, PatchTST, and TimesNet.

SARIMA is used as a classical statistical model capable of capturing autoregressive and seasonal dependencies [6]. Prophet is used as an additive model that captures trend and seasonality [19]. XGBoost is used as a machine learning model based on lagged values, rolling statistics, and calendar-derived features [20]. PatchTST and TimesNet are included as recent deep learning architectures for time series forecasting. PatchTST relies on a patch-based Transformer design, while TimesNet models temporal variations through multi-period representations [7, 21].

This set of models allows HEM to be assessed across statistical, additive, machine learning, and deep learning forecasting paradigms. Recent studies have shown that Transformer-based architectures such as Informer, Autoformer, PatchTST, and TimesNet have become important benchmarks for time series forecasting. Recent foundation models and surveys also illustrate the growing interest in general-purpose time-series models and cross-domain forecasting [18, 22-28].

4.4 Rolling-window evaluation protocol

All experiments use a rolling-window evaluation protocol. For each rolling window, the model is trained on historical observations and evaluated on the following forecast horizon. This avoids relying on a single train-test split and provides repeated paired observations for statistical testing.

For the household energy dataset, the training window is 180 days, the forecast horizon is 24 hours, and the rolling step is 7 days. For the Netflix dataset, the training window is approximately two years, the forecast horizon is five business days, and the rolling step is one month. For the Jena Climate dataset, the training window is one year, the forecast horizon is 24 hours, and the rolling step is two weeks.

4.5 Evaluation metrics and hybrid error metric configuration

Each model is evaluated using RMSE, NRMSE, MAE, MAPE, sMAPE, MASE, Pearson correlation, and HEM [1-3]. HEM is computed for $\alpha \in \{0.0, 0.1, ..., 1.0\}$. The value $\alpha = 0.5$ is reported as the balanced reference configuration, while the full α-sensitivity analysis is used to examine whether model rankings are stable or dependent on the evaluation preference.

Low α values emphasize trajectory fidelity, while high α values emphasize magnitude accuracy. This interpretation is important because the appropriate evaluation preference may vary by domain. For example, a profile-oriented energy forecasting task may favor lower α values, whereas numerical demand estimation may require higher α values.

4.6 Statistical validation

Two statistical validation procedures are used. First, bootstrap confidence intervals are computed for the mean HEM score of each model. These intervals quantify uncertainty across rolling windows. Second, Wilcoxon signed-rank tests are used for paired model comparisons. For each pair of models, the test compares HEM values window by window under $\alpha = 0.5$. This determines whether observed performance differences are statistically significant. The Wilcoxon signed-rank test is appropriate because it does not require assuming normality of the paired differences.

4.7 Controlled perturbation experiments

Controlled perturbation experiments are used to analyze the behavior of HEM under known types of forecast degradation. Artificial forecasts are generated from the observed test series using amplitude scaling, additive bias, noise, outlier contamination, smoothing, temporal shifts, and trajectory inversion. These experiments test whether HEM behaves consistently in interpretable cases. A perfect forecast should obtain HEM = 0. A forecast with preserved shape but biased amplitude should be penalized through the NRMSE component. A temporally shifted forecast should be penalized because it affects both magnitude error and trajectory agreement. An inverted trajectory should receive a strong penalty due to negative correlation.

5. RESULTS AND DISCUSSION

This section presents the empirical results obtained on the

three forecasting domains: household electricity consumption, Netflix stock prices, and Jena Climate temperature. The results are analyzed using both classical forecasting metrics and the proposed HEM. Particular attention is given to the role of α, since this parameter determines whether the evaluation favors magnitude accuracy or temporal trajectory fidelity.

The results show that HEM does not systematically favor a single forecasting model across all datasets. Instead, the preferred model depends on the domain and on the evaluation preference represented by α. This behavior is important because it confirms the role of HEM as a decision-oriented metric rather than a fixed ranking rule.

Figure 1 compares the mean HEM values under the balanced configuration $\alpha = 0.5$. Lower values indicate a better compromise between normalized magnitude error and temporal trajectory fidelity. The figure shows that the best-performing model changes across datasets: XGBoost performs best on the energy dataset, PatchTST performs best on the Netflix dataset, and SARIMA performs best on the Jena Climate dataset.

5.1 Household energy forecasting results

The household energy dataset provides a particularly relevant case for analyzing the behavior of HEM because energy demand forecasting can be interpreted from two complementary perspectives. In some applications, the objective is to reproduce the global consumption profile and follow the temporal dynamics of demand. In other applications, the objective is to estimate the numerical level of electricity consumption as accurately as possible. HEM allows these two objectives to be explicitly controlled through the parameter α.

Under the balanced configuration $\alpha = 0.5$, XGBoost obtains the lowest mean HEM, with a score of 0.5929. SARIMA ranks second with a mean HEM of 0.6141, followed by Prophet, PatchTST, and TimesNet. XGBoost also achieves the lowest RMSE and NRMSE, which indicates better magnitude accuracy. However, SARIMA obtains the highest Pearson correlation, which indicates better preservation of the temporal consumption profile.

This result illustrates the main purpose of HEM. A purely magnitude-based evaluation would favor XGBoost, whereas a purely trajectory-based evaluation would favor SARIMA. HEM makes this trade-off explicit. Under $\alpha = 0.5$, XGBoost is preferred because its lower normalized magnitude error compensates for its slightly lower correlation. However, when α is reduced and trajectory fidelity becomes more important, SARIMA becomes the preferred model.

This behavior is confirmed by the α-sensitivity analysis. For $\alpha = 0.0$ and $\alpha = 0.1$, SARIMA obtains the best HEM score, because the evaluation gives dominant importance to trajectory fidelity. From $\alpha = 0.2$ to $\alpha = 1.0$, XGBoost becomes the best model, because the normalized magnitude error receives increasing weight. Therefore, the energy dataset clearly shows that the preferred model depends on the operational objective: SARIMA is more appropriate for profile-oriented forecasting, while XGBoost is more appropriate for quantity-oriented forecasting.

The bootstrap confidence intervals of XGBoost and SARIMA overlap, and the Wilcoxon signed-rank test indicates that their difference is not statistically significant at the 5% level. Therefore, these two models should be interpreted as the strongest group on the energy dataset. The final preference between them depends on whether the application prioritizes temporal trajectory fidelity or numerical demand accuracy.

Table 2 shows that XGBoost achieves the best HEM score under the balanced configuration $\alpha = 0.5$. This is mainly due to its lower RMSE and NRMSE. However, SARIMA achieves the highest correlation, which means that it better follows the temporal evolution of the electricity consumption profile. This contrast motivates a deeper analysis of α.

Table 3 reports the HEM values obtained for all models across the α grid on the household energy dataset. The results show that SARIMA obtains the best HEM score when $\alpha = 0.0$ and $\alpha = 0.1$, meaning that it is preferred when the evaluation strongly prioritizes trajectory fidelity. From $\alpha = 0.2$ to $\alpha = 1.0$, XGBoost becomes the best model, showing that it is preferred as soon as normalized magnitude accuracy receives moderate or high importance. This transition illustrates the practical role of α as an interpretable decision parameter.

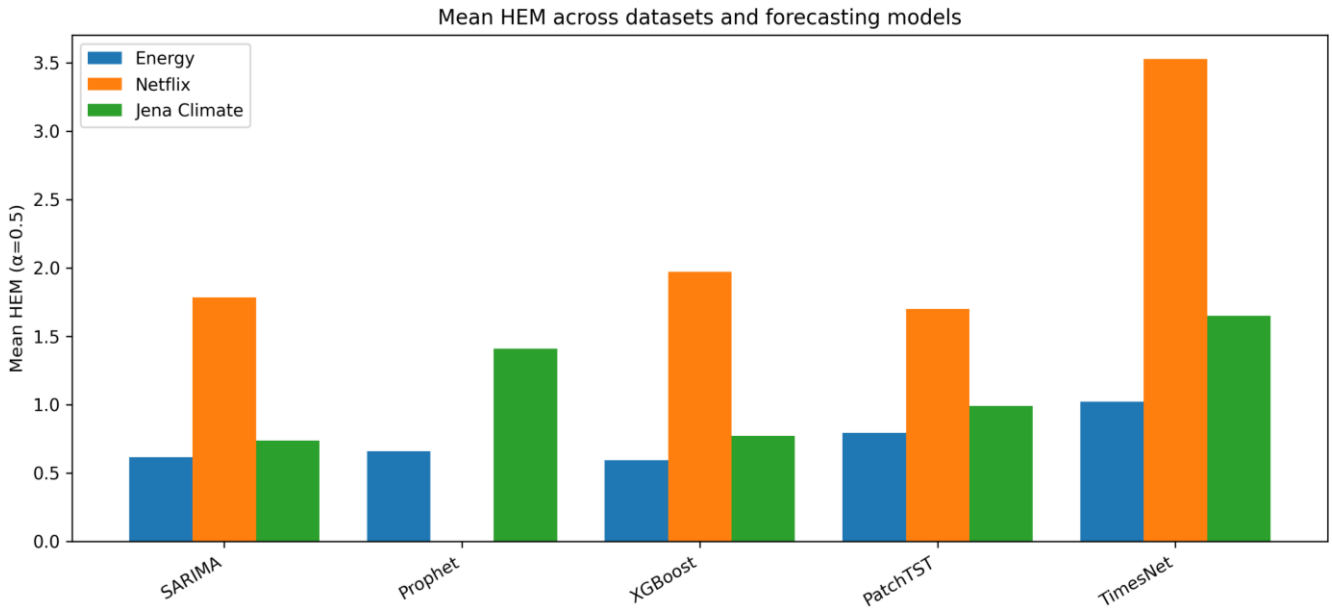

Figure 1. Mean hybrid error metric (HEM) at $\alpha = 0.5$ across datasets and forecasting models

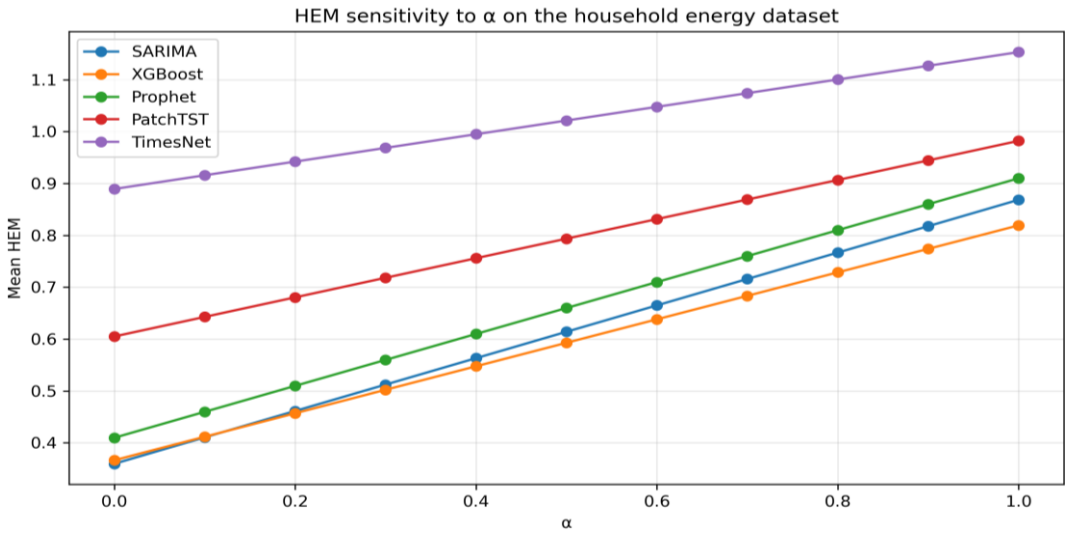

Figure 2. Hybrid error metric (HEM) sensitivity to α on the household energy dataset

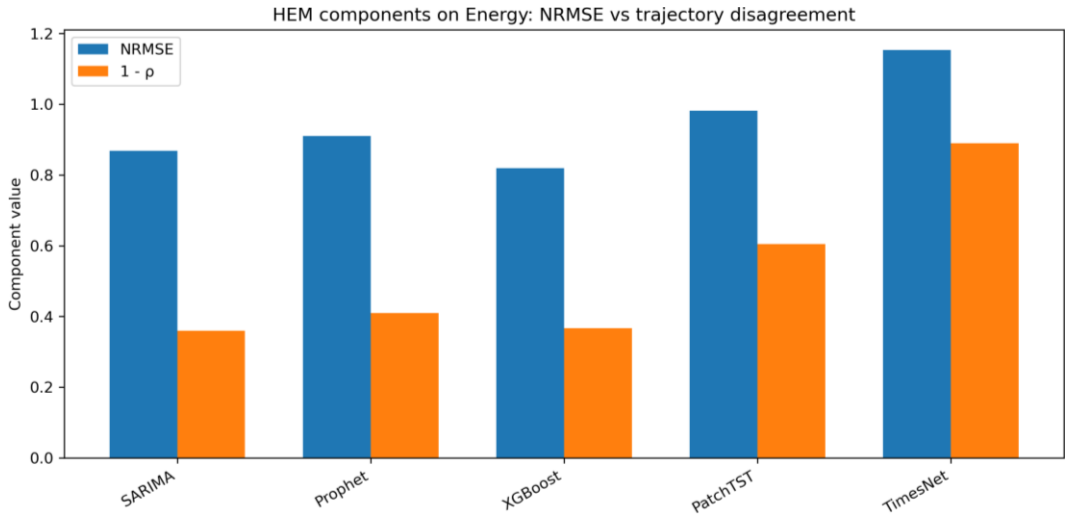

Figure 3. Decomposition of hybrid error metric (HEM) into Normalized Root Mean Square Error (NRMSE) and trajectory disagreement on the household energy dataset

Table 2. Average forecasting performance on the household energy dataset

Rank	Model	RMSE	NRMSE	MAE	ρ	HEM $\alpha = 0.5$
1	XGBoost	0.7135	0.8192	0.5332	0.6335	0.5929
2	SARIMA	0.7719	0.8685	0.5466	0.6403	0.6141
3	Prophet	0.7933	0.9100	0.6081	0.5902	0.6599
4	PatchTST	0.8866	0.9822	0.6612	0.3949	0.7936
5	TimesNet	1.0326	1.1532	0.7535	0.1106	1.0213

Note: RMSE = Root Mean Squared Error, MAE = Mean Absolute Error, HEM = hybrid error metric, NRMSE = Normalized Root Mean Square Error, SARIMA = Seasonal Autoregressive Integrated Moving Average.

Table 3. Hybrid error metric (HEM) sensitivity to α on the household energy dataset

α	Sarima	Prophet	XGBoost	PatchTST	TimesNet	Best Model
0.0	0.3597	0.4098	0.3665	0.6051	0.8894	SARIMA
0.1	0.4106	0.4598	0.4118	0.6428	0.9158	SARIMA
0.2	0.4615	0.5099	0.4570	0.6805	0.9422	XGBoost
0.3	0.5123	0.5599	0.5023	0.7182	0.9686	XGBoost
0.4	0.5632	0.6099	0.5476	0.7559	0.9950	XGBoost
0.5	0.6141	0.6599	0.5929	0.7936	1.0213	XGBoost
0.6	0.6650	0.7099	0.6381	0.8313	1.0477	XGBoost
0.7	0.7159	0.7600	0.6834	0.8690	1.0741	XGBoost
0.8	0.7668	0.8100	0.7287	0.9067	1.1005	XGBoost
0.9	0.8177	0.8600	0.7739	0.9445	1.1269	XGBoost
1.0	0.8685	0.9100	0.8192	0.9822	1.1532	XGBoost

Note: SARIMA = Seasonal Autoregressive Integrated Moving Average.

Figure 2 shows the sensitivity of HEM to α on the household energy dataset. SARIMA obtains the lowest HEM for $\alpha = 0.0$ and $\alpha = 0.1$, which corresponds to trajectory-oriented evaluation. XGBoost becomes the best model from $\alpha = 0.2$ to $\alpha = 1.0$, showing that it is preferred as soon as magnitude accuracy receives moderate or high importance. This figure illustrates how HEM makes the model selection process dependent on the forecasting objective.

Figure 3 decomposes HEM into its two components: NRMSE and trajectory disagreement, represented by $1 - \rho$. XGBoost has the lowest NRMSE, which explains why it is favored when α increases. SARIMA has the lowest trajectory disagreement because it achieves the highest correlation, which explains why it is favored when α is close to zero. This decomposition confirms that HEM provides an interpretable compromise between magnitude accuracy and temporal trajectory fidelity.

5.2 Netflix stock forecasting results

The Netflix stock forecasting experiment represents a noisy financial setting. Under $\alpha = 0.5$, PatchTST obtains the lowest mean HEM, followed by SARIMA and XGBoost. Prophet and TimesNet obtain higher HEM values.

The correlation values are generally weak across the models, which reflects the difficulty of short-term financial forecasting. In this context, HEM shows that the ranking is mainly driven by the ability to control normalized magnitude error rather than by strong trajectory alignment.

The Wilcoxon signed-rank test indicates that the difference between PatchTST and SARIMA is not statistically significant. Therefore, their performance should be interpreted as comparable under the balanced HEM configuration. PatchTST is significantly better than Prophet and TimesNet, while its difference with XGBoost is not statistically significant.

The α-sensitivity analysis shows that TimesNet ranks first only when $\alpha = 0.0$, corresponding to a purely correlation-oriented evaluation. Once magnitude accuracy is included, PatchTST becomes the most stable model across most α values. This suggests that PatchTST provides the best compromise for the Netflix forecasting task.

Table 4 shows that the Netflix dataset is more difficult to forecast than the energy and weather datasets. This is reflected by weak or negative correlation values for several models. PatchTST obtains the best HEM score under $\alpha = 0.5$ because it provides the best compromise between normalized error and trajectory disagreement in this noisy financial setting.

5.3 Jena Climate forecasting results

The Jena Climate dataset provides a smoother weather forecasting setting. Under $\alpha = 0.5$, SARIMA obtains the lowest mean HEM, followed closely by XGBoost. PatchTST ranks third, while Prophet and TimesNet obtain higher HEM values.

The comparison between SARIMA and XGBoost again illustrates the usefulness of HEM. XGBoost achieves the highest average correlation, meaning that it follows the temperature trajectory slightly better. However, SARIMA achieves lower RMSE and NRMSE. Under the balanced HEM setting, SARIMA is preferred because its magnitude advantage compensates for its slightly lower correlation.

The Wilcoxon signed-rank test shows that the difference between SARIMA and XGBoost is not statistically significant. Therefore, both models should be considered competitive on the Jena Climate dataset. SARIMA significantly outperforms Prophet and TimesNet under $\alpha = 0.5$.

The α-sensitivity analysis shows that XGBoost ranks first for low α values, where trajectory fidelity dominates, while SARIMA becomes first when α increases. This confirms that HEM provides an interpretable way to understand why the preferred model changes when the evaluation objective changes.

Table 5 shows that SARIMA and XGBoost are the two most competitive models on the weather dataset. XGBoost follows the trajectory slightly better, as indicated by its higher correlation, but SARIMA has lower magnitude error. Under $\alpha = 0.5$, HEM ranks SARIMA first because the reduction in normalized error compensates for the slightly lower correlation.

5.4 Cross-domain comparison

The three datasets show that HEM does not systematically favor a single model family.

Table 4. Average forecasting performance on the Netflix stock dataset

Rank	Model	RMSE	NRMSE	MAE	ρ	HEM $\alpha = 0.5$
1	PatchTST	1.8967	2.4674	1.7334	0.0679	1.6997
2	SARIMA	1.9023	2.4905	1.7416	-0.0769	1.7837
3	XGBoost	2.1404	2.9094	1.9574	-0.0310	1.9702
4	Prophet	2.7304	4.7428	2.5853	-0.0280	2.8854
5	TimesNet	3.2302	6.1511	3.0887	0.1001	3.5255

Note: RMSE = Root Mean Squared Error, MAE = Mean Absolute Error, HEM = hybrid error metric, NRMSE = Normalized Root Mean Square Error, SARIMA = Seasonal Autoregressive Integrated Moving Average.

Table 5. Average forecasting performance on the Jena Climate dataset

Rank	Model	RMSE	NRMSE	MAE	ρ	HEM $\alpha = 0.5$
1	SARIMA	2.3913	1.2582	2.0573	0.7825	0.7379
2	XGBoost	2.5271	1.3550	2.1617	0.8121	0.7714
3	PatchTST	2.7051	1.5362	2.2813	0.5536	0.9913
4	Prophet	3.5533	2.5967	3.2240	0.7818	1.4075
5	TimesNet	4.0584	2.4461	3.4610	0.1486	1.6488

Note: RMSE = Root Mean Squared Error, MAE = Mean Absolute Error, HEM = hybrid error metric, NRMSE = Normalized Root Mean Square Error, SARIMA = Seasonal Autoregressive Integrated Moving Average.

Table 6. Best model by dataset under the balanced hybrid error metric (HEM) Etting $\alpha = 0.5$

Dataset	Best Model	HEM $\alpha = 0.5$	Main Interpretation
Energy	XGBoost	0.5929	Best balanced score due to lower normalized error
Netflix	PatchTST	1.6997	Best compromise in a noisy financial setting
Jena Climate	SARIMA	0.7379	Best balanced score due to lower magnitude error

Note: HEM = hybrid error metric, SARIMA = Seasonal Autoregressive Integrated Moving Average.

Table 6 shows that in the energy dataset, XGBoost and SARIMA form the strongest group, with the preferred model depending on whether magnitude accuracy or trajectory fidelity is prioritized. In the Netflix dataset, PatchTST provides the best average compromise under most α values. In the Jena Climate dataset, SARIMA and XGBoost are again the strongest models, but their ranking changes with α.

This cross-domain behavior is important because it shows that HEM is not merely a score that always selects the same type of model. Instead, HEM reveals how model preference depends on the forecasting objective. The metric is therefore useful for interpreting model performance rather than only ranking models mechanically.

Across the three domains, the strongest models are not always the most recent deep learning architectures. PatchTST performs well on the Netflix dataset, while SARIMA and XGBoost remain highly competitive in energy and weather forecasting. This indicates that model selection should remain data-dependent and that evaluation metrics should be able to compare different model families transparently.

5.5 α-sensitivity and ranking stability

The α-sensitivity analysis is central to the interpretation of HEM. It shows how the ranking of forecasting models changes when the evaluation preference moves from trajectory fidelity to magnitude accuracy.

HEM was computed over the discrete grid $\alpha \in \{0.0, 0.1, 0.2, ..., 1.0\}$. For compactness, the main tables emphasize representative values $\alpha = 0.3$, $\alpha = 0.5$, $\alpha = 0.7$, and $\alpha = 1.0$, corresponding respectively to shape-oriented, balanced, magnitude-oriented, and pure normalized-error configurations. The complete grid is used to analyze ranking stability.

Table 7 shows that: in the energy dataset, SARIMA ranks first for $\alpha = 0.0$ and $\alpha = 0.1$, while XGBoost ranks first from $\alpha = 0.2$ to $\alpha = 1.0$. This means that SARIMA is preferable for profile-oriented energy forecasting, where preserving the consumption trajectory is the main goal, while XGBoost is preferable when numerical accuracy becomes more important.

In the Netflix dataset, TimesNet ranks first only when $\alpha = 0.0$, whereas PatchTST ranks first from $\alpha = 0.1$ to $\alpha = 1.0$. This suggests that PatchTST is the most stable model once magnitude accuracy is included in the evaluation.

In the Jena Climate dataset, XGBoost is preferred for $\alpha = 0.0$, $\alpha = 0.1$, and $\alpha = 0.2$, whereas SARIMA becomes preferable from $\alpha = 0.3$ to $\alpha = 1.0$. This shows that, in weather forecasting, the preferred model also depends on whether trajectory agreement or magnitude accuracy is emphasized.

Overall, the α-sensitivity analysis demonstrates that α should not be treated as a fixed universal value. Instead, α should be selected according to the forecasting objective. The balanced value $\alpha = 0.5$ provides a useful reference, but the full α grid is necessary to understand the robustness of the model ranking.

5.6 Statistical validation

Statistical validation is used to avoid overinterpreting small differences in average HEM values. Bootstrap confidence intervals provide uncertainty estimates for the mean HEM score, while Wilcoxon signed-rank tests compare models window by window.

The bootstrap confidence intervals in Figures 4, 5 and 6 show that the best and second-best models often have overlapping uncertainty ranges. This is the case for XGBoost and SARIMA in the energy dataset, PatchTST and SARIMA in the Netflix dataset, and SARIMA and XGBoost in the Jena Climate dataset. Therefore, even when one model has the best average HEM, the difference must be interpreted cautiously if the confidence intervals overlap.

Table 7. Best model across α ranges

Dataset	α Range	Best Model	Interpretation
Energy	0.0–0.1	SARIMA	Best when trajectory fidelity dominates
Energy	0.2–1.0	XGBoost	Best when magnitude accuracy receives moderate/high importance
Netflix	0.0	TimesNet	Best only in pure correlation-oriented setting
Netflix	0.1–1.0	PatchTST	Most stable model once magnitude is included
Jena Climate	0.0–0.2	XGBoost	Best when trajectory fidelity dominates
Jena Climate	0.3–1.0	SARIMA	Best when magnitude accuracy receives moderate/high importance

Note: SARIMA = Seasonal Autoregressive Integrated Moving Average.

Table 8. Summary of Wilcoxon signed-rank tests for hybrid error metric (HEM) at $\alpha = 0.5$

Dataset	Main Comparison	*p*-Value	Interpretation
Energy	XGBoost vs SARIMA	0.2286	Not significant
Energy	XGBoost vs Prophet	0.0384	Significant
Netflix	PatchTST vs SARIMA	0.6102	Not significant
Netflix	PatchTST vs Prophet	0.0141	Significant
Jena Climate	SARIMA vs XGBoost	0.7611	Not significant
Jena Climate	SARIMA vs Prophet	0.0005	Significant

Note: SARIMA = Seasonal Autoregressive Integrated Moving Average.

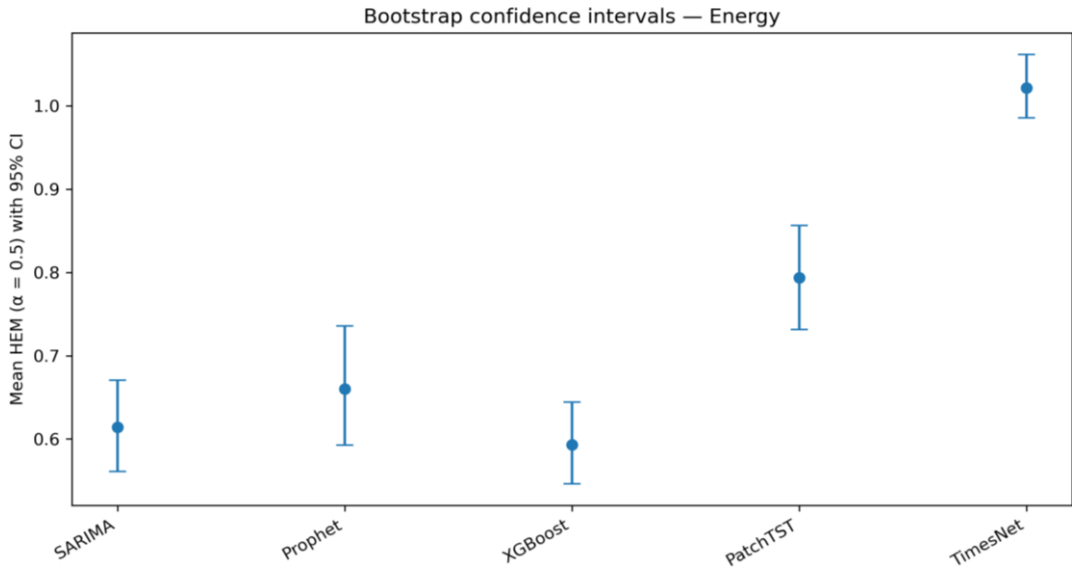

Figure 4. Bootstrap confidence intervals for mean hybrid error metric (HEM) at $\alpha = 0.5$ (Energy)

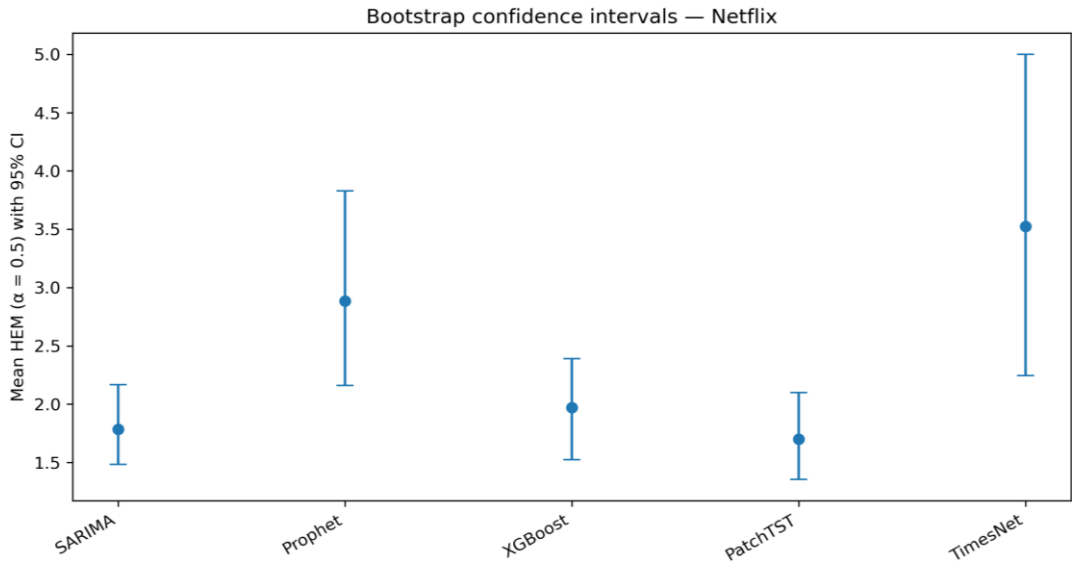

Figure 5. Bootstrap confidence intervals for mean hybrid error metric (HEM) at $\alpha = 0.5$ (Netflix)

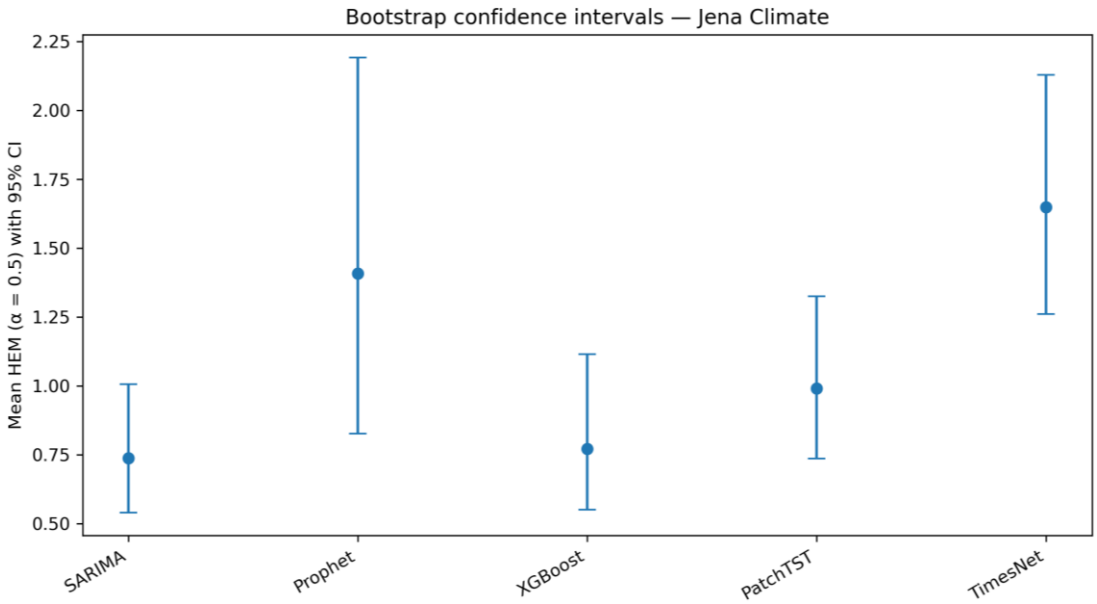

Figure 6. Bootstrap confidence intervals for mean hybrid error metric (HEM) at $\alpha = 0.5$ (Jena Climate)

Table 8 shows that the Wilcoxon signed-rank tests confirm this interpretation. On the energy dataset, XGBoost and SARIMA are not significantly different at the 5% level. On the Netflix dataset, PatchTST and SARIMA are also not significantly different. On the Jena Climate dataset, SARIMA and XGBoost are not significantly different. These results show that HEM should be interpreted together with statistical validation, especially when model scores are close.

5.7 Controlled perturbation analysis

The controlled perturbation experiments provide insight into the behavior of HEM under known forecast distortions. The perfect forecast obtains HEM = 0, confirming the lower-bound property of the metric.

Additive bias and amplitude scaling preserve the shape of the series and therefore maintain high correlation. However, HEM remains positive because the NRMSE component penalizes the magnitude distortion. This confirms that HEM does not confuse shape preservation with full forecasting accuracy.

Noise and outlier contamination lead to intermediate HEM values. These perturbations preserve part of the temporal structure but introduce local deviations. Smoothing also increases HEM because it reduces local variability and weakens the reproduction of peaks and troughs.

Temporal shifts are strongly penalized. A one-step shift already increases HEM, while a larger shift produces a stronger penalty. This demonstrates that HEM is sensitive to temporal misalignment. Finally, the inverted trajectory receives the strongest penalty because the correlation becomes negative. These results confirm that HEM reacts coherently to different types of forecast degradation.

Figures 7–9 show that HEM behaves coherently under controlled distortions. It remains zero for a perfect forecast, increases under amplitude bias and noise, strongly penalizes temporal shifts, and reaches its highest values for inverted trajectories.

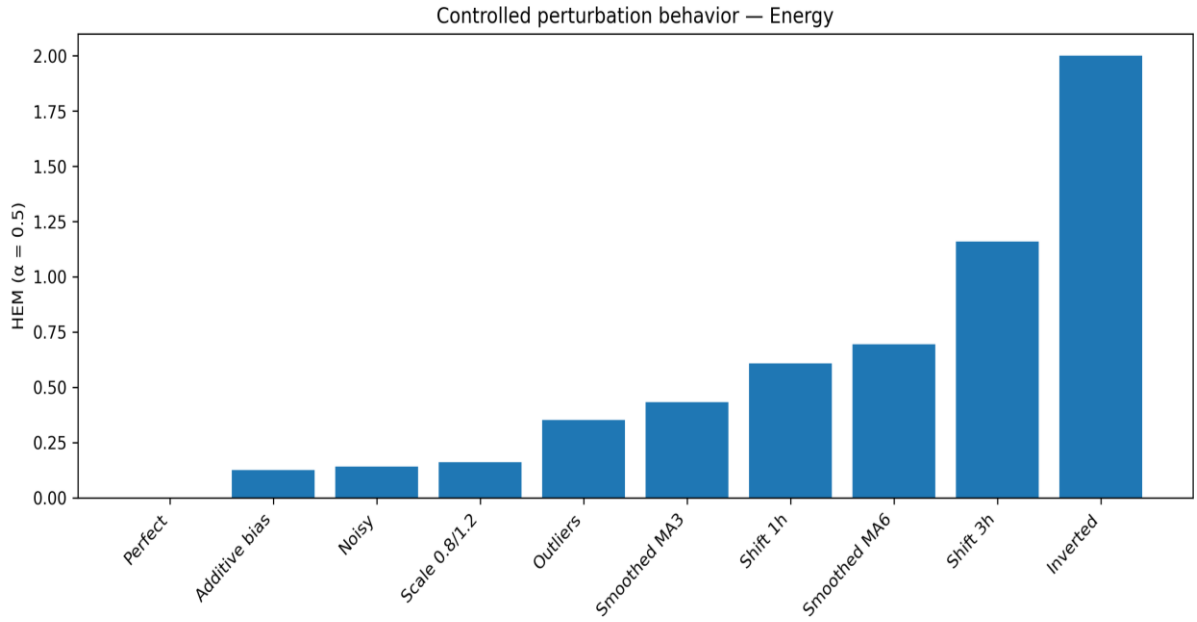

Figure 7. Controlled perturbation behavior of hybrid error metric (HEM) (Energy)

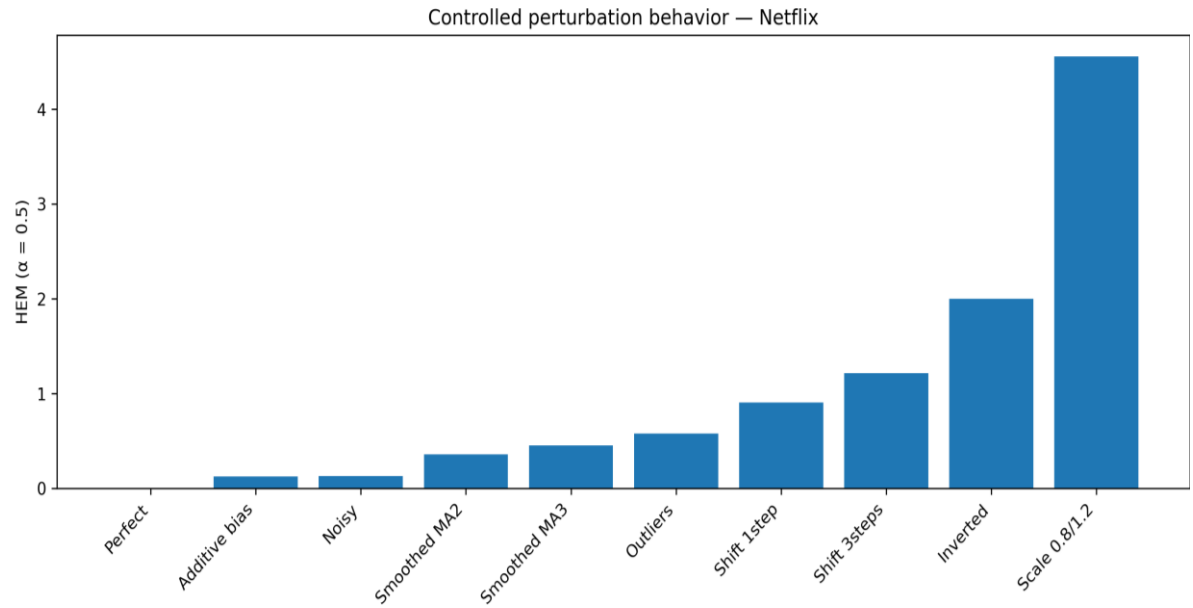

Figure 8. Controlled perturbation behavior of hybrid error metric (HEM) (Netflix)

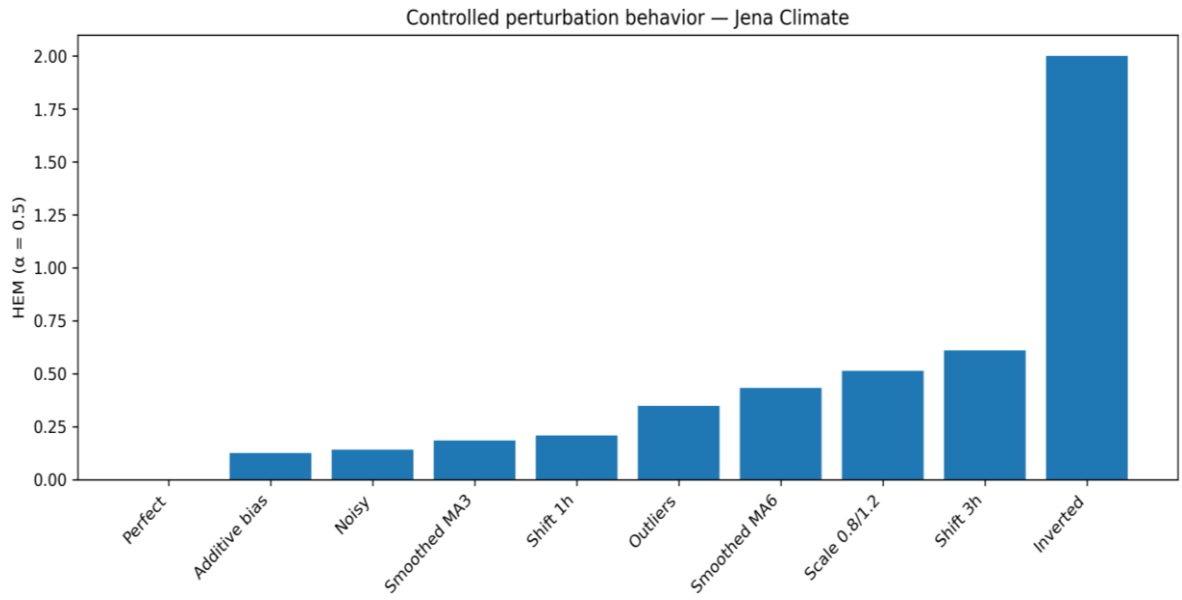

Figure 9. Controlled perturbation behavior of hybrid error metric (HEM) (Jena Climate)

6. CONCLUSION AND FUTURE WORK

This paper presented the HEM, an interpretable evaluation metric for time series forecasting that combines normalized magnitude error and temporal trajectory fidelity. HEM is based on two complementary components: a normalized RMSE term, which evaluates numerical accuracy, and a correlation-based disagreement term, which evaluates whether the forecast preserves the temporal evolution of the observed series. The parameter α controls the relative importance assigned to these two components.

The results show that HEM is useful when classical metrics lead to different interpretations. In the household energy dataset, SARIMA achieves stronger trajectory fidelity, while XGBoost achieves lower normalized magnitude error. HEM makes this distinction explicit: SARIMA is preferred when α is low and the evaluation prioritizes the consumption profile, whereas XGBoost is preferred when α increases and magnitude accuracy receives more importance. This demonstrates that α can be interpreted as a decision parameter rather than as an arbitrary weight.

The multi-domain evaluation also shows that the preferred forecasting model depends on the dataset. XGBoost obtains the best balanced HEM score on the household energy dataset, PatchTST performs best on the Netflix stock dataset, and SARIMA obtains the best balanced HEM score on the Jena Climate dataset. These results suggest that HEM does not systematically favor a specific model family. Instead, it provides a common framework for comparing statistical, machine learning, and deep learning models under different evaluation preferences.

The statistical validation provides a more cautious interpretation of the results. Bootstrap confidence intervals show that the best and second-best models often have overlapping uncertainty ranges. Wilcoxon signed-rank tests also indicate that some differences in average HEM are not statistically significant, such as XGBoost versus SARIMA on the energy dataset and SARIMA versus XGBoost on the Jena Climate dataset. Therefore, HEM should be interpreted together with uncertainty estimates and paired statistical tests, especially when model scores are close.

The controlled perturbation experiments further clarify the behavior of HEM. The metric reaches zero for a perfect forecast, increases under amplitude bias and noise, penalizes temporal shifts, and assigns strong penalties to inverted trajectories. These results support the theoretical interpretation of HEM as a metric that reacts to both magnitude errors and trajectory distortions.

Overall, HEM should be viewed as a decision-oriented metric rather than a universal replacement for existing forecasting metrics. Its main contribution is to make explicit the trade-off between predicting the correct numerical values and preserving the temporal trajectory of the series. This is useful in applications where the preferred forecasting model depends on the operational objective, such as profile-oriented energy forecasting, quantity-oriented demand estimation, financial trend analysis, or weather prediction.

Several limitations remain. First, HEM depends on the selection of α, and this parameter should be interpreted according to the application context. Although this study reports a full α-sensitivity analysis, future work could investigate automatic α calibration using explicit operational cost functions. Second, HEM currently uses Pearson correlation to measure trajectory fidelity. Other shape-aware similarity measures, such as DTW or differentiable alignment-based distances, could be explored in future extensions. Third, the deep learning models used in this study were evaluated under a lightweight configuration. Further tuning or larger-scale experiments may improve their performance.

Future work may extend HEM to probabilistic forecasting, multivariate forecasting, and anomaly-sensitive forecasting tasks. Another direction is to evaluate HEM on additional domains such as traffic, renewable energy generation, industrial sensor data, and healthcare time series. These extensions would help determine how the balance between magnitude accuracy and trajectory fidelity should be adapted to different operational settings.

REFERENCES

[1] Willmott, C.J., Matsuura, K. (2005). Advantages of the

mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1): 79-82. https://doi.org/10.3354/cr030079

[2] Hyndman, R.J., Koehler, A.B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4): 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

[3] Hyndman, R.J., Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

[4] Manandhar, P., Rafiq, H., Rodriguez-Ubinas, E., Palpanas, T. (2024). New forecasting metrics evaluated in prophet, random forest, and long short-term memory models for load forecasting. Energies, 17(23): 6131. https://doi.org/10.3390/en17236131

[5] Hebrail, G., Berard, A. (2012). Individual household electric power consumption data set. UCI Machine Learning Repository. https://doi.org/10.24432/C58K54

[6] Max Planck Institute for Biogeochemistry. (2020). Jena Climate Dataset. Kaggle. https://www.kaggle.com/datasets/mnassrib/jena-climate.

[7] Chen, T., Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA, pp. 785-794. https://doi.org/10.1145/2939672.2939785

[8] Wu, H., Xu, J., Wang, J., Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In Proceedings of the 35th International Conference on Neural Information Processing Systems, pp. 22419-22430.

[9] Makridakis, S., Spiliotis, E., Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting, 36(1): 54-74. https://doi.org/10.1016/j.ijforecast.2019.04.014

[10] Makridakis, S., Spiliotis, E., Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. International Journal of Forecasting, 38(4): 1346-1364. https://doi.org/10.1016/j.ijforecast.2021.11.013

[11] Gneiting, T., Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477): 359-378. https://doi.org/10.1198/016214506000001437

[12] Pinson, P., Tastu, J. (2013). Discrimination ability of the energy score. Technical University of Denmark.

[13] Miettinen, K. (1999). Nonlinear Multiobjective Optimization (Vol. 12). Springer Science & Business Media.

[14] Keogh, E., Ratanamahatana, C.A. (2005). Exact indexing of dynamic time warping. Knowledge and Information Systems, 7(3): 358-386. https://doi.org/10.1007/s10115-004-0154-9

[15] Cuturi, M., Blondel, M. (2017). Soft-DTW: A differentiable loss function for time-series. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 70: 894-903.

[16] Paparrizos, J., Gravano, L. (2015). K-shape: Efficient and accurate clustering of time series. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Australia, pp. 1855-1870. https://doi.org/10.1145/2723372.2737793

[17] Berndt, D.J., Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Seattle WA, pp. 359-370.

[18] Lim, B., Zohren, S. (2021). Time-series forecasting with deep learning: A survey. Philosophical Transactions of the Royal Society A, 379(2194): 20200209. https://doi.org/10.1098/rsta.2020.0209

[19] Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M. (2015). Time Series Analysis: Forecasting and Control. John Wiley & Sons.

[20] Taylor, S.J., Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1): 37-45. https://doi.org/10.1080/00031305.2017.1380080

[21] Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., Long, M. (2022). Timesnet: Temporal 2d-variation modeling for general time series analysis. arXiv preprint arXiv:2210.02186. https://doi.org/10.48550/arXiv.2210.02186

[22] Ansari, A.F., Stella, L., Turkmen, C., et al. (2024). Chronos: Learning the language of time series. arXiv preprint arXiv:2403.07815. https://doi.org/10.48550/arXiv.2403.07815

[23] Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., Dubrawski, A. (2024). Moment: A family of open time-series foundation models. In Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, pp. 16115-16152.

[24] Das, A., Kong, W., Sen, R., Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. In Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, pp. 10148-10167.

[25] Kottapalli, S.R.K., Hubli, K., Chandrashekhara, S., et al. (2025). Foundation models for time series: A survey. arXiv preprint arXiv:2504.04011. https://doi.org/10.48550/arXiv.2504.04011

[26] Hewamalage, H., Bergmeir, C., Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. International Journal of Forecasting, 37(1): 388-427. https://doi.org/10.1016/j.ijforecast.2020.06.008

[27] Liu, X., Wang, W. (2024). Deep time series forecasting models: A comprehensive survey. Mathematics, 12(10): 1504. https://doi.org/10.3390/math12101504

[28] Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. In International Conference on Learning Representations (ICLR).