Evaluating Accuracy and Probabilistic Reliability of Zero-Shot Time Series Foundation Models

Panagiotis Michael[1][0009−0008−9660−6117], Moysis Symeonides[2][0009−0007−2711−1949], and Demetris Trihinas[1][0000−0002−9540−7342]

[1] Department of Computer Science, University of Nicosia, Cyprus
michael.p15@live.unic.ac.cy, trihinas.d@unic.ac.cy
[2] Department of Computer Science, University of Cyprus, Cyprus
msymeo03@ucy.ac.cy

Abstract. Time Series Foundation Models (TSFMs) promise a paradigm shift toward zero-shot forecasting by eliminating task-specific training. However, existing works often overlook trade-offs between predictive accuracy and probabilistic calibration. This paper presents a benchmark study of six TSFMs evaluated on energy, traffic, and financial datasets. We contrast their performance against statistical baselines and a supervised DL model. The study reveals that while TSFMs outperform statistical methods and supervised models, they are subject to a fundamental trade-off between point accuracy and probabilistic reliability. Specifically, xLSTM architectures provide robust probabilistic calibration across horizons. In contrast, patch-based transformers offer competitive accuracy but face calibration issues at long horizons, while transformer-based models exhibit context saturation points for optimal zero-shot reasoning. These findings offer evidence-based guidance for balancing generalization and uncertainty quantification in real-world deployments.

Keywords: Time Series · Foundation Models · Zero-Shot Learning

1 Introduction

Time series forecasting is important for decision-making in critical domains, including energy grid management, supply chain logistics, and traffic planning. Although statistical models are common, the focus has shifted toward task-specific DL models setting new benchmarks for supervised forecasting [9]. However, these models require extensive historical data, considerable compute for optimization, and frequent retraining under new distributions or domains [13].

The success of large language models (LLMs) has inspired a new wave of forecasting research known as Time Series Foundation Models (TSFMs). TSFMs are neural networks pre-trained on thousands of time series to enable zero-shot forecasting [4]. This approach bypasses task-specific training, enabling direct deployment on unseen data through mechanisms such as token patching [6], numerical quantization [1], and generative flow matching [7]. Consequently, TSFMs offer a practical alternative for cold-start scenarios, such as forecasting demand for

new retail products or power generation for newly installed renewable microgrids where historical data is limited [5]. From a data engineering perspective, TSFMs enable a model-as-a-service paradigm, integrating directly into data pipelines and reducing the complexity of maintaining localized ML/AI workflows.

However, current TSFM evaluations are sparse and fragmented. Recent benchmarks show that while TSFMs are competitive, they do not exceed finely tuned task-specific baselines [8] and under-perform statistical methods for irregular time series [12]. Moreover, with TSFMs trained on public data, their performance can be artificially inflated on standard benchmarks. Another critical gap in the literature is the emphasis on point accuracy, neglecting probabilistic calibration. Our work bridges the gap between point accuracy and probabilistic reliability by benchmarking trade-offs among recurrent, generative, and transformer TSFMs, and evaluating them on unseen data to verify their generalization.

The contributions of this work are summarized as follows:

- A comprehensive, reproducible, and open-source[3] benchmark study on zero-shot time series forecasting, where we evaluate 6 TSFMs (Chronos-2, TiRex, Moirai-2.0, Sundial, TimesFM, and Toto) against statistical baselines and a state-of-the-art supervised DL architecture (PatchTST).
- We contrast model performance across diverse data representations and utilize newly collected datasets to ensure zero-shot forecasting on data unseen by TSFMs during their training to assess model robustness.
- The evaluation extends beyond point-error metrics, by using Interval Coverage Error (ICE) and Interval Mean Absolute Error (IMAE) to quantify the reliability of the uncertainty estimates from the TSFMs.

The rest of this paper is organized as follows: Section 2 reviews related work, Section 3 describes datasets and experimental setup, Section 4 presents the experiments and key takeaways, and Section 5 concludes and outlines future work.

2 Related Work

Prior to foundation models, the state-of-the-art was dominated by task-specific DL architectures trained on target datasets. The Transformer advanced this generation through self-attention, enabling models to capture long-range dependencies, with early adaptations such as Informer reducing the quadratic complexity of attention to handle longer contexts [13]. The paradigm matured with PatchTST [9], which segments time steps into discrete patches and processes variables independently, reducing memory overhead and setting a new supervised benchmark. Despite high accuracy, these models are constrained by large volumes of domain-specific data and significant optimization overhead, limiting their utility in zero-shot or cold-start scenarios [5].

TSFMs overcome the limitations of task-specific training by pre-training on large, diverse datasets and using zero-shot inference to generalize across unseen

[3] https://github.com/unic-ailab/TSFM-Eval

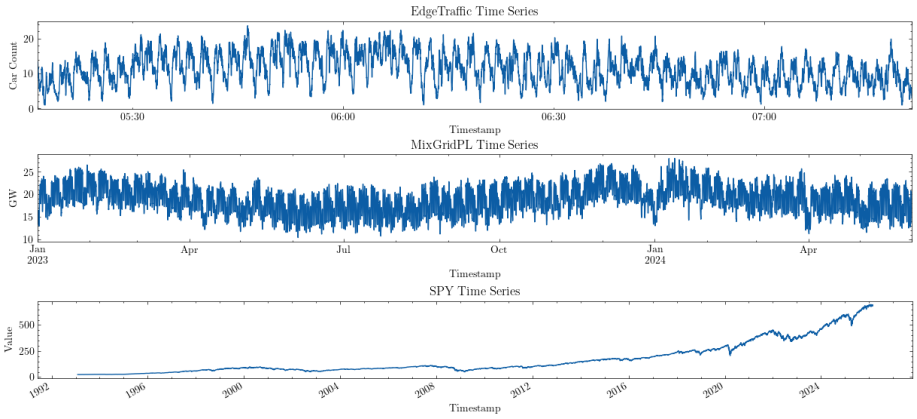

Fig. 1. High-Level Overview of the Datasets Embraced for the Benchmarking Study

distributions and domains. A key challenge is translating continuous numerical data into a format compatible with architectures designed for natural language. Chronos-2 [1] applies strict quantization, mapping continuous values into a discrete token vocabulary and treating forecasting as language modeling. Moirai-2.0 [6] and TimesFM [4] instead adopt patch-based, decoder-only autoregressive architectures, where Moirai-2.0 predicts multiple future patches simultaneously via multi-token prediction, and TimesFM lets the output patch length exceed the input patch length for long-horizon generation. Toto [3], a transformer-based TSFM embeds a Student-T mixture output head to capture heavy-tailed distributions common in time series observability data.

While attention transformers dominate, recent work proposes generative and state-tracking architectures to better model predictive uncertainty and temporal continuity. Sundial [7] approaches forecasting through generative modeling, using continuous tokenization and a flow-matching objective to sample multiple future trajectories and construct probabilistic bounds without a fixed parametric distribution, making it adaptable to volatile data. In contrast, TiRex [2] abandons attention in favor of an extended LSTM (xLSTM) architecture, where its recurrent backbone explicitly tracks hidden states over time, maintaining continuous internal memory for consistent long-range predictions.

3 Experimental Methodology

3.1 Datasets

We evaluate zero-shot generalization across three modalities (Fig. 1): discrete traffic counts (EdgeTraffic), continuous energy load (MixGridPL), and stochastic financial indices (SPY). These span a wide range of characteristics, including high versus low volatility, periodic versus stationary behavior, and temporal

Table 1. Experimental configurations for varied context (C) and horizon (H)

Dataset	Scenario	Fixed Value	Varying Range
EdgeTraffic	FixedC	$C = 900$	$H \in \{10, 60, 300, 600, 900, 1,200, 1,800\}$
	FixedH	$H = 60$	$C \in \{120, 300, 600, 900, 1,200, 1,800, 2,400\}$
MixGridPL	FixedC	$C = 240$	$H \in \{12, 24, 72, 144, 288, 576, 672\}$
	FixedH	$H = 24$	$C \in \{96, 120, 168, 240, 336, 480, 600\}$
SPY	FixedC	$C = 60$	$H \in \{1, 5, 21, 42, 63, 126, 252\}$
	FixedH	$H = 5$	$C \in \{20, 40, 60, 126, 252, 378, 504\}$

granularity. EdgeTraffic and MixGridPL are recently curated data outside known training corpora, isolating true zero-shot reasoning from data leakage.

The **EdgeTraffic** dataset [10] provides high-frequency vehicle counts from a road intersection in Iasi, Romania (7478 observations at 1s resolution over 2h window). The data (Fig. 1a) exhibits high volatility, periodic patterns, and a discrete modality that is challenging for continuous-value foundation models.

The **MixGridPL** dataset [11] tracks aggregate electricity for Poland's national power grid (12407 observations at 1h resolution) over 17 months in 2024-25 (Fig. 1b). A continuous, highly seasonal series whose recent collection ensures its temporal patterns were unavailable during TSFM development.

The Yahoo! **SPY** dataset consists of 8324 daily closing prices for SPDR S&P 500 ETF from 1993 to 2025 (Figure 1c). SPY is non-stationary with a long-term positive trend but significant short-term volatility, and percentage shifts are evaluated to gauge model responsiveness to economic trends.

3.2 Problem Description

We define a univariate time series as an ordered sequence of observations $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$. For any given reference time T, which represents the current moment in the series, the goal of zero-shot forecasting is to predict a future sequence of observations $\hat{\mathcal{X}}_{fut} = \{\hat{x}_{T+1}, \ldots, \hat{x}_{T+H}\}$ for a defined horizon H by utilizing a historical context of length C $(x_{T-C+1}, \ldots, x_T)$ without performing any gradient updates or fine-tuning on the target dataset.

To evaluate this goal across our experimental scenarios, we employ a non-overlapping rolling window evaluation scheme where each instance is constructed from a segment of length $L = C + 2H$. Each segment is divided into three parts: a context window of length C and two consecutive horizon windows of length H. The first horizon window provides ground-truth data to fairly train the supervised DL baseline (PatchTST, Section 3.4), while for the zero-shot models it is merged with the context to form an extended input of $C + H$. The second horizon window remains the consistent inference target across all models, ensuring all models are evaluated on the exact same unseen data. Subsequent segments are produced by sliding the start index forward by H, continuing while a full segment of length $C + 2H$ fits within the series.

3.3 Experiment Scenarios

To evaluate model performance, two parameters are varied for experimentation: the historical context C and prediction length H, as shown in Table 1.
– **Scenario 1 - Fixed Context (FixedC)**. For each dataset, we fix a representative context C while varying the forecast horizon H from short to long-range. This assesses how accuracy scales as the prediction distance increases.
– **Scenario 2 - Fixed Horizon (FixedH).** We keep the horizon H constant and vary the context length C. This identifies the point of "context saturation", beyond which additional history no longer improves zero-shot reasoning.

3.4 Benchmark Baselines and TSFMs

We evaluate three statistical baselines: (i) a fixed-order **ARIMA** model with local autocorrelation and non-stationarity; (ii) a **Running Average (RA)** adopting the context mean over the prediction horizon; and (iii) a **Random Walk with Drift (RWD)**, setting each future value to the previous plus a drift estimated from historical variance. These are configured in their best settings across the datasets. We also use a supervised DL baseline, **PatchTST**, configured for architectural and computational parity with the TSFMs (single encoder layer and attention head, hidden dimension of 32, and 0% dropout), to achieve inference latency comparable to benchmarked TSFMs. It also uses robust scaling to manage outliers, and is optimized with Multi-Quantile Loss (MQLoss) to estimate 10th, 50th, and 90th percentiles for direct probabilistic comparison.

We evaluate six state-of-the-art TSFMs that span diverse architectures and scales (detailed in Section 2), enabling a comparative analysis of how internal structure influences forecasting accuracy across the evaluated modalities. All adopt optimal pre-trained configurations to ensure standardized zero-shot evaluation and only adjust each model to emit the quantile forecasts required by the calibration metrics in Section 3.5. The benchmarked models and their parameter counts are: **Moirai-2** (11.4M), **TiREX** (35M), **Chronos-2** (120M), **Sundial** (128M), **Toto** (151M), and **TimesFM** (200M).

3.5 Testbed and Evaluation Metrics

Experiments are run in a virtual realm with an Nvidia T4 GPU (15GB VRAM) and 12GB of system RAM. Implementation is based on PyTorch, with all TSFMs originating from Hugging Face in their default settings. Inference is performed without task-specific fine-tuning to evaluate zero-shot forecasting.

We employ three metrics for predictive accuracy and calibration. For point accuracy, the Symmetric Mean Absolute Percentage Error (sMAPE, eq. 1) provides a percentage error measure robust to scale variations normalizing the absolute error by the mean magnitude of the observed and predicted values. For probabilistic reliability, we use Interval Coverage Error (ICE) and Interval Mean Absolute Error (IMAE) to evaluate prediction intervals. We use an 80% prediction interval with quantiles $\alpha = 0.1$ and $\beta = 0.9$. Fig. 2 depicts this setting, with

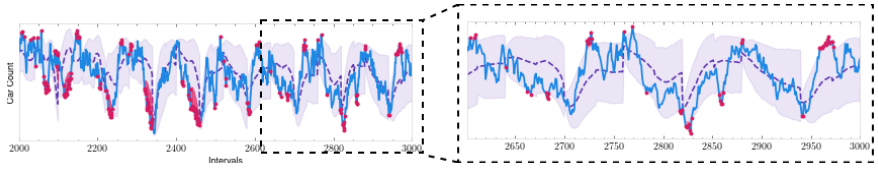

Fig. 2. Example TSFM with EdgeTraffic dataset depicting ground truth (blue line), forecast (purple line), confidence interval (shaded area), and interval violations (pink points). Zoomed-in highlight of last 500 datapoints provides a detailed overview.

the ground truth, point predictions, and shaded intervals between the quantile bounds. ICE in eq. (2) measures calibration error by comparing the observed miscoverage, the fraction of points outside the shaded band, with the nominal miscoverage of 0.20. A value of zero indicates perfect calibration. IMAE in eq. (3) complements ICE by measuring the magnitude of these violations, the vertical distance between each out-of-interval point and the nearest interval bound in Fig. 2. Thus, ICE captures how often the interval fails, while IMAE captures how severe these failures are. A model may therefore achieve low ICE but high IMAE if only a few observations fall outside the interval, but by a large margin.

$$\text{sMAPE} = \frac{100}{H} \sum_{h=1}^{H} \begin{cases} 0, & \text{if } |x_{T+h}| + |\hat{x}_{T+h}| = 0 \\ \dfrac{2\,|x_{T+h} - \hat{x}_{T+h}|}{|x_{T+h}| + |\hat{x}_{T+h}|}, & \text{otherwise} \end{cases} \tag{1}$$

$$\text{ICE}_{\alpha,\beta} = \left| \frac{1}{H} \sum_{h=1}^{H} (\mathbb{1}_{\{x_{T+h} < \hat{x}^{\alpha}_{T+h} \vee x_{T+h} > \hat{x}^{\beta}_{T+h}\}}) - (1 - (\beta - \alpha)) \right| \tag{2}$$

$$\text{IMAE}_{\alpha,\beta} = \frac{1}{H} \sum_{h=1}^{H} \left[\max(0, \hat{x}^{\alpha}_{T+h} - x_{T+h}) + \max(0, x_{T+h} - \hat{x}^{\beta}_{T+h}) \right] \tag{3}$$

Here, $\hat{x}^{\alpha}_{T+h}$ and $\hat{x}^{\beta}_{T+h}$ represent the predicted α- and β-quantiles at horizon step h, respectively. In turn, T denotes the reference time, while $h = 1, \ldots, H$ indexes the forecast horizon of length H. The variable x_{T+h} represents the ground-truth, and $\hat{x}_{T+h}$ denotes the corresponding median point prediction.

4 Evaluation

This section examines the benchmark results for predictive accuracy and probabilistic calibration, visualized per dataset in Figures 3-5. For brevity, findings are presented as plots. Full tabular data and reproduction instructions are openly available in the benchmark repository.

4.1 Predictive Accuracy Evaluation of Forecasting Models

The accuracy results show that TSFMs perform strongly across diverse settings, though their advantage depends on the structure of the underlying series. In

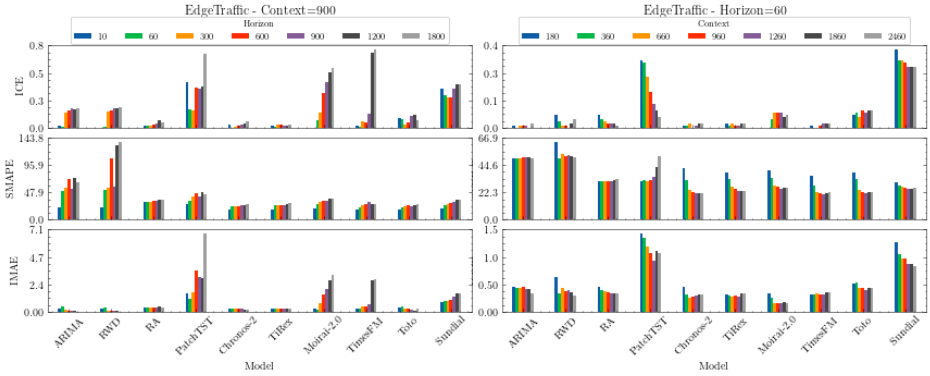

Fig. 3. Overview of Benchmark Results for the EdgeTraffic Dataset

volatile realms they remain more robust than traditional statistical methods, particularly as the forecast horizon extends.

For the discrete EdgeTraffic dataset and FixedC scenario, TimesFM and Chronos-2 achieve the best short-horizon sMAPE (17%) at $H = 10$, surpassing ARIMA (22%) and RA (32%). The gap widens at long horizons: at $H = 1800$ Chronos-2 remains the most stable (28%), while ARIMA, RA, and RWD exceed 65%, 35%, and 136%. Under FixedH, TSFMs show non-monotonic sensitivity to context. TimesFM improves from 36% at $C = 180$ to 21% at $C = 1260$, then marginally degrades at $C = 2460$ (22.5%), indicating that excessive history can introduce noise rather than improve accuracy. In contrast, supervised PatchTST collapses as context grows, rising from 31.5% at $C = 180$ to 52% at $C = 2460$. This exposes a key limitation of task-specific architectures, where they lack the "context reasoning" of TSFMs and overfit as input dimensionality grows.

Gains are more pronounced in the MixGridPL dataset, where TSFMs substantially outperform statistical baselines for all scenarios. For FixedC, Chronos-2, TiRex, and Sundial reach 3.92% to 3.95% at $H = 12$, versus 14% for ARIMA and 13% for RA; at $H = 672$, Chronos-2 and Moirai-2.0 stay at 6.4% and 6.6% while ARIMA reaches 19%. The FixedH scenario confirms that TSFMs exploit longer contexts for seasonal data, with Chronos-2 improving from 6.70% at $C = 120$ to 4.4% at $C = 624$ and Moirai-2.0 following a similar trend.

For SPY, short-horizon results are nearly identical across methods (at $H = 1$, RWD 0.79%, TiRex 0.80%, Chronos-2 0.82%), but TSFMs become competitive at long horizons, where TiRex (8.34%) and Chronos-2 (8.64%) at $H = 252$ improve over ARIMA (9.68%) and RWD (9.74%). Under FixedH ($H = 5$), most TSFMs remain stable as context grows (TiRex from 1.40% at $C = 25$ to 1.26% at $C = 509$), whereas the RA worsens from 2.50% to 14.83%, confirming that a constant-mean approach is unsuitable for trending financial data.

Key Takeaway: *TSFMs consistently match or exceed traditional baselines, with the largest gains in seasonal and structurally rich series where long-context reasoning provides a clear zero-shot advantage.*

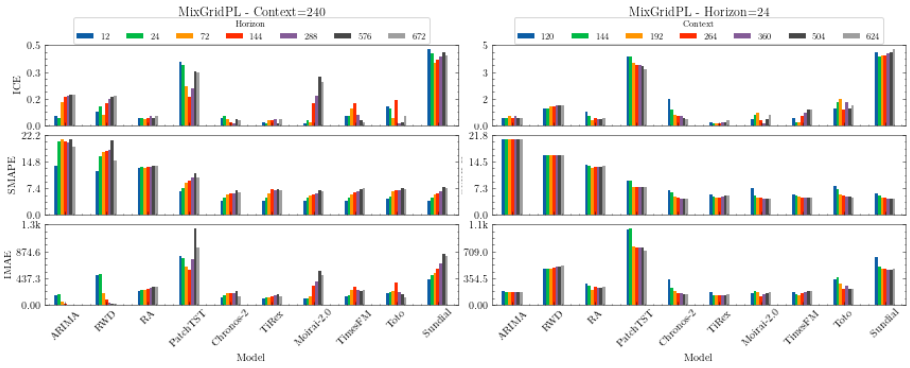

Fig. 4. Overview of Benchmark Results for the MixGridPL Dataset

4.2 Probabilistic Reliability and Interval Calibration

Several of the TSFMs maintain low sMAPE at extended horizons, but their uncertainty estimates deteriorate, leading to significant miscalibration, as shown by their ICE and IMAE scores.

For EdgeTraffic (FixedC), several transformer models suffer a *calibration collapse* as the horizon extends. TimesFM and Moirai rise from near-perfect calibration at $H = 10$ to ICE values of 0.73 and 0.56 at $H = 1800$, indicating severe over-confidence and a failure to capture observed traffic peaks where their IMAE also scales poorly (reaching 2.8 and 3.2). This denotes that missed values are missed by a wide margin. Toto is the exception among transformers, holding an ICE of 0.07 at $H = 1800$, which suggests its Student-T mixture head captures heavy-tailed distributions more robustly than the standard density heads of other patch-based models. The best performers are Chronos-2 and TiRex, both keeping ICE below 0.06 across all horizons and their slightly higher IMAE (averaging $\sim$0.28) reflects a conservative profile that is preferable when missing a traffic burst is costlier than overestimating variance. PatchTST performs worst, with an ICE of 0.7 and IMAE of 6.8 at $H = 1800$, showing that task-specific training fails to generalize uncertainty to long-range and out-of-distribution shifts.

For MixGridPL, high seasonality yields more stable behavior, although differences persist. TiRex has an ICE between 0.01 and 0.04 across all scenarios. Under FixedH, Chronos-2 improves its calibration as context grows, with ICE dropping from 0.17 ($C = 120$) to 0.04 ($C = 624$). In contrast, Sundial and PatchTST are over-confident regardless of context (ICE $\sim$0.45). Moirai achieves good short-horizon calibration (ICE 0.01 at $H = 12$), but long-range reliability degrades sharply, with IMAE rising from 109 to 479, indicating patch-based transformers struggle to propagate uncertainty through long autoregressive chains.

For SPY, the stochastic signal exposes narrow interval risk. Sundial and PatchTST exceed a 0.50 ICE, missing half of the price movements they target, while TimesFM and Moirai-2.0 show similar decay as the horizon extends. TiRex and Chronos-2 are the most robust. Under FixedH, as context grows to $C = 509$,

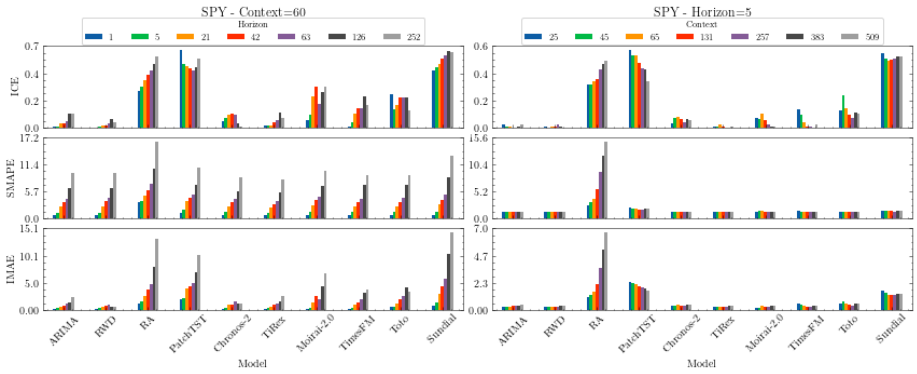

Fig. 5. Overview of Benchmark Results for the SPY Dataset

TiRex reaches a near-perfect ICE of 0.01 with a low IMAE (0.34), confirming its intervals are safe without being excessively loose. This balance makes recurrent (TiRex) and quantized-encoder (Chronos-2) architectures more suitable for risk-sensitive financial applications than purely generative or patch-based ones.
Key Takeaway: *Point accuracy does not guarantee probabilistic. TimesFM and Moirai often lead in sMAPE yet become over-confident at long horizons, whereas Chronos and TiRex provide calibrated uncertainty across modalities.*

5 Conclusion and Future Directions

This paper presented a benchmark of zero-shot TSFMs across different data modalities. TSFMs emerge as a strong alternative to statistical and lightweight supervised baselines, especially in seasonal settings where longer context windows improve performance. However, point accuracy alone is insufficient to assess real-world deployment where forecasting quality depends on a trade-off between predictive accuracy and uncertainty calibration. TSFM architecture drives this trade-off. The xLSTM-based TiRex gives the most consistent, conservative calibration across modalities, staying stable as the horizon extends, whereas patch-based transformers such as Moirai-2.0 and TimesFM often lead on accuracy yet suffer a calibration collapse, becoming over-confident at long horizons. Chronos-2 is the most balanced, pairing competitive accuracy with well-calibrated intervals. Success is also signal-dependent, where seasonal data (MixGridPL) is most favorable, discrete high-frequency traces (EdgeTraffic) challenge calibration, and stochastic financial series (SPY) with short-term volatility remain hardest. TSFM selection should thus weigh forecast horizon, signal volatility, and the cost of under-estimating uncertainty, not sMAPE alone.

Future work includes extending the benchmark to multivariate forecasting and to robustness under missing data, irregular sampling, and concept drift. A dedicated study of computational efficiency is left to forthcoming work, alongside calibration-aware forecasting and adaptive model-selection strategies.

References

1. Ansari, A.F., Shchur, O., Küken, J., Auer, A., Han, B., Mercado, P., Rangapuram, S.S., Shen, H., Stella, L., Zhang, X., Goswami, M., Kapoor, S., Maddix, D.C., Guerron, P., Hu, T., Yin, J., Erickson, N., Desai, P.M., Wang, H., Rangwala, H., Karypis, G., Wang, Y., Bohlke-Schneider, M.: Chronos-2: From univariate to universal forecasting (2025), https://arxiv.org/abs/2510.15821
2. Auer, A., Podest, P., Klotz, D., Böck, S., Klambauer, G., Hochreiter, S.: Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
3. Cohen, B., Khwaja, E., Doubli, Y., Lemaachi, S., Lettieri, C., Masson, C., Miccinilli, H., Ramé, E., Ren, Q., Rostamizadeh, A., et al.: This time is different: An observability perspective on time series foundation models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
4. Das, A., Kong, W., Sen, R., Zhou, Y.: A decoder-only foundation model for time-series forecasting. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)
5. Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J.Y., Shi, X., Chen, P.Y., Liang, Y., Li, Y.F., Pan, S., Wen, Q.: Time-LLM: Time series forecasting by reprogramming large language models. In: The Twelfth International Conference on Learning Representations (2024)
6. Liu, C., Aksu, T., Liu, J., Liu, X., Yan, H., Pham, Q., Savarese, S., Sahoo, D., Xiong, C., Li, J.: Moirai 2.0: When less is more for time series forecasting (2026), https://arxiv.org/abs/2511.11698
7. Liu, Y., Qin, G., Shi, Z., Chen, Z., Yang, C., Huang, X., Wang, J., Long, M.: Sundial: A family of highly capable time series foundation models. In: Forty-second International Conference on Machine Learning (2025)
8. Meyer, M., Gonzalez, D.Z., Kaltenpoth, S., Müller, O.: Benchmarking time series foundation models for short-term household electricity load forecasting. IEEE Access **13**, 218141–218153 (2025)
9. Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. In: The Eleventh International Conference on Learning Representations (2023)
10. Symeonides, M., Trihinas, D., Cleju, N.: Edgetraffic: An edgeai dataset integrating road traffic dynamics with system telemetry. In: Proceedings of the 4th International Workshop on Testing Distributed Internet of Things Systems. p. 13–18. TDIS '26, Association for Computing Machinery, New York, NY, USA (2026)
11. Symeonides, M., Tsiopani, N., Maouris, G., Trihinas, D., Pallis, G., Dikaiakos, M.D.: Carbonoracle: Automating energy mix & renewable energy source forecast modeling for carbon-aware micro data centers. In: 2024 IEEE/ACM 17th International Conference on Utility and Cloud Computing (UCC). pp. 246–255 (2024)
12. Toner, W., Lee, T.L., Joosen, A., Singh, R., Asenov, M.: Performance of zero-shot time series foundation models on cloud data. arXiv preprint arXiv:2502.12944 (2025)
13. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W.: Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35 (2021)