BEYOND EARTHBOUND BENCHMARKS: EVALUATING TIME SERIES FOUNDATION MODELS ON SATELLITE TELEMETRY DATA

Geremia Pompei **Deepan Anbarasan**
ContinualIST, Pisa, Italy
{geremia.pompei, deepan.anbarasan}@continualist.ai

ABSTRACT

Time series foundation models (TSFMs) excel on terrestrial benchmarks but remain untested in the space domain. We present the first systematic zero-shot evaluation of four TSFMs - Credence, TimesFM-2.5, Chronos-2, and TiRex - on the ESA Anomaly Detection Benchmark dataset. Results demonstrate generalization, with most models outperforming seasonal naive baselines. While probabilistic accuracy (CRPS) improves at lower sampling frequencies, point accuracy (MASE) degrades significantly. Overall, **Chronos-2** achieves the strongest point accuracy (leading MASE), while **Credence** provides superior probabilistic performance (leading CRPS). These findings validate TSFMs for automated mission operations and establish a baseline for future space forecasting benchmarks.

Track: Industry & Applications

1 INTRODUCTION

While Time Series Foundation Models (TSFMs) claim universal capabilities and have been rigorously tested on benchmarks like GIFT-Eval (Aksu et al., 2024) and fev-bench (Shchur et al., 2024), the space domain remains an unexplored frontier. This gap is critical given the aerospace sector's operational reliance on forecasting for early anomaly detection and preventive maintenance (Hundman et al., 2018; Kotowski et al., 2025). To address this, we present the first systematic evaluation of the forecasting performance of TSFMs in the space domain, utilizing the European Space Agency Anomaly Detection Benchmark (ESA-ADB) dataset (De Canio et al., 2024).

2 DATASET AND METHODOLOGY

We present the data sources, preprocessing steps, and evaluation protocols established to benchmark the forecasting performance of time series foundation models against a standard space industry baseline.

2.1 ESA SATELLITE TELEMETRY DATA

We utilize the *ESA Anomaly Detection Benchmark (ESA-ADB)* (Kotowski et al., 2025) dataset, a curated dataset of operational telemetry from two ESA missions spanning 14 years. This dataset is complex with over **1.5 billion data points** across 176 channels and varying sampling rates (30s for Mission 1, 18s for Mission 2), and both gradual and abrupt concept drifts.

2.2 DATASET PREPARATION AND RESAMPLING

To balance computational feasibility with sufficient context, we adopted a frequency-dependent evaluation window. To evaluate performance across multiple temporal resolutions, we downsampled the data to minute, hourly, and daily granularity. For high-frequency tasks, shown as "High Freq" in results (native and 1-minute resolutions), we utilized the final 12 months of data for Mission 1 and the final 6 months for Mission 2. Conversely, for hourly and daily resolutions (Low freq), we retained the full dataset history to ensure adequate sequence length for the input context. This

multi-resolution approach mimics operational scenarios where engineers view telemetry at varying granularities for different monitoring tasks. For each channel, we take the full resampled time series and segment it into fixed-length input sequences of 16,392 time steps to construct the evaluation samples.

2.3 Models and Evaluation Protocol

We performed zero-shot forecasting evaluation of four foundation models against a standard **Seasonal Naive** baseline. The models include **Credence**, our agentic foundation model; **TimesFM-2.5**, a decoder-only transformer (Das et al., 2024); **Chronos-2**, an encoder-only transformer (Ansari et al., 2024); and **Ti-Rex**, an xLSTM-based forecasting model (Auer et al., 2025).We report performance using **relative metrics**. We compute the Mean Absolute Scaled Error (MASE) for point forecasts and the Continuous Ranked Probability Score (CRPS) for probabilistic forecasts. All results are normalized with respect to the Seasonal Naive baseline. We assessed the models' forecasting capabilities over three distinct horizons: short-, medium-, and long-term. These correspond to progressively increasing forecast lengths, constructed as multiples of the base prediction horizon. Our evaluation protocol follows the benchmarking methodology introduced in GIFT-Eval benchmark (Aksu et al., 2024).

3 Results

Table 1 indicates that foundation models generally outperform the Seasonal Naive baseline, though performance varies significantly by metric and frequency. A notable divergence was observed in sampling rates: while **CRPS values effectively improved** in low-frequency regimes, **MASE performance degraded**, often exceeding 1.0 in downsampled forecasts.

Across the 16 measured categories, **Chronos-2** achieved the highest number of best-performing results (9), primarily demonstrating strong point forecast accuracy (MASE). In contrast, **Credence** (4 leading results) led in probabilistic reliability (CRPS), while **TiRex** surprisingly excels at short horizons but degrades on medium- and long-term forecasts, which is counter-intuituive given its xLSTM background (Beck et al., 2024).

Table 1: Zero-shot performance breakdown by sampling frequency and forecasting horizon. Results are reported for all terms (short-, medium-, and long-term) and separately for high-frequency (second- and minute-level) and low-frequency (hourly and daily) data. Lower values are better, with bold indicating the best performing model per category.

		CRPS		MASE	
Term	**Model**	**High Freq**	**Low Freq**	**High Freq**	**Low Freq**
All	Credence	**0.198051**	**0.080300**	0.541946	0.978217
	TimesFM-2.5	0.218934	0.082899	0.700942	1.206667
	Chronos-2	0.198563	0.084017	**0.495415**	**0.846023**
	TiRex	0.243099	0.099785	0.608531	0.882128
Long	Credence	0.198812	0.078308	0.622387	0.986209
	TimesFM-2.5	0.230512	**0.075646**	0.871986	**0.953756**
	Chronos-2	**0.183432**	0.086770	**0.540243**	1.050199
	TiRex	0.276504	0.093869	0.750242	1.089285
Medium	Credence	0.181260	**0.072398**	0.555847	0.791524
	TimesFM-2.5	0.204812	0.076418	0.754371	0.789535
	Chronos-2	**0.172012**	0.080687	**0.494571**	**0.780197**
	TiRex	0.232173	0.091541	0.635229	0.940313
Short	Credence	**0.215570**	0.091329	0.460100	1.199147
	TimesFM-2.5	0.222273	0.098553	0.523543	2.333208
	Chronos-2	0.248119	**0.084708**	**0.455083**	0.739045
	TiRex	0.223788	0.115625	0.472843	**0.670164**

4 Conclusion and Future Work

This study provides evidence that general-purpose time series foundation models demonstrate **robust zero-shot generalization** to space telemetry without fine-tuning. Despite unique signal characteristics and anomalies, models such as Credence and Chronos-2 proved effective. Moving forward, we aim to expand this work by integrating additional telemetry datasets, such as the Soil Moisture Active Passive (**SMAP**) satellite and the Mars Science Laboratory (**MSL**) rover datasets from **NASA** (Hundman et al., 2018) and the **OPS-SAT** benchmark by Ruszczak et al. (2025), which contains telemetry from the European Space Agency (ESA) satellite of the same name. Our goal is to consolidate such datasets into a comprehensive, multi-agency Space Forecasting Benchmark.

In addition, future work will investigate the impact of **light fine-tuning and domain adaptation** on these models, such as LoRA fine-tuning (Hu et al., 2021), evaluating whether modest amounts of mission-specific data can further improve performance or alter model rankings observed in the zero-shot setting.

Finally, while operational anomaly detection was outside the scope of this initial study, future research must investigate whether the strong forecasting accuracy of these TSFMs translates to robust downstream anomaly detection, particularly concerning thresholding strategies, false-alarm reduction, and compute/latency constraints.

References

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex: Zero-shot forecasting across long and short horizons with enhanced in-context learning, 2025. URL https://arxiv.org/abs/2505.23719.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory, 2024. URL https://arxiv.org/abs/2405.04517.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 10620–10641. PMLR, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Gabriele De Canio, Krzysztof Kotowski, and Christoph Haskamp. ESA anomaly dataset, June 2024. URL https://doi.org/10.5281/zenodo.12528696.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 387–395, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219845. URL https://doi.org/10.1145/3219819.3219845.

Krzysztof Kotowski, Christoph Haskamp, Jacek Andrzejewski, Bogdan Ruszczak, Jakub Nalepa, Daniel Lakey, Peter Collins, Aybike Kolmas, Mauro Bartesaghi, Jose Martinez-Heras, and Gabriele De Canio. European space agency benchmark for anomaly detection in satellite telemetry, 2025. URL https://arxiv.org/abs/2406.17826.

Bogdan Ruszczak, Krzysztof Kotowski, David Evans, and Jakub Nalepa. The OPS-SAT benchmark for detecting anomalies in satellite telemetry. *Scientific Data*, 12(1):710, 2025. doi: 10.1038/s41597-025-05035-3. URL https://doi.org/10.1038/s41597-025-05035-3. Erratum in: Sci Data. 2025;12(1):1445.

Oleksandr Shchur, Ali Caner Turkmen, Tim Januschowski, and Jan Gasthaus. FEV-Bench: A holistic benchmark for zero-shot time series foundation models. *arXiv preprint arXiv:2410.10393*, 2024.