A Comparative Analysis of Forecasting Foundation Models for Spacecraft Multivariate Time Series Anomaly Detection

Anonymous CVPR submission

Paper ID 47

Abstract

Detecting anomalies in spacecraft telemetry is critical for early warning of potential system failures. Forecasting foundation models have shown success across various domains. However, their effectiveness for time series anomaly detection has not been thoroughly evaluated, especially for spacecraft telemetry data. We explore the use of multiple pre-trained forecasting foundation models for anomaly detection in spacecraft telemetry, including MOIRAI (Masked Encoder-based Universal Time Series Forecasting Transformer), Toto, Chronos and TimesFM. We use these models to forecast multivariate time series and estimate the forecast error between actual and predicted values. We evaluate both zero-shot and fine-tuned approaches, comparing univariate models (Chronos, TimesFM) with multivariate models (MOIRAI, Toto). We evaluate the performance on a Mars Reconnaissance Orbiter (MRO) dataset with multiple telemetry channels and four injected anomaly types. Results show that Toto achieves the best zero-shot performance with an F1-score of 0.752, while fine-tuned MOIRAI achieves the best overall performance with an F1-score of 0.859. Domain-specific fine-tuning and multivariate processing improve anomaly detection performance for spacecraft telemetry.

1. Introduction

Spacecraft time series anomaly detection is critical for mission success and safety, as early detection of anomalous behavior can prevent catastrophic failures and enable timely corrective actions [10]. Traditional approaches rely on threshold-based methods or classical machine learning techniques, but these often struggle with the complex, multivariate nature of spacecraft telemetry data [15]. The high dimensionality, temporal dependencies and domain-specific patterns in spacecraft data present unique challenges that require sophisticated modeling approaches.

Despite the effectiveness of specialized approaches, their deployment in spacecraft operations faces a significant scalability bottleneck. Modern spacecraft generate telemetry across thousands of distinct channels and training individual, specialized models for each channel is computationally prohibitive and operationally complex [3]. Furthermore, specialized models often struggle with the cold-start problem in new missions where historical training data is scarce [4]. Forecasting foundation models offer a solution to these challenges by leveraging transfer learning from massive, diverse pre-training datasets [22]. If these models can successfully generalize to spacecraft telemetry, they could serve as universal backbones, reducing the need for channel-specific architecture tuning and enabling anomaly detection in scenarios where traditional methods fail.

Forecasting foundation models can learn generalizable temporal patterns and dependencies [7]. By forecasting expected behavior, deviations between predictions and actual observations can reveal anomalies. However, the application of forecasting foundation models to spacecraft anomaly detection remains understudied. While forecasting foundation models aim to minimize forecast error by capturing all temporal dynamics, anomaly detection relies on the inability of the model to predict anomalous events [5]. Additionally, these models face challenges including limited training data scale and potential generalization issues when applied to specialized domains like spacecraft telemetry [2].

In this work, we present a comparative analysis of four forecasting foundation models for spacecraft multivariate time series anomaly detection: MOIRAI (Masked Encoder-based Universal Time Series Forecasting Transformer) [22], Toto [7], Chronos [1] and TimesFM [8]. We assess both zero-shot and fine-tuned strategies. We show that anomaly detection performance is driven by context length, model capacity and multivariate processing capability. We also show that fine-tuning on domain-specific data improves performance, and that a fine-tuned smaller model can outperform larger zero-shot models. We conclude that forecasting foundation models are viable for spacecraft anomaly detection, although traditional specialized models currently require lower computational resources for inference.

Table 1. Characteristics of the forecasting foundation models MOIRAI (Small), Toto, Chronos-T5 (Small) and TimesFM.

Characteristic	**MOIRAI (Small)** [22]	**Toto** [7]	**Chronos-T5 (Small)** [1]	**TimesFM** [8]
Univariate	✓	✓	✓	✓
Multivariate	✓	✓	×	×
Type of forecasting	Probabilistic	Probabilistic	Probabilistic	Point
Transformer architecture	Encoder-Only	Decoder-Only	Encoder-Decoder	Decoder-Only
Number of parameters	14M	103M	46M	200M
Evaluation settings	Zero-Shot + Fine-Tuning	Zero-Shot	Zero-Shot + Fine-Tuning	Zero-Shot
Pre-training dataset size	∼27B	∼1T	∼84B	∼100B
Pre-training dataset domains	Energy, Transport, Weather/Climate, Economics & Finance, Healthcare, Web, IT/CloudOps, Nature, Sales	Energy, Transport, Weather/Climate, Economics & Finance, Healthcare, Web, IT/CloudOps, Nature, Sales	Energy, Transport, Weather/Climate, Economics & Finance, Healthcare, Retail, Web	Energy, Transport, Weather/Climate, Economics & Finance, Demographic, Industry, Web
Open weights	✓	✓	✓	✓

2. Related Work

2.1. Anomaly Detection via Forecasting

Anomaly detection in time series is often framed as a self-supervised forecasting task [20], where a model learns the nominal dynamics of a system and identifies anomalies as significant deviations between predicted and actual values. Traditional statistical methods like autoregressive integrated moving average (ARIMA) are effective for linear, low-dimensional data but struggle with complex non-linear dependencies [19]. Deep learning approaches, utilizing Recurrent Neural Networks (RNNs) [18], specifically Long Short-Term Memory (LSTM) networks [10], advanced the field by capturing long-term temporal correlations [4]. More recently, transformer architectures [21] were introduced for handling sequential data, leveraging self-attention mechanisms to capture both long and short-term dependencies. The potential of transformers in anomaly detection has been demonstrated in [12] [6]. However, training these deep learning architectures from scratch requires extensive domain-specific data, presenting a cold-start challenge for new spacecraft missions where historical telemetry is scarce [17].

2.2. Foundation Models for Time Series

To mitigate the data scarcity and generalization limitations of task-specific models, foundation models leverage representations learned from extensive pre-training on diverse datasets. We examine four forecasting foundation models: MOIRAI (Small) [22], Toto [7], Chronos-T5 (Small) [1] and TimesFM [8].

2.2.1. MOIRAI

MOIRAI (Masked Encoder-based Universal Time Series Forecasting Transformer) [22] is designed to address heterogeneity in frequency and variate count. The method utilizes a transformer encoder-only architecture where an any-variate attention mechanism flattens multivariate data into a single 1D sequence. This structure enables the learning of correlations both across time and between channels within a unified attention mechanism. Pre-trained on the LOTSA dataset [22] (containing 27 billion observations), MOIRAI predicts the parameters of a parametric mixture distribution (Student's t, Negative Binomial, Log-Normal and Low Variance Normal) for flexible probabilistic forecasting. The model is available in three sizes: small, base and large, with 14M, 91M and 311M parameters respectively. We evaluate the small model.

2.2.2. Toto

Toto [7] utilizes a decoder-only transformer architecture trained on a dataset of over 1 trillion data points from Datadog's cloud infrastructure and LOTSA [22]. The pre-training on real-world infrastructure data allows the framework to accommodate the ragged and irregular sampling characteristic of system telemetry. A proportional factorized space-time attention mechanism is used to separate temporal attention from cross-variate attention. This strategy captures dependencies in high-dimensional multivariate data while mitigating the computational overhead of full attention mechanism [21].

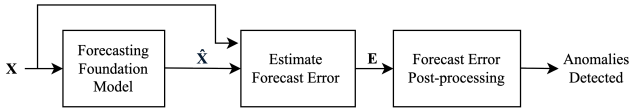

Figure 1. Block diagram of our approach for spacecraft multivariate time series anomaly detection using forecasting foundation models. $\mathbf{X}$ is the multivariate time series input. $\hat{\mathbf{X}}$ is the forecast output. $\mathbf{E}$ is the estimated forecast error.

2.2.3. Chronos

Chronos [1] approaches time series forecasting as a language modeling task. The framework quantizes continuous values into a fixed vocabulary of discrete tokens using a scaling and binning strategy. Constructed upon the T5 encoder-decoder architecture [16], the network was trained on a mixture of real-world public datasets and synthetically generated time series to enhance generalization. The joint distribution of token sequences is learned by minimizing Categorical Cross-Entropy (CCE) loss. Probabilistic forecasts are subsequently generated via sampling, a strategy that avoids the need for explicit parametric assumptions regarding the data distribution. The pre-trained models range from 20M to 710M parameters. We evaluate the small model (46M parameters).

2.2.4. TimesFM

TimesFM [8] employs a decoder-only transformer architecture trained on a corpus of over 100 billion time points, incorporating real-world and synthetic data. A patching mechanism segments input time series into tokens which are subsequently processed by a transformer decoder. To predict a specified future length in a single forward pass, the framework applies horizon-conditional forecasting [24]. The model is trained with a point forecasting objective, minimizing Mean Squared Error (MSE) loss.

3. Proposed Approach

Figure 1 shows the block diagram of our approach for multivariate time series anomaly detection using forecasting foundation models. There are three parts in our system: we first use a forecasting foundation model (pre-trained or fine-tuned) for time series forecasting. Secondly, we estimate the forecast error as the squared difference between the multivariate time series and the forecast. Finally, we post-process the forecast error to determine the temporal location and the affected channel of the anomaly.

3.1. Probabilistic Time Series Forecasting

Let $\mathbf{X} \in \mathbb{R}^{M \times N}$ denote the multivariate spacecraft telemetry time series (multiple channels of telemetry signals from the same spacecraft) consisting of M channels and N time samples. Uppercase variables correspond to the $M \times N$ samples of a multivariate time series. Bold lowercase variables with subscripts denote vectors of length M (channel state at a specific time), while bold lowercase variables with superscripts denote vectors of length N (full time series of a specific channel). Therefore, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ where $\mathbf{x}_t \in \mathbb{R}^M$ represents the state of all telemetry channels at time step t. Similarly, $\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)}]^T$ where $\mathbf{x}^{(i)} \in \mathbb{R}^N$.

The forecasting task is formulated as a conditional generation problem over a sliding window. A context window of length c is utilized to forecast a horizon of p future steps. For a given time index t, the input context is defined as $\mathbf{W}_t = \mathbf{X}_{t-c+1:t} \in \mathbb{R}^{M \times c}$. The forecasting foundation model, denoted as a function f_θ parameterized by weights θ, estimates the parameters Φ_t of a joint probability distribution over the future horizon:

$$\Phi_t = f_\theta(\mathbf{W}_t), \tag{1}$$

where Φ_t characterizes the conditional distribution $P(\mathbf{X}_{t+1:t+p} | \mathbf{W}_t; \Phi_t)$. For example, MOIRAI predicts the parameters of a parametric mixture distribution.

To obtain a point forecast $\hat{\mathbf{X}} \in \mathbb{R}^{M \times p}$ from the probabilistic output, the median (0.5-quantile) of the predicted distribution is computed:

$$\hat{x}_\tau^{(i)} = \text{median}\left(P(x_\tau^{(i)} | \mathbf{W}_t; \Phi_t)\right), \tag{2}$$

for each channel $i \in \{1, \ldots, M\}$ and future time step $\tau \in \{t+1, \ldots, t+p\}$.

3.2. Forecast Error Estimation

Anomalies are detected based on the deviation between the actual telemetry $\mathbf{X}$ and the point forecast $\hat{\mathbf{X}}$. The forecast error matrix $\mathbf{E} \in \mathbb{R}^{M \times N}$ is defined using the squared difference between actual and predicted values. For the i-th channel at time t, the error is given by:

$$e_t^{(i)} = \left(x_t^{(i)} - \hat{x}_t^{(i)}\right)^2. \tag{3}$$

The squared difference emphasizes large deviations, facilitating the detection of abrupt telemetry changes while mitigating the impact of nominal noise.

3.3. Post-Processing and Anomaly Detection

Following [14], large periodic spikes are removed from the forecast error of each channel in $\mathbf{E}$, based on the assumption that anomalies are not periodic. Anomalies in spacecraft telemetry typically lack periodicity as they originate from unpredictable events such as environmental disturbances (e.g., radiation spikes, micrometeoroid impacts) [14]. The resulting sequence is defined as the updated forecast error $\mathbf{E}_u$.

Finally, the updated forecast error $\mathbf{E}_u$ is thresholded to classify anomalies. Let τ_{thresh} be the detection threshold. The binary anomaly indicator matrix $\mathbf{A} \in \{0,1\}^{M \times N}$ is defined as:

$$a_t^{(i)} = \begin{cases} 1 & \text{if } e_{u,t}^{(i)} > \tau_{\text{thresh}} \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where $e_{u,t}^{(i)}$ is the updated forecast error for channel i at time t and $a_t^{(i)} = 1$ indicates a detected anomaly in channel i at time t.

4. Experimental Results

4.1. Datasets

The SMAP/MSL dataset [10] is commonly used to evaluate spacecraft anomaly detection. However, concerns have been raised regarding the validity of this dataset [23]. We instead use a Mars Reconnaissance Orbiter (MRO) [9] dataset, which comprises telemetry data collected from thousands of distinct channels encompassing all spacecraft subsystems. We take a subset of the MRO dataset containing 50 channels. Each channel is sampled irregularly, with measurements recorded only when a signal change occurs. Additionally, the dataset is completely unlabeled, as the precise locations of anomalous behaviors are unknown. Given that the data for each channel concludes at a system reboot, we operate under the assumption that anomalous behavior occurs in proximity to the reboot.

To address the challenges posed by irregular sampling, we extracted a one-week window of data preceding the system reboot. From this extracted window, we used the first five days for our experiments. We normalized each channel to the range $[-1, 1]$. To account for channels with a lot of unchanging values, channels with a standard deviation less than 0.1 were removed. We then organized the data into seven distinct groups, with each group consisting of 10 randomly selected telemetry channels.

To evaluate model performance while preserving temporal dependencies, we use a chronological splitting strategy. The processed time series for each channel is partitioned into three sequential subsets: the first 13,000 samples are allocated to the training set, followed by 1,000 samples for the validation set and the final 1,000 samples for the test set.

Since the original dataset lacks ground truth for specific anomaly types, we introduced four distinct types of synthetic anomalies into the test set. These anomalies, with durations ranging from 10 to 40 samples, were distributed across all seven test groups, resulting in a total of 27 injected anomalies. The specific anomaly types are defined as follows:

Noise: A sequence with added uniformly distributed noise at the anomaly location.

Peak: A large value is added at the center of the anomaly location, with the remaining values replaced by linear interpolation.

Pit: A small value is added at the center of the anomaly location, with the remaining values replaced by linear interpolation.

Data Lost: All values within the anomaly location are set to the most recent previous value, *i.e.* freeze the data until the anomaly ends.

All experiments were conducted using a single NVIDIA A40 GPU with 48 GB of memory.

4.2. Evaluation Criteria

To evaluate the forecasting performance, Mean Square Error (MSE) and Mean Absolute Error (MAE) are computed, averaged across all channels and samples:

$$MSE = \frac{1}{NM} \sum_{t=1}^{N} \sum_{i=1}^{M} \left(x_t^{(i)} - \hat{x}_t^{(i)} \right)^2, \tag{5}$$

$$MAE = \frac{1}{NM} \sum_{t=1}^{N} \sum_{i=1}^{M} \left| x_t^{(i)} - \hat{x}_t^{(i)} \right|, \tag{6}$$

where:

- N is the number of time samples,
- M is the number of channels,
- $x_t^{(i)}$ is the actual value of the i-th channel at time t,
- $\hat{x}_t^{(i)}$ is the predicted value of the i-th channel at time t.

We define the evaluation criteria for anomaly detection identical to the rules used in [10]:

- True positive (TP): Any portion of a true anomaly sequence overlaps with any detected anomaly sequence
- False positive (FP): Detected anomaly sequence does not overlap with any true anomaly sequence
- False negative (FN): True anomaly sequence does not overlap with any detected anomaly sequence

We now define our primary metrics to evaluate the anomaly detection performance:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{8}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{9}$$

High precision minimizes false positives, whereas high recall reduces false negatives. The F1-score evaluates the balance between these two, measuring the ability of the model to identify true positives while limiting both types of errors.

Table 2. Zero-shot forecasting and anomaly detection performance of MOIRAI (Small), Toto, Chronos-T5 (Small) and TimesFM across context lengths $c \in \{128, 256, 512, 1024\}$ and prediction lengths $p \in \{10, 20, 40\}$. For each metric, bold values indicate the best result across all (c, p) configurations.

		$c = 128$			$c = 256$			$c = 512$			$c = 1024$		
Model	**Metric**	p=10	p=20	p=40	p=10	p=20	p=40	p=10	p=20	p=40	p=10	p=20	p=40
MOIRAI (Small)	MSE (↓)	0.063	0.117	0.268	**0.056**	0.101	0.204	0.058	0.102	0.203	0.060	0.107	0.211
	MAE (↓)	0.102	0.155	0.273	**0.090**	0.135	0.225	0.093	0.140	0.229	0.101	0.150	0.239
	Precision (↑)	0.547	0.370	0.168	0.626	0.557	0.190	0.578	0.523	0.200	**0.705**	0.472	0.212
	Recall (↑)	0.814	0.814	0.838	0.857	0.843	0.843	**0.871**	0.843	0.843	0.843	0.843	0.843
	F1 (↑)	0.487	0.329	0.180	0.597	0.545	0.204	0.546	0.496	0.197	**0.626**	0.416	0.208
	Latency (↓)	1.12s	0.79s	0.59s	1.61s	0.99s	0.70s	2.67s	1.52s	0.96s	4.72s	2.58s	1.48s
Toto	MSE (↓)	0.059	0.109	0.226	0.054	0.097	0.194	0.049	0.072	0.083	0.038	0.041	**0.033**
	MAE (↓)	0.087	0.134	0.230	0.080	0.114	0.187	0.073	0.095	0.114	**0.065**	0.070	0.072
	Precision (↑)	0.530	0.598	0.184	0.657	0.602	0.236	0.795	0.573	0.671	0.773	0.767	**0.821**
	Recall (↑)	0.814	0.871	0.857	0.829	0.871	0.857	0.871	0.857	0.857	**0.914**	0.900	0.886
	F1 (↑)	0.488	0.571	0.188	0.605	0.570	0.242	0.742	0.527	0.583	**0.752**	0.708	0.744
	Latency (↓)	17.13s	8.86s	4.73s	29.61s	15.10s	7.81s	55.30s	27.89s	14.17s	123.20s	56.21s	28.30s
Chronos-T5 (Small)	MSE (↓)	0.060	0.102	0.193	0.061	0.104	0.179	0.057	0.092	0.152	**0.056**	0.092	0.155
	MAE (↓)	0.093	0.139	0.222	0.091	0.135	0.205	0.082	0.118	0.177	**0.081**	0.118	0.180
	Precision (↑)	0.430	0.358	0.145	**0.474**	0.355	0.201	0.413	0.362	0.268	0.389	0.375	0.258
	Recall (↑)	0.824	0.829	0.829	0.838	0.833	0.814	0.843	0.843	**0.900**	0.838	0.829	0.871
	F1 (↑)	0.386	0.347	0.162	**0.451**	0.334	0.163	0.388	0.368	0.262	0.362	0.365	0.248
	Latency (↓)	10.14s	9.80s	9.21s	10.21s	9.50s	9.34s	11.71s	10.52s	9.70s	24.30s	11.00s	9.67s
TimesFM	MSE (↓)	0.062	0.115	0.241	0.057	0.101	0.202	0.053	0.085	0.169	**0.048**	0.077	0.135
	MAE (↓)	0.099	0.151	0.256	0.088	0.135	0.228	0.087	0.125	0.206	**0.085**	0.120	0.184
	Precision (↑)	0.510	0.393	0.189	0.721	0.626	0.199	0.707	0.619	0.242	**0.750**	0.695	0.270
	Recall (↑)	0.843	0.829	0.829	0.800	0.857	0.829	0.843	0.886	0.843	0.843	0.886	**0.914**
	F1 (↑)	0.490	0.390	0.200	0.638	0.629	0.199	0.638	0.617	0.230	**0.695**	0.682	0.294
	Latency (↓)	1.60s	3.22s	1.65s	0.81s	3.22s	1.61s	0.81s	4.00s	2.01s	1.01s	9.19s	4.59s

4.3. Zero-Shot Performance Comparison

In this section, we evaluate the zero-shot performance of MOIRAI (Small), Toto, TimesFM and Chronos-T5 (Small) on the MRO dataset. Each model is evaluated using four context lengths ($c \in \{128, 256, 512, 1024\}$) and three prediction lengths ($p \in \{10, 20, 40\}$). The context lengths are chosen to range from half a period to several periods of the quasi-periodic signal of the dataset ($\sim$230 samples). The prediction lengths are chosen to cover the full range of injected anomaly durations, from the shortest (10 samples) to the longest (40 samples). Table 2 shows the MSE, MAE, precision, recall, F1-score and latency for each (c, p) configuration, with bold values denoting the best result per metric. The forecasting metrics, MSE and MAE, were evaluated on the anomaly-free validation dataset. Latency is defined as the inference time required for the model to forecast 1,000 samples. The anomaly detection metrics (precision, recall and F1-score) were computed using the test dataset, which contains the injected anomalies.

4.3.1. Impact of context and prediction length

Table 2 shows two consistent trends across models. First, F1-score improves with longer context length: for example, the F1 of Toto rises from 0.488 at c=128 to 0.752 at c=1024 and that of TimesFM from 0.490 to 0.695 over the same range (both at p=10). This is also reflected

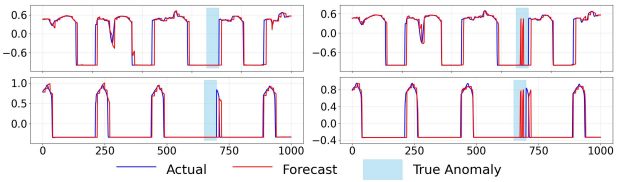

Figure 2. Toto zero-shot forecasts (p=10) for channel ten of group 1 (top row) and channel one of group 2 (bottom row), at context length c=128 (left column) and c=1024 (right column). At c=128, the forecast copies recent input samples, suppressing the error in the anomalous region (shaded in blue) and making the anomaly undetectable. At c=1024, the model learns the nominal periodic behavior and diverges from the input during the anomaly, producing large forecast errors that enable detection.

in the forecasting metrics, with MSE and MAE decreasing as context length increases for most models. Second, shorter prediction lengths yield higher F1-scores, with p=10 outperforming p=40 across all models. As a result, MOIRAI, Toto and TimesFM all achieve their best F1-score at (c=1024, p=10), with scores of 0.626, 0.752 and 0.695 respectively. Chronos is the exception, reaching its highest F1-score earlier at (c=256, p=10) with F1=0.451 and degrading for larger context lengths.

Figure 2 shows the zero-shot forecasts with context lengths c=128 (left column) and c=1024 (right column) of Toto for two channels from group 1 and 2 of the test dataset. At c=128, the forecast closely follows the input with a small phase delay. This means the model copies recent samples instead of learning the periodic structure of the signal. This copying behavior continues in the anomalous region (shaded in blue). The model predicts the anomalous signal, keeping the forecast error low. This suppresses the error exactly where a large deviation is needed to detect the anomaly. At c=1024, the model observes multiple complete oscillation cycles and learns the nominal behavior. When an anomaly changes this expected pattern, the forecast diverges from the input. We can see this as rapid oscillations of the red line in the anomalous samples in the right column. This divergence produces large forecast errors that stand out clearly against the low errors during nominal operation. This difference between nominal and anomalous forecast errors translates into the F1-score improvements seen at larger context lengths.

4.3.2. Impact of model capacity and multivariate processing

The ranking of the highest F1-scores is Toto (0.752), TimesFM (0.695), MOIRAI (0.626) and Chronos (0.451). This ranking aligns with two architectural factors shown in Table 1: model capacity and multivariate processing capability. Grouping by capacity, the two larger models, Toto (103M parameters) and TimesFM (200M parameters), outperform the two smaller models, MOIRAI (14M parameters) and Chronos (46M parameters). Within each capacity group, the multivariate model outperforms its univariate counterpart, with Toto exceeding TimesFM by 0.057 F1 and MOIRAI exceeding Chronos by 0.175 F1.

4.4. Fine-Tuning Performance

We fine-tune MOIRAI and Chronos on all seven groups of the MRO training dataset, using 70 different channels in total. These models were selected due to the availability of their open-source fine-tuning frameworks. Both models are evaluated across context lengths $c \in \{64, 128, 256, 512, 768, 1024\}$ and prediction lengths $p \in \{10, 20, 40\}$. Figure 3 shows the resulting MSE and MAE on the validation set across all evaluated context and prediction lengths.

4.4.1. MOIRAI Fine-Tuning

We fine-tune MOIRAI using the Adam optimizer [11] with a learning rate of 10^{-4}, a batch size of 32, a patch size of 32 and early stopping with a patience of 3 epochs monitoring the validation loss.

MOIRAI predicts the parameters of a mixture distribution composed of four components: a Student's t-distribution, a Negative Binomial, a Log-Normal and a Low Variance Normal distribution:

$$P(x_t^{(i)}|\mathbf{W}_t) = \sum_{k=1}^{4} w_{t,k}^{(i)} \cdot \mathcal{D}_k(x_t^{(i)}|\theta_{t,k}^{(i)}), \qquad (10)$$

where $w_{t,k}^{(i)}$ are the time-step-specific mixing weights for channel i summing to 1 over k and $\mathcal{D}_k$ is the k-th component distribution with parameters $\theta_{t,k}^{(i)}$. The fine-tuning objective minimizes the Negative Log-Likelihood (NLL):

$$\mathcal{L}_{\text{MOIRAI}} = -\sum_{i=1}^{M}\sum_{t=1}^{p} \log\left(\sum_{k=1}^{4} w_{t,k}^{(i)} \cdot \mathcal{D}_k(x_t^{(i)}|\theta_{t,k}^{(i)})\right). \qquad (11)$$

As shown in Figure 3 (a-c), MSE and MAE decrease as c increases up to c=256 after which gains plateau. This is consistent with the quasi-periodicity of approximately 230 samples in the dataset.

4.4.2. Chronos Fine-Tuning

Since Chronos does not natively support multivariate time series, we fine-tune it by processing each of the 70 channels independently as univariate sequences. We use the AdamW optimizer [13] with a learning rate of 10^{-3}, a batch size of 32 and early stopping with a patience of 3 epochs monitoring the validation loss.

Chronos quantizes the continuous time series values x_t into discrete tokens z_t using a fixed vocabulary. The fine-tuning objective minimizes the Categorical Cross-Entropy (CCE) loss:

$$\mathcal{L}_{\text{Chronos}} = -\sum_{t=1}^{T} \log P_\theta(z_{t+1}|z_1, \ldots, z_t), \qquad (12)$$

where P_θ is the probability distribution over vocabulary tokens. Figure 3 (d-f) shows that Chronos follows the same trend as MOIRAI, with MSE and MAE decreasing up to c=256 and plateauing beyond this point.

4.5. Anomaly Detection Performance

Table 3 shows the F1-score for both fine-tuned models across all evaluated configurations. Fine-tuning improves anomaly detection performance compared to zero-shot: MOIRAI improves from 0.626 to 0.859 and Chronos from 0.451 to 0.624 in F1-score. Across all configurations, MOIRAI consistently achieves higher F1-scores than

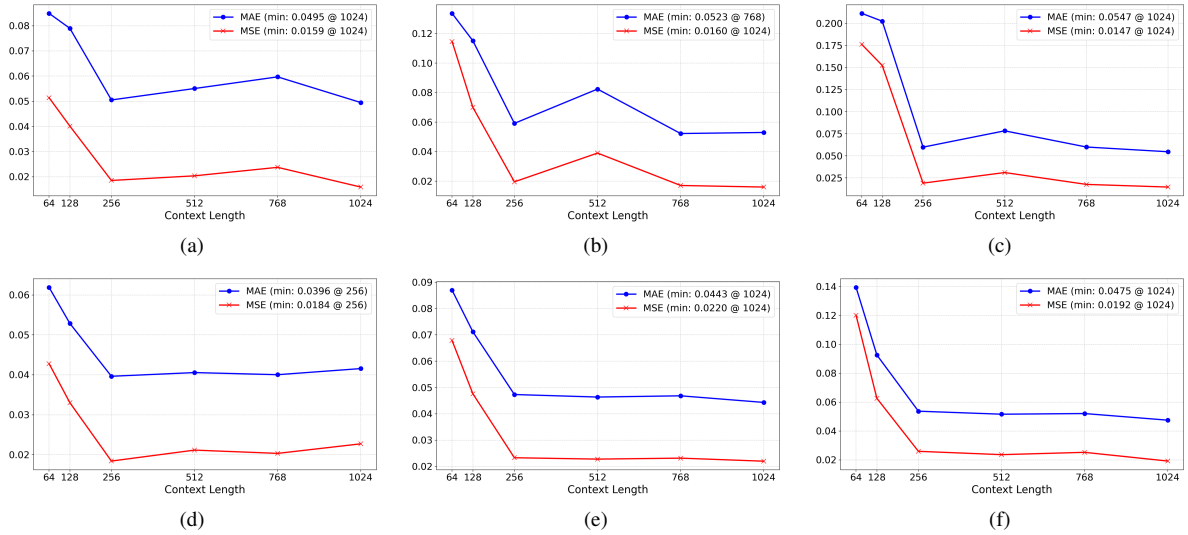

Figure 3. MSE and MAE of forecast performance with context lengths $c \in \{64, 128, 256, 512, 768, 1024\}$ in the MRO validation dataset. Top row (a-c): MOIRAI (Small). Bottom row (d-f): Chronos-T5 (Small). Columns show prediction lengths $p \in \{10, 20, 40\}$ respectively.

Chronos. We also observe that lower MSE and MAE do not necessarily correspond to better anomaly detection performance. For example, Chronos achieves its lowest MSE at $(c{=}256, p{=}10)$, but its best F1-score is obtained at $(c{=}1024, p{=}10)$, where the forecasting error is slightly higher.

Figure 4 shows the forecast performance of Chronos for the first group of the MRO test set at $(c{=}256, p{=}10)$. Since Chronos processes each channel independently, it cannot leverage cross-channel dependencies. Channels 4 and 10 exhibit false positives immediately after the anomalous samples, as the forecast accuracy degrades when anomalies appear in the context window. This produces large forecast errors that persist beyond the anomaly window, complicating anomaly detection from the error sequence.

MOIRAI processes all channels simultaneously as a multivariate time series and leverages cross-channel dependencies. Figure 5 shows the forecast performance of MOIRAI for the first group of the MRO test set at $(c{=}512, p{=}10)$. All three anomalies are detected without any false positives. The multivariate processing prevents error propagation when an anomaly appears in the context window of one channel, resulting in a lower false positive rate compared to the univariate approach used by Chronos.

5. Conclusion

We evaluated four forecasting foundation models, MOIRAI, Toto, Chronos and TimesFM, on a Mars Reconnaissance Orbiter (MRO) dataset in both zero-shot and fine-tuned settings, addressing the underexplored use of forecasting foundation models for multivariate spacecraft telemetry anomaly detection.

In the zero-shot setting, performance is driven by context length, model capacity and multivariate processing capability. At short context lengths, all models copy recent input samples rather than learning the nominal behavior of the signal. This keeps forecast errors low in anomalous regions, making anomalies undetectable. Longer context lengths allow the models to learn the nominal behavior and produce large forecast errors when anomalies occur. Larger capacity models achieve higher F1-scores, with Toto (103M parameters, F1=0.752) and TimesFM (200M parameters, F1=0.695) outperforming MOIRAI (14M parameters, F1=0.626) and Chronos (46M parameters, F1=0.451). We also observed that multivariate models outperform univariate models of similar capacity.

We fine-tuned MOIRAI and Chronos on domain-specific data, with MOIRAI improving from F1=0.626 to 0.859 and Chronos from 0.451 to 0.624. These results show that fine-tuning a smaller model can match or exceed the zero-shot performance of a larger one. Fine-tuned MOIRAI (14M parameters, F1=0.859) outperforms zero-shot Toto (103M parameters, F1=0.752), and fine-tuned Chronos (46M parameters, F1=0.624) reaches performance comparable to zero-shot TimesFM (F1=0.695). We also observed that when an anomaly falls within the context window of a univariate model, it affects the forecast accuracy of that channel, generating false positives that persist beyond the anomalous region. Multivariate models prevent this by leveraging cross-channel dependencies.

Our results show that forecasting foundation models are viable for spacecraft anomaly detection. In the zero-shot setting, Toto achieves an F1-score of 0.752 without domain-specific training, and fine-tuned MOIRAI reaches

Table 3. Fine-tuned anomaly detection performance of MOIRAI (Small) and Chronos-T5 (Small) across context lengths $c \in \{128, 256, 512, 1024\}$ and prediction lengths $p \in \{10, 20, 40\}$. For each metric, bold values indicate the best result across all (c, p) configurations.

		$c = 128$			$c = 256$			$c = 512$			$c = 1024$		
Model	**Metric**	p=10	p=20	p=40	p=10	p=20	p=40	p=10	p=20	p=40	p=10	p=20	p=40
MOIRAI (Small) fine-tuned	Precision (↑)	0.714	0.535	0.249	0.831	0.871	0.764	0.844	0.775	0.800	0.833	0.838	**0.896**
	Recall (↑)	0.843	0.843	0.829	0.886	0.829	0.800	0.914	0.843	0.814	**0.938**	0.914	0.929
	F1 (↑)	0.650	0.501	0.232	0.752	0.729	0.638	0.781	0.653	0.648	0.786	0.771	**0.859**
Chronos-T5 (Small) fine-tuned	Precision (↑)	0.482	0.445	0.333	0.619	0.522	0.437	0.512	0.560	**0.659**	0.633	0.610	0.558
	Recall (↑)	0.857	0.886	0.886	0.886	0.867	0.914	0.910	0.900	0.929	0.895	**0.943**	0.900
	F1 (↑)	0.441	0.433	0.335	0.604	0.505	0.421	0.501	0.526	0.620	**0.624**	0.607	0.532

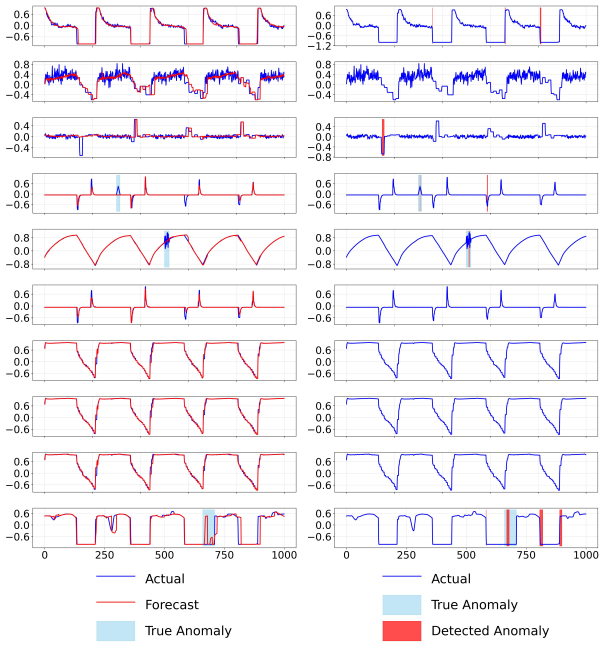

Figure 4. Forecast performance of Chronos for the ten channels in the first group of the MRO testing dataset for p=10 and c=256. The three anomalies are detected but there are several false positives.

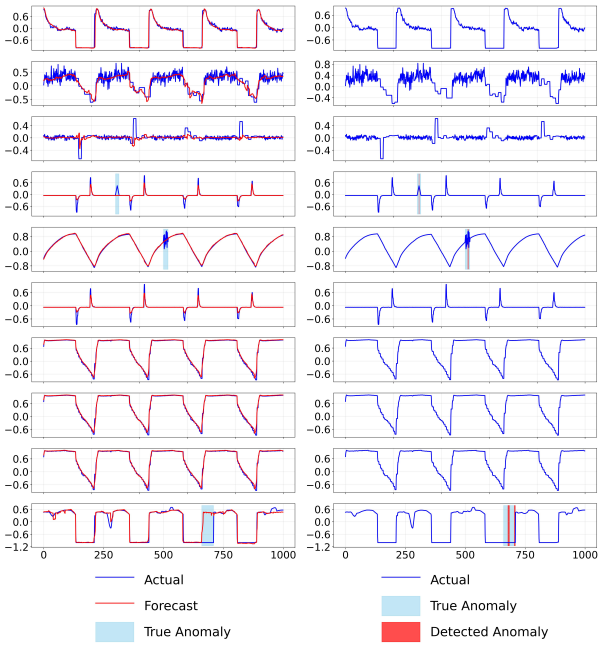

Figure 5. Forecast performance of MOIRAI for the ten channels in the first group of the MRO testing dataset for p=10 and c=512. The three anomalies are detected with no false positives.

an F1-score of 0.859. However, computational resources are an important consideration. Latency scales with model size, with Toto requiring 123s per 1,000 samples at $(c{=}1024, p{=}10)$ compared to 4.72s for MOIRAI. Fine-tuned MOIRAI offers the best balance between anomaly detection performance and latency among the evaluated foundation models, although traditional specialized models currently require lower computational resources for inference.

6. Acknowledgments

In this section we will add the acknowledgments.

References

[1] A. F. Ansari, L. Stella, and C. Turkmen et al. Chronos: Learning the Language of Time Series. *arXiv*, 2024. 1, 2, 3

[2] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant Risk Minimization. *arXiv*, 2020. 1

[3] S. Baireddy, S. Desai., J. Mathieson, R. Foster, M. Chan, M. Comer, and E. Delp. Spacecraft Time-Series Anomaly Detection Using Transfer Learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2021. Virtual. 1

[4] S. Baireddy, S. Desai, R. Foster, M. Chan, M. Comer, and E. Delp. Spacecraft Time-Series Online Anomaly Detection Using Deep Learning. *Proceedings of the IEEE Aerospace*

Conference, pages 1–9, 2023. Big Sky, Montana, United States. 1, 2

[5] V. Chandola, A. Banerjee, and V. Kumar. Anomaly Detection: A Survey. *ACM Computing Surveys*, 41(3):1–58, 2009. 1

[6] Z. Chen, D. Chen, X. Zhang, Z. Yuan, and X. Cheng. Learning Graph Structures with Transformer for Multivariate Time-Series Anomaly Detection in IoT. *IEEE Internet of Things Journal*, 9(12):9179–9189, 2021. 2

[7] B. Cohen, E. Khwaja, K. Wang, C. Masson, E. Ramé, Y. Doubli, and O. Abou-Amal. Toto: Time Series Optimized Transformer for Observability. *arXiv*, 2024. 1, 2

[8] A. Das, W. Kong, R. Sen, and Y. Zhou. A Decoder-Only Foundation Model for Time-Series Forecasting. *arXiv*, 2024. 1, 2, 3

[9] J. Graf, R. Zurek, R. Jones, H. Eisen, M. Johnston, B. Jai, and B. Mateer. An Overview of the Mars Reconnaissance Orbiter Mission. *Proceedings of the IEEE Aerospace Conference*, 2002. Big Sky, MT. 4

[10] K. Hundman, V.Constantinou, C.Laporte, I.Colwell, and T.Soderstrom. Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding. *Proceedings of the ACM SIGKDD international conference on knowledge discovery & data mining*, pages 387–395, 2018. London, United Kingdom. 1, 2, 4

[11] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, 2015. San Diego, CA. 6

[12] T. Li, S. Baireddy, M. Comer, E. Delp, S. Desai, R. Foster, and M. Chan. Multi-Channel Anomaly Detection for Spacecraft Time Series Using MAP Estimation. *IEEE Transactions on Aerospace and Electronic Systems*, pages 1–15, 2024. 2

[13] I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*, pages 1–15, 2017. Freiburg, Germany. 6

[14] A. Olive, S. Desai, R. Foster, M. Chan, M. Comer, and E. Delp. Spacecraft Multivariate Time Series Anomaly Detection in the Presence of Non-Anomalous Spikes. *Proceedings of the IEEE Aerospace Conference*, 2025. Big Sky, Montana, United States. 3

[15] G. R. Prakash, M. Comer, E. Delp, M. Chan, R. Foster, and S. Desai. A Graph Neural Network for Anomaly Detection in Multi-Channel Time Series Data. *Proceedings of IEEE Military Communications Conference*, 2025. Los Angeles, CA. 1

[16] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *The Journal of Machine Learning Research*, 21(140):5485–5551, 2020. 3

[17] T. Reiss, G. Kour, N. Zwerdling, A. Anaby-Tavor, and Y. Hoshen. From Zero to Hero: Cold-Start Anomaly Detection. *arXiv*, 2024. 2

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning Representations by Back-Propagating Errors. *Nature*, 323 (6088):533–536, 1986. 2

[19] R. Samsudin, A. Shabri, , and P. Saad. A Comparison of Time Series Forecasting Using Support Vector Machine and Artificial Neural Network Model. *Journal of Applied Sciences*, 10(11):950–958, 2010. 2

[20] A. Sánchez-Ferrera, B. Calvo, and J. Lozano. A Review on Self-Supervised Learning in Time Series Anomaly Detection: Recent Advances and Open Challenges. *ACM Computing Surveys*, 58(5):1–35, 2025. 2

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, A Gomez L. Jones, and Łukasz Kaiser. Attention is All you Need. *Advances in Neural Information Processing Systems*, 2017. Long Beach, CA. 2

[22] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo. Unified Training of Universal Time Series Forecasting Transformers. *arXiv*, 2024. 1, 2

[23] R. Wu and E. J. Keogh. Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress. *arXiv preprint arXiv:2009.13807*, pages 1–17, 2021. Riverside, California, United States. 4

[24] A. Zeng, M. Chen, L. Zhang, and Q. Xu. Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. Washington DC. 3