Time Series Challenge

Improving Time Series Prediction via Auxiliary Data and Cross-Signal Context

WARA AI TRICS

September 15, 2026

1 Motivation

WARA AI TRICS aims to bring real-world time series problems to the academic community, recognizing their cross-domain relevance in modern industrial, telecommunication, and financial systems. To foster methodological innovation in this space, we present the first WARA AI TRICS Open Challenge for PhD students. This challenge is open to PhD student teams within WASP, with all required computational infrastructure and compute resources provided by WARA AI TRICS. Participating teams will have a 3-month development period, spanning from **[Start Date]** to **[End Date]**.

While time series modeling may fall outside some participants' primary research domains, tackling multimodal, cross-signal temporal problems provides a direct pathway to high-impact research. Modern machine learning research is rapidly transitioning from isolated, single-variable forecasting toward contextual architectures that integrate heterogeneous side information, spatial topographies, and unstructured logs. This shift is reflected across premier learning theory and domain-specific venues:

- **Top-Tier ML Venues:** Specialized tracks and workshops - such as the upcoming NeurIPS 2026 Workshops in Paris, as well as recent events like the NeurIPS Workshop on Time Series in the Age of Large Models and the ICML Time Series Workshop - frequently highlight advancements in temporal foundation models, multimodal alignment, and non-stationary dynamics.
- **Domain-Specific AI Venues:** For financial and economic applications, the flagship ACM International Conference on AI in Finance (ICAIF) actively solicits contributions on regime-aware modeling, extreme risk evaluation, and cross-asset side information. We recommend the participants to consider 3rd Workshop on LLMs and Generative AI for Finance (AI for Finance Summit) as a potential venue for submission.

As all challenge datasets are open access, participants are strongly encouraged to formalize their methodologies and empirical findings into workshop or main-track submissions. Participation equips researchers with novel benchmark datasets, and industry-relevant pre-formulated problem setups - providing the foundational empirical validation required for competitive conference papers.

Ultimately, WARA AI TRICS aims to build an active collaborative community bringing together academic researchers and industry practitioners around complex temporal systems. The challenge aims to establish direct bridges between academic researchers and industry partners including Ericsson, SEB, and IPercept.

2 Challenge Overview

Traditional time series forecasting models often operate primarily on the historical observations of a target signal. However, in complex real-world dynamics, a single time series rarely contains enough information to explain all structural shifts, tail events, or localized anomalies. A sudden deviation, peak, or drop in a target signal is frequently driven by underlying contextual regimes, interactions with related signals, or external multimodal information.

This challenge investigates whether incorporating such auxiliary context can improve downstream time series prediction. Participants are encouraged to develop frameworks that combine primary target signals with heterogeneous contextual information, including related time series, cross-signal interactions, and multimodal data. The objective is to evaluate whether and to what extent this additional context improves predictive performance, particularly under non-linear dynamics, extreme events, regime changes, and other difficult prediction regimes.

To demonstrate the applicability and generality of this paradigm, participants will tackle solutions across three distinct, real-world domain case studies:

1. **Quantitative Finance (SEB):** Modeling structural asset price deviations using an auxiliary regime variates dataset containing macroeconomics, rates, and volatility signals.
   Contact: Alireza Javid (alireza.javid@seb.se).

2. **Telecom Systems (Ericsson):** Forecasting critical network Key Performance Indicator (KPI) behavior by capturing spatial, temporal, and contextual interactions across neighboring cells, concurrent KPIs, alarms, logs, and topology.
   Contact: Hannes Larsson (hannes.larsson@ericsson.com).

3. **Industrial Machines Monitoring (IPercept):** Anomaly detection in repetitive industrial robot motion using time series foundation models. A robotic system executes a structured motion pattern under controlled, repeatable conditions. Participants build models that learn normal motion from a large unlabeled corpus and detect mechanical degradation patterns in a held-out evaluation window.
   Contact: Peter Troll (peter@ipercept.io).

The three case studies share a common methodological objective but differ in their domain, target signal, auxiliary context, and downstream prediction task. The table 1 presents the three cases at a glance.

Table 1: Overview of the three challenge case studies.

Case	Domain	Primary Signal	Auxiliary Context	Main Task
SEB	Quantitative Finance	Asset prices	Macroeconomic and market regime variables	Max deviation prediction
Ericsson	Telecom	Network KPIs	Cells, alarms, topology, logs	KPI forecasting and network state prediction
IPercept	Industrial Machines	6-channel IMU	Session metadata	Anomaly detection (AUROC)

3 Core Theme & General Problem Formulation

3.1 Auxiliary Context Framework

Let $S_t^{(i)} \in \mathbb{R}$ represent the primary target time series for sample i at time step t. Given a lookback window of length M, the core target signal sequence is defined as $X_i = \{S_t^{(i)}\}_{t=T-M}^{T}$.

In traditional setups, models map $X_i \to Y_i$, where Y_i is a downstream target defined over a lookforward horizon N. Under our "Auxiliary Context Framework", model $\mathcal{M}$ is provided with an auxiliary context $\mathcal{A}_i(t)$, which may include heterogeneous time series, spatial graph topologies, or unstructured multimodal logs/events available up to anchor time T:

$$\hat{Y}_i = \mathcal{M}\left(X_i, \mathcal{A}_i(t)\right) \tag{1}$$

3.2 Common Problem Formulation

The common problem across all case studies is to evaluate whether incorporating auxiliary context can improve downstream prediction compared with models operating only on the primary target signal. Let the baseline model, provided by the challenge organizers for each case study, be defined as

$$\hat{Y}_i^{\text{base}} = \mathcal{M}_{\text{base}}(X_i). \tag{2}$$

The primary objective is to assess whether and to what extent the inclusion of $\mathcal{A}_i$ improves predictive performance over the corresponding baseline, particularly under extreme events, regime changes, and off-mean dynamics. The case studies in the next section provide distinct real-world settings in which participants can investigate this question.

3.3 Recommended Resources

Participants may refer to existing open-source time series forecasting and foundation model implementations as starting points for their solutions. The following resources are provided as optional implementation references and are not intended to prescribe a particular modeling approach. Participants are free to use alternative architectures and hybrid frameworks subject to the challenge rules.

- **TSMixer:** An all-MLP multivariate forecasting architecture that mixes information across both temporal and feature dimensions. It provides a simple and efficient starting point for exploiting cross-variate and auxiliary information in time series forecasting [Chen et al.(2023)].
- **TimesFM:** Google Research's time series foundation model, including implementations for zero-shot forecasting, fine-tuning, and covariate-based forecasting [Sen et al.(2024)].
- **TimeMixer:** An open-source multiscale time series forecasting architecture with implementations and benchmark scripts for long- and short-term forecasting [Wang et al.(2024)].
- **Kairos:** An open-source, parameter-efficient time series foundation model designed for heterogeneous time series and zero-shot forecasting [Feng et al.(2025)].
- **Aurora:** A multimodal time series foundation model designed to incorporate information from multiple modalities for forecasting [Wu et al.(2026)].

4 Evaluation Framework

All challenge submissions will be evaluated on held-out data that is not available to participants for model training. The evaluation framework is designed to assess both overall predictive accuracy and the ability of models to capture difficult dynamics, including extreme events, regime changes, and off-mean behavior.

4.1 Leaderboard / Ranking

The challenge will maintain a separate leaderboard for each of the three case studies: **SEB**, **Ericsson**, and **IPercept**. Participants will be ranked independently for each case study according to its primary evaluation metric.

An **overall challenge ranking** across the three case-study leaderboards will also be provided for participants who submit valid results for all three case studies. Participants submitting results for only a subset of the case studies will remain eligible for the corresponding case-specific leaderboards but will not be included in the overall ranking.

4.2 Innovation Award

In addition to the performance-based rankings, the challenge will include an **Innovation Award** to recognize particularly creative, novel, or promising approaches. The award is evaluated independently of both the case-specific and overall rankings, allowing participants to be recognized for a strong idea even if their solution does not achieve the highest predictive performance. The award will consider factors such as **novelty, technical creativity, quality of reasoning, and potential practical impact**.

4.3 Primary and Secondary Metrics

Each case study will define a **primary metric** that determines the principal leaderboard ranking and, where appropriate, one or more **secondary metrics** to provide additional insight into model behavior on specific subsets of interest, such as extreme values, tail events, or other difficult prediction regimes. The exact metrics and their calculation procedures are specified within each case study. If two or more submissions achieve identical performance according to the primary leaderboard metric, secondary metrics will be used to determine the ranking.

4.4 Live / Out-of-Sample Evaluation

Where applicable, selected submissions will be evaluated on a strictly out-of-sample period that is not available to participants during model development. This evaluation is intended to measure the ability of submitted models to generalize beyond the historical data used during development and to assess robustness under previously unseen temporal conditions. For the SEB case study, the Live Out-of-Sample Forward Test uses a fixed evaluation period immediately following the official challenge submission deadline.

5 Submission Guidelines & Rules

5.1 Submission Package

Each participating team is required to submit the following materials:

- **Technical Paper:** A technical paper of up to 4 pages, following an IEEE/ACM style format, describing the proposed methodology, model architecture, feature engineering and auxiliary data integration framework, as well as the achieved performance.
- **Code Repository:** A fully containerized code repository to reproduce the submitted results.

The exact prediction format and any additional case-specific submission requirements will be specified separately for each case study if needed.

5.2 Data Leakage

Participants must ensure that all feature engineering, preprocessing, model training, and prediction pipelines strictly respect the temporal structure of the corresponding task. For a prediction anchor time T, all information used to generate the model input must be available at or before T. The use of future information from the lookforward period for feature construction, preprocessing, model selection, or training is strictly prohibited. Any use of lookforward information that results in data leakage will constitute a violation of the challenge rules and may result in immediate disqualification.

5.3 Data Usage Policy

Participants may use only the datasets and data sources explicitly permitted for the corresponding case study. For the SEB case study, participants **are permitted** to use any external data source for the prediction task beyond the provided Auxiliary Regime Variates Dataset.

5.4 Compute Resources

Details of compute resources available to the participants are TBD.

6 Case Studies

6.1 SEB: Quantitative Modeling of Asset Price Deviations

6.1.1 Task Description

In quantitative finance, modeling asset price dynamics relative to historical price trends is critical for both risk management and opportunistic alpha generation. This case study tasks participants with predicting the maximum relative structural deviation (the "Max Deviation %") of an asset's price from its 95% statistical confidence interval (CI) over a lookforward horizon.

6.1.2 Historical Trend & Boundary Definition

Let $P(t)$ be the price of a financial asset at a discrete trading day t. For any given prediction anchor day T, we consider sample i to be comprised of X_i, a historical lookback window of $P(t)$ with size $M = 252$, and its corresponding target variable Y_i to be defined later based on a lookforward window of size $N = 21$. To define a historical trend, we establish the asset's baseline trajectory by fitting an ordinary least squares (OLS) linear regression model over the training interval $t \in [T - M, T]$:

$$L(t) = \alpha + \beta \cdot t \tag{3}$$

The standard deviation of the residuals from this linear fit is denoted as σ. The 95% confidence interval boundaries for the baseline trend are defined as:

$$CI_{\text{upper}}(t) = L(t) + 1.96\sigma, \quad CI_{\text{lower}}(t) = L(t) - 1.96\sigma \tag{4}$$

6.1.3 Target Variable (Max Deviation %)

Let N denote the lookforward forecast horizon ($t \in [T + 1, T + N]$). The objective function measures the maximum percentage deviation outside the established confidence interval boundaries, extrapolated through the lookforward window. We define the target variable, Y_i (Max Deviation %), at anchor day T for sample i as follows:

1. **Upper Violation:** If the price exceeds the upper boundary during the lookforward window, the maximum upper deviation is:

$$\Delta_{\text{upper}} = \max_{t \in [T+1, T+N]} \left(\frac{P(t) - CI_{\text{upper}}(t)}{CI_{\text{upper}}(t)} \right) \tag{5}$$

2. **Lower Violation:** If the price drops below the lower boundary during the lookforward window, the maximum lower deviation is:

$$\Delta_{\text{lower}} = \max_{t \in [T+1, T+N]} \left(\frac{CI_{\text{lower}}(t) - P(t)}{CI_{\text{lower}}(t)} \right) \tag{6}$$

The final target variable Y_i is defined piecewise based on boundary breaches:

$$Y_i = \begin{cases} \Delta_{\text{upper}} \times 100 & \text{if } \exists t \text{ s.t. } P(t) > CI_{\text{upper}}(t) \text{ and } \Delta_{\text{upper}} > \Delta_{\text{lower}} \\ 0 & \text{if } CI_{\text{lower}}(t) \leq P(t) \leq CI_{\text{upper}}(t) \quad \forall t \in [T + 1, T + N] \\ -\Delta_{\text{lower}} \times 100 & \text{if } \exists t \text{ s.t. } P(t) < CI_{\text{lower}}(t) \text{ and } \Delta_{\text{lower}} > \Delta_{\text{upper}} \end{cases} \tag{7}$$

6.1.4 Dataset Configuration

The target variable Y_i is bounded dynamically by the asset classes, economical regime at anchor day T, and macro economical factors such inflation rates, Treasury rates, Gold price, etc. Therefore, in order to predict the maximum deviation, it can be beneficial to incorporate such auxiliary variates as side information for the downstream regression task. You can read more on how to construct example economical regimes in Appendix A. Two distinct datasets are provided to the participants:

1. **Core Stock Universe Dataset:** A comprehensive dataset consisting of daily OHLCV historical asset data for US equities, spanning from `1991-01-01` to `2020-12-31`.

2. **Auxiliary Regime Variates Dataset:** A heterogeneous macroeconomic and cross-asset dataset, curated to assist in market regime classification and macro signal extraction. The granular descriptions of the provided auxiliary features are detailed in Appendix B.

Data Policy: Participants **are permitted** to incorporate additional fast-moving data sources beyond those provided above—such as options-implied volatility and skew metrics, industry ETF volatility benchmarks (e.g., `VXFXI`, `VXGDX`), or any other custom built indices (e.g., `CESI`, `FSI`, `EPU`, `GPR`) or technical indicators (e.g., `RSI`, `MACD`). However, all feature engineering and external data ingestion must ensure zero data leakage. Any use of future or lookforward information within the feature engineering pipeline will result in immediate disqualification.

6.1.5 Evaluation Tracks

Participants will develop solutions that predict the target variable Y_i. The submitted solutions will be evaluated in two distinct timelines and market regimes to test model versatility:

1. **Historical High-Volatility Backtest (Track A):** This evaluation spans the historical period from `2020-01-01` to `2020-12-31` on a daily rolling-window basis. The evaluation dataset comprises 500 randomly selected US stocks, capturing severe market shocks, rapid drawdowns, and subsequent volatility regimes.

2. **Live Out-of-Sample Forward Test (Track B):** This evaluation encompasses a fixed evaluation period of exactly one month immediately following the official challenge submission deadline on a daily rolling-window basis. Predictions will be evaluated and averaged over 500 randomly selected US stocks to assess the model's generalizability.

To maintain the integrity of the Track A evaluation, the historical period from `2020-01-01` to `2020-12-31` strictly **cannot** be used as training data for fitting model weights; however, it is permissible to use this period as a validation set for hyperparameter tuning. The participants are required to provide trained model checkpoints as described in the repository:
Git Repository: `https://github.com/banque-nouveau/WARA_SEBx_TS_Challenge`

6.1.6 Evaluation Metric

The primary leaderboard ranking metric will be the standard **Root Mean Squared Error (RMSE)** calculated over the test set instances $\mathcal{D}_{\text{test}}$:

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i \in \mathcal{D}_{\text{test}}} (Y_i - \hat{Y}_i)^2} \tag{8}$$

Where Y_i is the true Max Deviation % value, and $\hat{Y}_i$ is the participant's model prediction. If two or more participants achieve the same RMSE, Tie-breaking criteria will utilize the **Mean Absolute Error (MAE)** calculated exclusively on non-zero boundary violation events ($Y_i \neq 0$). An example of how an accurate prediction looks like is shown in Appendix C.

6.2 Ericsson: Telecom Network State & KPI Dynamics

6.2.1 Background and Motivation

Modern telecom networks generate large volumes of operational data from radio access, core network, transport, services, configurations, alarms, logs, and customer-facing systems. Much of what operators want to predict or monitor is expressed as time series, such as traffic, latency, throughput, availability, or other network KPIs.

The difficulty is that the behavior of a KPI is often not explained by that signal alone. A peak or drop may be caused by interactions with other KPIs, nearby cells, configuration changes, alarms, logs, topology, tickets, or even visual and map-based information. This makes telecom a natural setting for both multivariate modeling, where several related time series are considered together, and multimodal modeling, where time series are combined with other forms of context.

Recent time-series foundation models show promising generalization, but they mostly operate on isolated signals. The central motivation for this challenge is to understand how such models can be extended, adapted, or combined with other pre-trained models so that they capture a richer view of network state and perform better on the situations where telecom predictions matter most.

6.2.2 Research Challenge

How can foundation models move from understanding individual time series to understanding network state through multiple modalities and interacting signals?

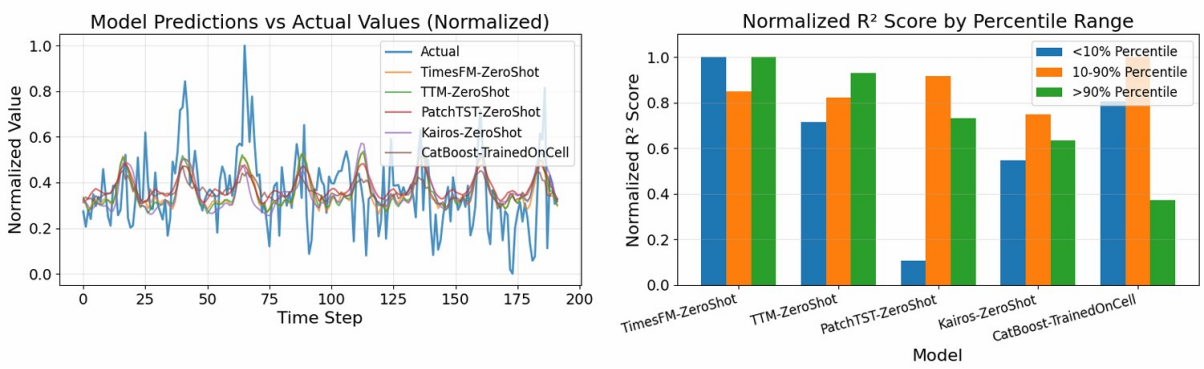

Figure 1: Visual representation of several time-series foundation models on KPI forecasting task

The figure shows a representative KPI forecasting task where several time-series foundation models are evaluated in zero-shot mode and compared with a CatBoost model trained for the specific KPI. The models capture the average signal behavior reasonably well, but struggle with common peaks and valleys away from the mean. These are not necessarily extreme outliers, but they are often the situations where accurate predictions matter most, for example when service-level agreements may be at risk.

Although the example focuses on forecasting, the same limitation is relevant for anomaly detection, root-cause analysis, and network state estimation. A likely reason is that a single KPI does not contain enough information to explain all relevant variations. Better performance may require additional context from related time series or other modalities such as logs, alarms, tickets, topology information, documentation, or image-based representations.

The challenge can be studied along two main dimensions. First, it can explore how models learn representations across different modalities, such as combining time-series KPIs with logs, alarms, tickets, topology information, or other contextual data. Second, it can study how models learn representations across multiple related time series, for example interacting KPIs,

neighboring cells, or cross-layer measurements. These directions can be approached either by combining and adapting existing pre-trained foundation models, or by training models that learn shared multimodal or multivariate representations directly.

6.2.3 Dataset Configuration

The participants can use telecom observability datasets that combine KPI time series with relevant contextual information, such as alarms, logs, topology, or configuration data. Where useful, public multimodal or multivariate benchmarks may also be included to test whether the methods generalize beyond telecom-specific settings.

6.2.4 Evaluation Track A: Covariate-Enhanced Forecasting

Input: Historical time series (168 hours context) + past and known-future covariates
Predict: Next 24 hours of the target variable
Evaluation:

- Primary: MAE, MASE (scale-independent)
- Secondary: CRPS (if probabilistic), coverage of prediction intervals

Tiered Evaluation (must report all tiers)

Tier	Input
0	Target only (univariate FM)
1	+ Calendar features (hour, day-of-week, holiday)
2	+ Weather covariates (temperature, precipitation)
3	+ All available covariates

Baselines

Method	Electricity MAE (kW)	Uber MAE (pickups/hr)
Chronos-2 (univariate, zero-shot)	12.65	22.1
CatBoost (univariate, lag features)	9.36	28.1

6.2.5 Dataset A1: Electricity (Portugal, 2011–2014)

Source	HuggingFace: `monash_tsf/electricity_hourly`
Scale	370 clients, hourly resolution
Time range	2011-01-01 to 2014-12-31
Target	Per-client kW consumption
License	Open

Provided covariates

Covariate	Type	Source
Temperature (°C)	Known future	NOAA Lisbon station, hourly
Relative humidity (%)	Known future	NOAA Lisbon station, hourly
Hour of day	Known future	Derived
Day of week	Known future	Derived
Portuguese public holidays	Known future	Public calendar

Participants may source additional covariates (e.g., electricity spot price, solar irradiance, school calendars) but must declare them.

Data split

Training	Validation	Evaluation
All data before Nov 1, 2012	Oct 2012	Undisclosed window(s) after training cutoff

6.2.6 Dataset A2: Uber TLC (NYC, Apr–Sep 2014)

Source	NYC Taxi & Limousine Commission / FiveThirtyEight
Scale	4,391 hours × 99 spatial zones, hourly
Time range	2014-04-01 to 2014-09-30
Target	Per-zone pickup count
License	Open

Provided covariates

Covariate	Type	Source
Temperature (°C)	Known future	NOAA JFK station, hourly
Precipitation (mm)	Known future	NOAA JFK station, hourly
Wind speed (m/s)	Known future	NOAA JFK station, hourly
Hour of day	Known future	Derived
Day of week	Known future	Derived

Participants may source additional covariates (e.g., NYC event calendars, subway disruptions, sports schedules) but must declare them.

Data split

Training	Validation	Evaluation
All data before Aug 15, 2014	Aug 1–14	Undisclosed window(s) after training cutoff

6.2.7 Track A Code Requirements

Code must accept `--train-end` and `--test-start` / `--test-end` arguments. Organizers evaluate on window(s) of their choosing—participants do not know which period is tested.

Scoring: $\text{Score}_A = 0.5 \times \text{rank}(\text{Electricity\_MAE}) + 0.5 \times \text{rank}(\text{Uber\_MAE})$

Bonus for covariate improvement: $(\text{MAE}_{\text{tier0}} - \text{MAE}_{\text{best\_tier}})/\text{MAE}_{\text{tier0}}$

6.2.8 Track B: Multi-Modal Anomaly Detection & Root Cause Analysis

Input: 18-channel multivariate time series (128 steps, 10 Hz) + optionally text and metadata

Tasks:

- Anomaly Detection—binary classification + temporal localization
- Root Cause Analysis—identify cause from 11 anomaly types
- Q&A (optional)—answer questions requiring joint time-series + text reasoning

Evaluation:

- Anomaly Detection: F1 (primary), Precision, Recall
- Root Cause: Accuracy (primary), Top-3 accuracy
- Q&A: MAE (numerical), Accuracy (categorical)

Tiered Evaluation (must report all tiers)

Tier	Input
0	Time series only (z-normalized)
1	Time series with absolute scale
2	+ Text descriptions
3	+ Troubleshooting tickets + metadata

Baselines (zero/few-shot, from Feng et al. 2026)

Model	Anomaly F1	RCA Acc
Mantis	0.800	0.590
Toto	0.615	0.848
Moment	0.397	0.550

6.2.9 Dataset B1: TelecomTS (5G Testbed, 2024)

Source	HuggingFace: `AliMaatouk/TelecomTS`
Scale	32,000 samples × 18 KPI channels × 128 timesteps (10 Hz)
Duration per sample	12.8 seconds
License	Open Database License
Anomaly types	11 categories

Provided covariates (included in dataset)

Covariate	Type	Content
Text description	Contextual	Natural language summary of KPI behavior per sample
Troubleshooting tickets	Contextual	Diagnosis narratives for anomalous samples
Zone label	Static metadata	A, B, C, or In motion
Application type	Static metadata	File, YouTube, Twitch
Congestion flag	Static metadata	Yes/No
Mobility flag	Static metadata	Static/Mobile

Data split

Temporal split at January 2028. Samples are ordered by timestamp; all data before the cutoff is training, all data after is evaluation. All 11 anomaly types appear in both splits, but training contains very few examples of each non-Jamming type (few-shot setting). This mirrors real deployment: a model trained on limited historical anomalies must detect them reliably in future data.

6.2.10 Track B Code Requirements

Code must accept a `--test-ids` argument.

Scoring: $\text{Score}_B = 0.6 \times F1_{\text{anomaly}} + 0.4 \times \text{Accuracy}_{\text{RCA}}$

6.2.11 Evaluation Metrics

Submissions are evaluated on held-out test splits prioritizing both global error and dynamic tail performance:

- **Primary Metric:** Root Mean Squared Error (RMSE) across overall test instances.
- **Tail / Extreme Performance (Segmented R^2 / MAE):**
- The models are benchmarked using normalized R^2 scores binned across percentile ranges ($<$ 10th, 10th–90th, and $>$ 90th percentiles) to enforce accurate forecasting of critical peaks and valleys.

6.3 IPercept: Anomaly Detection in Repetitive Industrial Motion

6.3.1 Background and Motivation

Many industrial systems execute the same motion pattern continuously under controlled conditions. When the mechanical state of the system changes, the motion signal changes with it. In practice these changes are subtle: they are small relative to the natural cycle-to-cycle variation of the machine itself; they may alter the structure of the signal without changing its overall energy.

6.4 IPercept: Anomaly Detection in Repetitive Industrial Motion

6.4.1 Background and Motivation

Many industrial systems execute the same motion pattern continuously under controlled conditions. When the mechanical state of the system changes, the motion signal changes with it. In practice these changes are subtle: they are small relative to the natural cycle-to-cycle variation of the machine itself, they may alter the structure of the signal without changing its overall energy, and they often develop gradually rather than appearing abruptly. Detectors based on amplitude thresholds or simple novelty scores tend to fail on all three counts.

This case study provides a large corpus of real multi-channel IMU data — accelerometer and gyroscope axes — from an industrial robotic system executing a structured, repeatable motion sequence across many recording sessions. The corpus is unlabeled. The task is to learn a model of normal motion and use it to identify mechanically meaningful deviations in a held-out evaluation window.

6.4.2 Dataset Configuration

Sensor	Multi-channel IMU: accelerometer and gyroscope axes
Scale	Many recording sessions over an extended period
Structure	Repeating motion cycles

Recordings are provided as independent segments and should not be assumed contiguous. Channels are not equally informative; identifying which carry usable information is part of the task.

6.4.3 Training Data and Contamination

The training corpus is unlabeled and predominantly, but not entirely, representative of normal operation. An unspecified small fraction of training cycles deviate from normal behaviour, and their positions are not disclosed. Methods that treat the entire training corpus as clean may therefore be biased by the deviations it contains. Robustness to this contamination is an explicit part of the challenge.

A labeled validation split is provided so that participants can develop and tune locally. Evaluation-window labels are withheld.

6.4.4 Evaluation Window

The held-out evaluation window contains sections where the motion signal departs from the normal pattern in physically motivated ways, interleaved with healthy sections. Several deviations are constructed to leave the overall signal energy of the affected cycles unchanged, so that they cannot be recovered from amplitude or energy statistics alone. The window also contains benign events that are unusual but mechanically harmless; these must not be flagged. The nature, timing and severity of all events are withheld and revealed only at evaluation.

6.4.5 Evaluation Tracks

Track A — TSFM-based anomaly detection. Apply a pretrained time series foundation model as a reconstruction-based or embedding-based anomaly detector. Fine-tuning on the provided corpus is permitted.

Track B — Cross-session generalisation. Train on a designated subset of recording sessions and evaluate on held-out sessions, testing robustness to session-level distribution shift. Session identity is available as a categorical label.

6.4.6 Evaluation Metric

- **Primary:** AUROC over the full evaluation window. Threshold-free and appropriate to the unlabeled training setting.
- **Secondary:** Average Precision (AP), used as tiebreaker.

A mean-template reconstruction baseline is released as starter code. Deviation severities are calibrated so that this baseline performs only moderately above chance; substantially exceeding it requires modelling the structure of the motion cycle rather than its magnitude.

6.4.7 Submission Format

Participants submit a CSV file with one row per cycle in the evaluation window, containing the cycle index and an anomaly score. Higher scores indicate greater deviation from the learned normal pattern.

References

[Chen et al.(2023)] Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O. Arik, and Tomas Pfister. 2023. TSMixer: An All-MLP Architecture for Time Series Forecasting. *Transactions on Machine Learning Research* (2023). `https://github.com/google-research/google-research/tree/master/tsmixer`

[Feng et al.(2025)] Kun Feng, Shaocheng Lan, Yuchen Fang, Wenchao He, Sihan Lu, Shuqi Gu, Lintao Ma, Xingyu Lu, and Kan Ren. 2025. Kairos: Toward Adaptive and Parameter-Efficient Time Series Foundation Models. *arXiv preprint arXiv:2509.25826* (2025). `https://github.com/foundation-model-research/Kairos`

[Sen et al.(2024)] Rajat Sen, Yichen Zhou, et al. 2024. A Decoder-Only Foundation Model for Time-Series Forecasting. In *International Conference on Machine Learning*. `https://github.com/google-research/timesfm`

[Wang et al.(2024)] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. 2024. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. In *International Conference on Learning Representations*. `https://github.com/kwuking/TimeMixer`

[Wu et al.(2026)] Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Chenjuan Guo. 2026. Aurora: Towards Universal Generative Multimodal Time Series Forecasting. In *International Conference on Learning Representations*. `https://github.com/decisionintelligence/Aurora`

A SEB Economic Regime Definitions

A.1 Theoretical Paradigm

In institutional asset management, regime identification is critical for risk awareness and strategic allocation. For portfolio managers, the main purpose is to create a "common language" for discussing the investment environment. A regime framework should help organize evidence, compare signals across horizons, and explain why certain portfolio risks may be attractive in the current regime. Any regime model rests on a few foundational core assumptions:

- **Persistency:** Regimes must last long enough to matter for portfolio decision-making.
- **Classification Uncertainty:** Regimes should generally be expressed as continuous probabilities or statistical confidence levels rather than as hard deterministic labels.
- **Asset Behavior Divergence:** Structural distributions of asset returns, volatility, cross-asset correlations, or drawdown risks should vary meaningfully across distinct regimes.

A.2 Taxonomy of Main Regime Models

There are three broad quantitative perspectives that are useful for regime modeling. They function as complementary architectural layers rather than mutually exclusive methods, as formalized in Table 2.

Table 2: Taxonomy and Structural Characteristics of Core Regime Frameworks

Model Type	Typical Inputs	Strengths	Limitations
Economic Regimes	Growth, inflation, policy, fiscal conditions, labor markets	Grounded in economic theory, highly interpretable, used for long-term strategy	Macro data are often lagged, heavily revised, and slow to capture structural turning points
Implicit/Market Regimes	Trends, credit spreads, yield curves, asset correlations	Highly timely, reflects investor behavior, closer to tradable conditions	Inherently noisy; prone to false signals, heavily affected by regulations and technical patterns
Risk Regimes	Volatility, drawdowns, liquidity stress, tail-risk indicators	Simple, robust, and relevant for risk awareness	Less nuanced; describes instant risk conditions without explaining underlying economic cause

A comprehensive, high-performing framework should ideally combine these heterogeneous perspectives. Economic regimes provide the narrative structure (the *"Why"*), market regimes capture real-time pricing and sentiment (the *"What"*), and risk regimes map the tail severity of the current environment (the *"How much"*).

B SEB Auxiliary Regime Dataset

The structural descriptions of the provided auxiliary features are detailed in Table 3.

Table 3: Auxiliary Regime Features and Macroeconomic Variates

Asset Name / Type	**Description**
`ISM Manufacturing PMI`	US factory demand pulse; the 50-line represents the expansion/contraction boundary. It serves as the most widely observed cyclical survey globally.
`BBG Cmdty ex-Prec Mtl`	Market-priced proxy for real industrial and energy demand, stripped of the monetary noise inherent in gold/silver.
`Gold Spot`	Safe-haven and currency debasement gauge; shifts primarily on fear and systemic liquidity rather than fundamentals.
`South Korea Exports YoY`	Earliest global-trade canary, situated upstream in supply chain; acts as a leading indicator for the PMI complex.
`China Manufacturing PMI`	An anchor of global industrial demand; China represents the marginal buyer for the majority of commodity basket.
`Ifo Pan-Germany Business Exp.`	Forward-looking European industrial read, serving as the export and capital expenditure (capex) bellwether for EU.
`Gold Spot $/Oz`	Spot pricing metric. The structural spread of commodities ex-gold vs. gold can be a market distress indicator.
`US 10y Real Rate`	Market-implied real yield tracker. Participants are encouraged to look at the continuous structural spread between the US Neutral Rate and the US 10-year Real Rate.
`Financial Index Generic`	Equity index futures (e.g., S&P 500, Euro Stoxx 50, FTSE 100). Exchange-traded contracts cash-settled daily via initial and variation margins rather than physical asset delivery.
`Fixed Income Index`	A rules-based benchmark tracking the cumulative performance of a specific basket of bonds (e.g., investment-grade or high-yield corporate debt segments).
`Return IDX`	A total return index measuring profit/loss from a currency carry trade (borrowing in low-yield currencies to invest in high-yield currencies), excluding pure spot exchange rate movements.
`Cross`	Currency cross rates reflecting relative values derived from respective exchange rates against a common reference currency (typically the US dollar) for trade or hedging metrics.
`Financial Commodity Generic`	Continuous price series derived from rolling front-month futures contracts, automatically switching to the next active expiry to avoid physical settlement traps.
`SWAP`	Interest rate swap agreements tracking fixed-to-floating rates (e.g., SONIA, EURIBOR). The fixed swap rate reflects market expectations of averaged future short-term rates.
`Equity Index`	Specialized equity sector indexes tracking the aggregate performance of industry-specific corporate baskets across broad geographic regions.
`Spot`	Real-time market exchange price of a currency pair or commodity asset (such as gold) designated for immediate delivery and physical settlement.

C SEB Example Linear Trend Prediction

This appendix provides a visual demonstration of an optimized model output compared against the true underlying asset price movement and the calculated statistical boundaries.

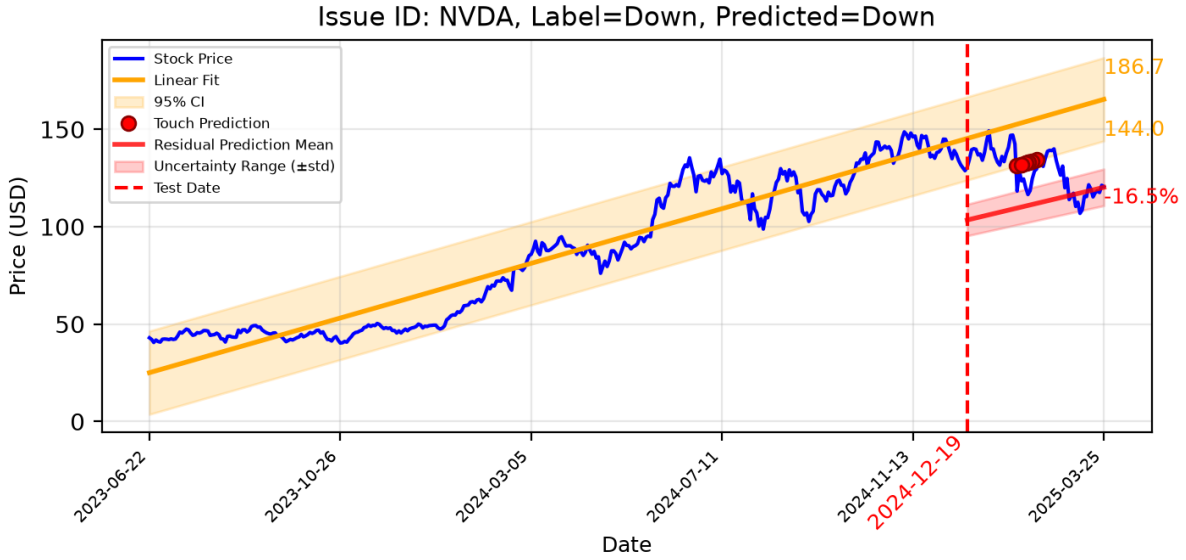

Figure 2: Visual representation of a highly accurate prediction during a significant linear trend boundary violation event.

C.1 Interpretation of an Accurate Prediction

As illustrated in Figure 2, the asset trend line $L(t)$ and its corresponding 95% confidence interval boundaries (CI_{upper} and CI_{lower}) are established deterministically over the historical lookback window M. The forecast challenge begins at the anchor day T and extends forward across a lookforward horizon of N days.

An ideal, high-performing machine learning model demonstrates success in this task by accurately predicting the following structural dynamics:

- **Directional Breakout Accuracy:** Successfully forecasting whether the asset will remain within the statistical bounds ($Y_T = 0$), or breakout structurally above the upper bound ($Y_T > 0$) or below the lower bound ($Y_T < 0$).
- **Magnitude Precision:** Capturing the exact peak percentage distance (Δ_{upper} or Δ_{lower}) separating the extreme price touch-point from the closest boundary threshold, rather than simply predicting a generalized volatility expansion.
- **Robustness to Intraday Noise:** Maintaining stability across consecutive anchor days, ensuring the model is robust against transient intraday noise. Rather than generating erratic, high-variance outputs on adjacent days, an optimal model captures the broader underlying regime to yield smooth, stable, and sequentially logical predictions.