FROM FORECASTS TO ANOMALY SCORES: REUSING FROZEN TIME SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen forecasting time series foundation models (TSFMs), when used for anomaly detection based on one-step forecast error (conventional scoring), have been reported to perform comparably to a model-free baseline that predicts by averaging the last three observations. We investigate whether their existing multi-step forecasts contain information discarded by conventional scoring. We propose **Forecast Disagreement Anomaly Scoring (FDAS)**, a post-hoc scoring layer that combines the one-step forecasting error with disagreement among forecasts for the same target time made at different forecast origins. FDAS smooths and normalises both signals using a reference segment presumed to contain normal observations, then takes their maximum as the anomaly score. FDAS reuses forecasts already produced by a TSFM, without fine-tuning it or using anomaly labels for the target series. On the full TSB-AD benchmarks, we evaluate FDAS on eight univariate and five multivariate TSFM configurations. Compared with conventional scoring, FDAS raises mean VUS-PR from 0.29–0.30 to 0.45–0.50 on the univariate benchmark and from 0.21–0.24 to 0.37–0.40 on the multivariate benchmark. These gains are statistically significant in every evaluated configuration. Ablations show that combining the smoothed forecasting-error and disagreement tracks outperforms either track alone. FDAS makes frozen forecasting TSFMs competitive with dedicated anomaly-detection TSFMs, indicating that their limitation lies largely in the scoring rule rather than the forecasts.

1 INTRODUCTION

Time series anomaly detection (TSAD) aims to identify time series observations that deviate from normal behaviour and to support the reliable operation of real-world systems (Blázquez-García et al., 2021; Schmidl et al., 2022). Anomalies are rare and their identification often requires domain expertise, thus labelled data are difficult to obtain (Goswami et al., 2023). This motivates methods that can detect anomalies without requiring anomaly labels for the target series (Zamanzadeh Darban et al., 2024).

Time series foundation models (TSFMs) are pre-trained on large-scale time series datasets and support downstream tasks such as forecasting, classification, and anomaly detection (Liang et al., 2024; Goswami et al., 2024). TSFMs are applicable to unseen time series, which makes them attractive for zero-shot TSAD. Prior work on zero-shot TSAD with TSFMs has followed two main approaches: developing TSFMs that directly support anomaly detection (Goswami et al., 2024; Ekambaram et al., 2026) and reusing forecasting TSFMs for anomaly detection (Liu & Paparrizos, 2024; Park & Yoon, 2026).

Prior evaluations have identified limitations in reusing forecasting TSFMs for anomaly detection. These studies used one-step forecasting error as the anomaly score, a practice we refer to as *conventional scoring* (Liu & Paparrizos, 2024; Zhu et al., 2026). Liu & Paparrizos (2024) reported that forecasting TSFMs under conventional scoring struggle to detect sequence anomalies. Zhu et al. (2026) further reported no significant difference in anomaly detection performance between forecasting TSFMs and *Last-3*, a simple baseline that predicts each value by averaging the three preceding observations, under conventional scoring. The same study identified the one-step forecast horizon as the primary limiting factor for detecting sequence anomalies. Forecasting TSFMs

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

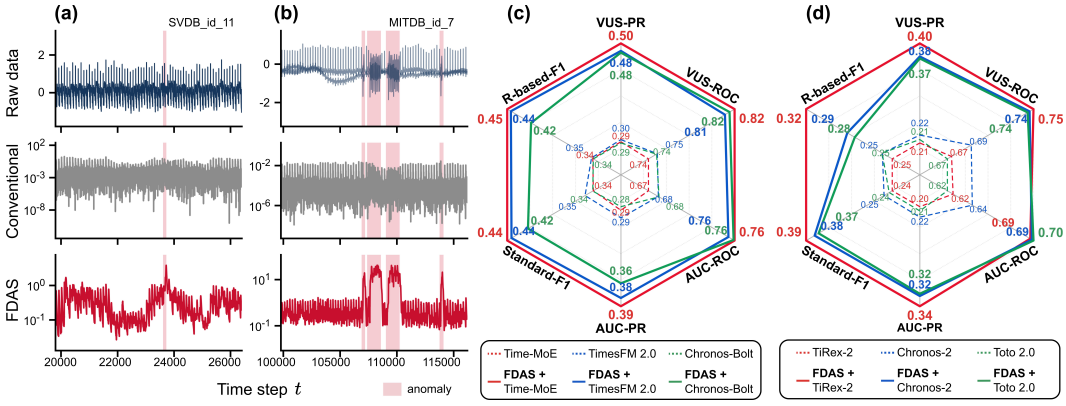


Figure 1: Conventional scoring and FDAS applied to the same backbone forecasts. (a) A univariate series with Chronos-Bolt and (b) a multivariate series with TiRex-2. Labelled anomalies are shaded. Six evaluation metrics under conventional scoring and FDAS for three backbones on (c) the univariate benchmark TSB-AD-U and (d) the multivariate benchmark TSB-AD-M.

produce multi-step forecasts, whereas conventional scoring uses only the first step and leaves longer-horizon predictions unused.

We therefore propose **Forecast Disagreement Anomaly Scoring (FDAS)**, a post-hoc scoring layer that leverages these unused predictions. FDAS uses two signals: the one-step forecasting error and disagreement among forecasts for the same target time step made at different forecast origins. The two signals are then smoothed and normalised, and their maximum is taken as the anomaly score. On the full TSB-AD benchmarks (Liu & Paparrizos, 2024), we evaluate FDAS using 13 TSFM backbone configurations: eight univariate and five multivariate. FDAS improves VUS-PR (volume under the precision–recall surface) over conventional scoring in every configuration and, unlike conventional scoring, significantly outperforms the Last-3 baseline and its smoothed variant (Section 5.2). Figure 1(c,d) shows improvements across six evaluation metrics for three representative backbones, one per model family, on the univariate and multivariate benchmarks. Our contributions are summarised as follows:

- **Forecasting TSFMs as competitive anomaly detectors.** We propose FDAS for frozen forecasting TSFMs. Its best configurations reach mean VUS-PR within 0.02 of the strongest compared TSFM that directly supports anomaly detection on each benchmark: 0.02 below on TSB-AD-U and 0.01 above on TSB-AD-M.
- **Anomaly scoring without additional model training or backbone inference.** FDAS reuses existing multi-step forecasts from either a univariate or a multivariate TSFM without fine-tuning or additional backbone inference. Its scoring layer has no learned parameters and requires no anomaly labels for the target series.
- **Greater gains from scoring than from backbone choice.** In our evaluation, replacing conventional scoring with FDAS raises mean VUS-PR by 0.15–0.21, while differences across backbones on the same benchmark remain within approximately 0.05 under either scoring rule. Ablations show gains from the disagreement track beyond smoothing, and its advantage over the forecasting-error track grows with average anomaly length.

2 Related Work

TSFM-based anomaly detection. Recent work has explored two approaches to TSFM-based anomaly detection: developing models that directly support anomaly detection and reusing forecasting TSFMs. In the former approach, TSFMs produce anomaly scores through dedicated mechanisms. With self-supervised pre-training, MOMENT (Goswami et al., 2024) scores anomalies using reconstruction error, and TSPulse (Ekambaram et al., 2026) derives scores from time-domain reconstruction, frequency-domain reconstruction, and forecasting heads. DADA (Shentu et al., 2025) is pre-trained on multi-domain data with injected anomalies and scores anomalies using the variance

across multiple reconstructions. TimeRCD (Lan et al., 2026) uses synthetic anomaly labels for supervised pre-training and scores anomalies using a learned anomaly head. In the latter approach, prior studies have explored reusing forecasting TSFMs by treating forecasting errors as anomaly scores (Liu & Paparrizos, 2024; Park & Yoon, 2026; Zhu et al., 2026). However, Zhu et al. (2026) reported no significant difference between forecasting TSFMs and the Last-3 baseline when anomaly scores were based on one-step forecasting error. Prior work has also used forecasts at multiple horizons for anomaly detection. Zhu et al. (2026) found that longer horizons could improve detection of longer anomalies but reduce sensitivity to shorter ones, and proposed taking the maximum of mean squared forecasting errors computed over several horizons from a single forecast origin. $\mathcal{W}_1$-ACAS (Martinez Gil et al., 2026) uses forecasts for the same target time step from different origins, converts each forecasting error into a conformal p-value, and takes their median. Both methods thus rely on forecasting-error scores, whereas our work additionally uses forecast disagreement.

Anomaly scoring in TSAD. In general, forecasting-based TSAD methods derive anomaly scores from forecasts in several ways. Deep learning-based forecasting methods commonly score each time step using the one-step forecasting error (Munir et al., 2019; Schmidl et al., 2022; Zamanzadeh Darban et al., 2024). Some methods additionally smooth the forecasting error using an exponentially weighted moving average before thresholding (Hundman et al., 2018). Beyond one-step scoring, several methods use information from multiple forecasts or longer forecast horizons. LSTM-AD (Malhotra et al., 2015) combines the forecasting errors for each time step from different forecast origins into an error vector and scores anomalies using a multivariate Gaussian fitted to error vectors from normal data. xLSTMAD-F (Faber et al., 2025) computes a single anomaly score over the entire forecast horizon, so the score is delayed until all observations within the horizon are available. Forecast disagreement and predictive uncertainty have also been used as anomaly signals. FATE (Kang et al., 2026) uses disagreement across an ensemble of forecasting models to detect anomaly precursors, and FOLD (Jhin & Park, 2026) uses input masking to obtain sensitivity signals and stochastic inference to estimate predictive uncertainty.

Like FATE, our work uses forecast disagreement as an anomaly signal, but derives it from a single forecasting TSFM rather than an ensemble of models. As in LSTM-AD and $\mathcal{W}_1$-ACAS, we use forecasts for the same target time step made at different forecast origins. Our work brings these ideas together by measuring disagreement across forecast origins within a single TSFM and combining it with the one-step forecasting error.

3 Background

Let $Y = (y_1, \ldots, y_T) \in \mathbb{R}^{T \times C}$ denote a time series with T time steps and C channels, where $y_t \in \mathbb{R}^C$ is the observation at time step t. The series is univariate when $C = 1$ and multivariate when $C > 1$, and $y_t^{(c)}$ denotes its value in channel c. The task is to assign an anomaly score $S_t \in \mathbb{R}_{\geq 0}$ to each time step t. Following the benchmark protocol (Liu & Paparrizos, 2024), each series Y includes a predefined train segment $Y_{\mathrm{tr}} = (y_1, \ldots, y_{t_{\mathrm{tr}}})$. FDAS uses this segment only for score normalisation while keeping the backbone frozen. We assume it contains no anomalies.

The backbone is a forecasting TSFM f_θ that takes a lookback window of the L most recent observations and produces forecasts for the next H time steps, where H is the forecast horizon. We call the last time step of the lookback window the *forecast origin*. The lookback window advances by one time step after each forecast, so the same target time step is forecast from different origins. For $i \in \{1, \ldots, H\}$, we write $\hat{y}_t^{(t-i)}$ for the *i-step-ahead* forecast of y_t made at origin $t - i$.

4 Proposed Method

FDAS overview. FDAS is a post-hoc scoring layer that uses the forecasts of a frozen forecasting TSFM to construct two anomaly signals, which we call *tracks*, and combines them into a single anomaly score (Figure 2). The first track is the disagreement among forecasts for the same target time step, and the second is the one-step forecasting error used by conventional scoring. FDAS smooths both tracks individually with an exponentially weighted moving average (EWMA) (Hundman et al., 2018), normalises them to a common scale, and takes their maximum as a per-channel score. For multivariate series, the maximum across channels gives the final score.

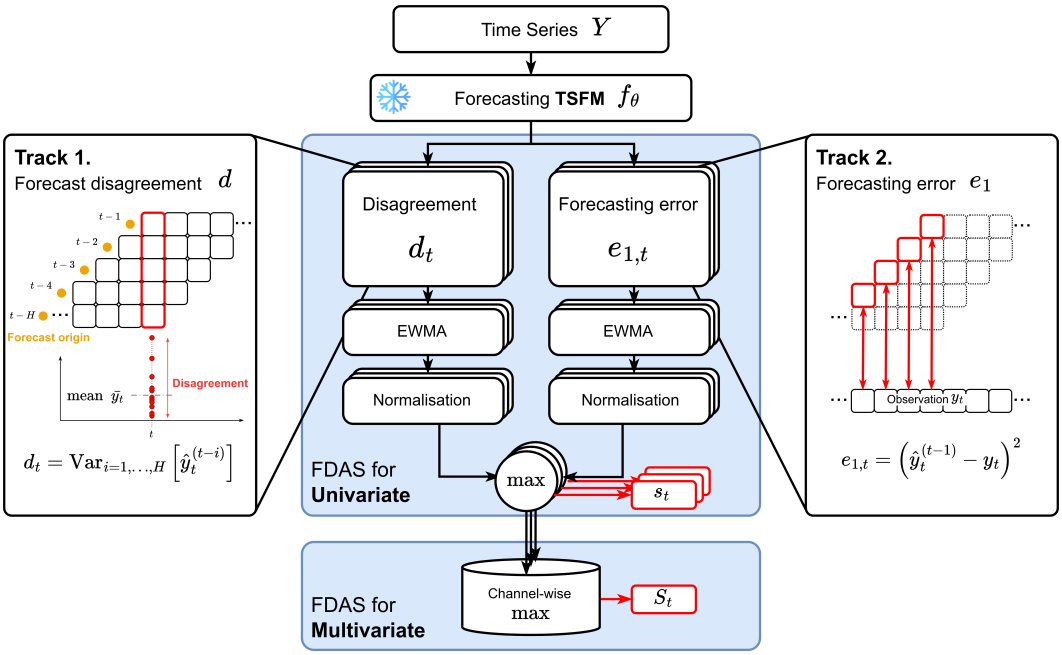


Figure 2: Overview of FDAS. Stacked outlines indicate per-channel processing.

4.1 Backbone assumptions

FDAS applies to frozen forecasting TSFMs that satisfy the following two assumptions:

- **A1 Single forecast sequence.** FDAS uses a single forecast sequence for each lookback window. For backbones with a quantile head (Ansari et al., 2024; 2025; Podest et al., 2026; Khwaja et al., 2026), we use the median forecast according to their documented procedures.
- **A2 Multi-step output per forward pass.** At each forecast origin, the backbone must produce an H-step forecast with $H \geq 2$ in a single forward pass.

4.2 Two tracks

Multi-origin forecasts. For a target time step t, the forecasts for y_t come from forecast origins $t-1, \ldots, t-H$ (Figure 2, Track 1). The lookback window ending at the earliest origin $t-H$ starts at $t-H-L+1$, so all H forecasts are available from $t_0 = L+H$. Let $f_\theta(\cdot)[i]$ denote the i-step-ahead forecast of the backbone. For $t \geq t_0$, the H forecasts for y_t are

$$\hat{y}_t^{(t-i)} = f_\theta\big(y_{t-i-L+1:\,t-i}\big)[\,i\,], \qquad i = 1, \ldots, H. \tag{1}$$

Track 1: Forecast disagreement. When an anomalous observation before t enters only some lookback windows, we expect greater disagreement among forecasts for y_t. We quantify this disagreement as the variance of the H forecasts:

$$\bar{y}_t = \frac{1}{H}\sum_{i=1}^{H} \hat{y}_t^{(t-i)}, \qquad d_t = \mathrm{Var}_{i=1,\ldots,H}\left[\hat{y}_t^{(t-i)}\right] = \frac{1}{H}\sum_{i=1}^{H}\left(\hat{y}_t^{(t-i)} - \bar{y}_t\right)^2. \tag{2}$$

Disagreement can be low when the forecasts are close to one another, even if they all deviate substantially from y_t. We therefore combine it with the one-step forecasting error, which measures deviation from y_t.

Track 2: One-step forecasting error. The second track measures the squared deviation of the one-step forecast from y_t (Figure 2, Track 2), as in conventional scoring:

$$e_{1,t} = \left(\hat{y}_t^{(t-1)} - y_t\right)^2. \tag{3}$$

4.3 SMOOTHING, NORMALISATION, AND TRACK AGGREGATION

EWMA smoothing. We apply an EWMA separately to both tracks. For the forecasting-error track, this reduces the influence of temporarily large errors that can occur even for normal observations. For $x \in \{d,\, e_1\}$, we define

$$\mathrm{EWMA}_\lambda(x)_t = \frac{\sum_{k=0}^{t-t_0} \lambda^k x_{t-k}}{\sum_{k=0}^{t-t_0} \lambda^k}, \qquad D = \mathrm{EWMA}_\lambda(d), \quad E_1 = \mathrm{EWMA}_\lambda(e_1). \tag{4}$$

Here, $\lambda \in [0, 1)$ is the decay factor, with a larger λ giving more weight to older values, and a symbol without a time subscript denotes the whole sequence. Both tracks share λ and t_0 for consistent smoothing and initialisation. We use $\lambda = 0.95$ across all backbones and time series (Section 5.1). Appendix C.1 reports the sensitivity to λ.

Normalisation. The scales of the two tracks depend on the backbone and the series, so we normalise each smoothed track by its normal reference level before track aggregation. Within the train segment Y_{tr}, we use the time steps from t_0 onward as the reference window $\mathcal{T}$. The normal reference level $q_{a,\mathcal{T}}$ is the empirical a-quantile of each smoothed track over $\mathcal{T}$, computed separately for each series:

$$\mathcal{T} \;=\; \{\, t_0, \ldots, t_{\mathrm{tr}} \,\}, \qquad \tilde{X}_t = \frac{X_t}{q_{a,\mathcal{T}}(X)}, \qquad X \in \{D,\, E_1\}. \tag{5}$$

We set $a = 0.99$ because a high quantile is less sensitive to outliers than the maximum. Appendix C.2 reports the sensitivity to a.

Track aggregation. FDAS defines the per-channel anomaly score s_t as the pointwise maximum of the normalised disagreement and forecasting-error tracks, so that a large value in either track is preserved even when the other is low. This aggregation applies only when both tracks can be normalised, which requires three conditions. First, the train segment must end at or after t_0 so that the reference window is nonempty ($\mathcal{T} \neq \varnothing$). Second, the normal reference level of the disagreement track must be positive ($q_{a,\mathcal{T}}(D) > 0$); for example, it is zero when the forecasts for y_t coincide at every time step in $\mathcal{T}$. Third, the same must hold for the forecasting-error track ($q_{a,\mathcal{T}}(E_1) > 0$); for example, it is zero when the one-step forecast equals y_t at every time step in $\mathcal{T}$. When any condition fails, the score falls back to $E_{1,t}$, the smoothed but unnormalised forecasting-error track.

$$s_t = \begin{cases} \max\left(\tilde{D}_t,\, \tilde{E}_{1,t}\right) & \text{if } \mathcal{T} \neq \varnothing \;\wedge\; q_{a,\mathcal{T}}(D) > 0 \;\wedge\; q_{a,\mathcal{T}}(E_1) > 0 \\ E_{1,t} & \text{otherwise} \end{cases} \tag{6}$$

For $t < t_0$, we set $s_t = s_{t_0}$. Appendix E.2 compares the maximum with mean, product, and minimum track aggregation.

4.4 CHANNEL AGGREGATION

We apply the scoring procedure in Sections 4.2 to 4.3 separately to each channel c and obtain scores $s_t^{(c)}$. At each time step, the channels share an anomaly label, so we aggregate their scores into one score. We take the maximum over the set $\mathcal{K}$ of channels that satisfy the three conditions of Eq. (6):

$$\mathcal{K} = \left\{\, c \in \{1, \ldots, C\} \;:\; \mathcal{T} \neq \varnothing \;\wedge\; q_{a,\mathcal{T}}\big(D^{(c)}\big) > 0 \;\wedge\; q_{a,\mathcal{T}}\big(E_1^{(c)}\big) > 0 \,\right\}. \tag{7}$$

When $\mathcal{K}$ is nonempty, channels outside $\mathcal{K}$ are dropped because their unnormalised scores are on a different scale. When $\mathcal{K}$ is empty, the score falls back to the maximum of the smoothed but unnormalised forecasting-error tracks across all channels.

$$S_t = \begin{cases} \max\limits_{c \in \mathcal{K}} s_t^{(c)} & \text{if } \mathcal{K} \neq \varnothing \\ \max\limits_{1 \le c \le C} E_{1,t}^{(c)} & \text{otherwise} \end{cases} \tag{8}$$

For $C = 1$, Eq. (8) reduces to Eq. (6). The channel maximum preserves high anomaly scores that occur in only a subset of channels. Appendix E.3 compares it with the channel mean and median. Appendix C.3 reports how often the series-level fallback is used and how including these series affects the reported gains. For backbones that forecast each channel independently, such as TTM-r2.1 (Ekambaram et al., 2024), this aggregation is the only step at which channels interact, even when $C > 1$.

Properties. FDAS introduces no learned parameters and requires no anomaly labels for the target series. Its scoring hyperparameters are λ and a, which we fix across all evaluated configurations. The lookback length L and forecast horizon H are forecast-generation hyperparameters selected for each backbone (Section 5.1). They determine $t_0 = L + H$, the first time step at which both tracks are available. The normal reference levels are computed once over $\mathcal{T}$ and remain fixed thereafter. Consequently, for $t \geq \max(t_0, t_{\text{tr}})$, scoring is causal: S_t depends only on $y_1, \ldots, y_t$ and can be computed online as y_t arrives. Algorithm 1 in Appendix A gives pseudocode for the procedure.

5 EXPERIMENTS

5.1 SETUP

Datasets and protocol. TSB-AD (Liu & Paparrizos, 2024) divides the series of each benchmark into disjoint tuning and evaluation splits. The tuning splits, 48 univariate and 20 multivariate series, are used only for hyperparameter selection. Following its official protocol, we evaluate on the full evaluation splits (350 univariate series from 23 datasets in TSB-AD-U and 180 multivariate series from 17 datasets in TSB-AD-M; Table 4) and score each series along its entire length including its train segment. Anomaly labels are used only to compute evaluation metrics, not for scoring.

Backbone models. We use all backbones as frozen forecasting TSFMs, and none was pre-trained for anomaly detection. The univariate evaluation includes eight backbones: Chronos-Bolt small and base (Ansari et al., 2024), Time-MoE base and large (Shi et al., 2025), TimesFM 1.0 and 2.0 (Das et al., 2024), and TTM-r2.1 L1 and L2 (Ekambaram et al., 2024). The multivariate evaluation includes the five backbones that support multivariate input: Chronos-2 (Ansari et al., 2025), TiRex-2 (Podest et al., 2026), Toto 2.0 (Khwaja et al., 2026), and TTM-r2.1 L1 and L2. TTM-r2.1 forecasts channels independently when used zero-shot. Appendix B.3 lists the checkpoint of each backbone.

With the global hyperparameters λ and a fixed, we tune only (L, H), over $L \in \{64, 128, 256, 512\}$ and $H \in \{32, 64, 96\}$ restricted to the pairs each backbone can produce. TTM-r2.1 instead uses the pairs for which its released checkpoints were trained. For each backbone–benchmark pair, we select (L, H) to maximise mean VUS-PR under FDAS on the tuning split and then apply both conventional scoring and FDAS to the same forecasts generated with the selected (L, H). We refer to a backbone–benchmark pair with its selected (L, H) as a *configuration*, with 13 configurations in total. Appendix B.4 lists the search space and tuning results for each backbone.

Baselines. We compare FDAS with three baseline groups. The first group is Last-3 (Zhu et al., 2026), which uses the mean of the three preceding observations as the one-step forecast and the squared forecasting error as the anomaly score. Last-3^{EWMA} smooths this score with the EWMA of Eq. (4), applied from the start of the series with the same λ as FDAS. Because Last-3 is defined for univariate series, we average its score over channels for multivariate series, following the forecasting-based detectors on the benchmarks (Liu & Paparrizos, 2024; Ekambaram et al., 2026):

$$\text{Last-3}_t \;=\; \frac{1}{C}\sum_{c=1}^{C}\Big(\frac{1}{3}\sum_{j=1}^{3} y_{t-j}^{(c)} \;-\; y_t^{(c)}\Big)^2, \qquad \text{Last-3}^{\text{EWMA}} = \text{EWMA}_\lambda(\text{Last-3}). \tag{9}$$

The second group is conventional scoring (denoted by *CS* in the results), which uses the one-step forecasting error $e_{1,t}$ of Eq. (3) as the anomaly score and averages it across channels for multivariate series.

The third group consists of TSFMs that directly support anomaly detection, with results taken from the TSB-AD leaderboard[1]: TSPulse (Ekambaram et al., 2026), MOMENT (Goswami et al., 2024), TimeRCD (Lan et al., 2026), and CHARM (Dutta et al., 2025). We include only entries evaluated on the full evaluation splits without fine-tuning on the target series. Table 1 also lists Lag-Llama (Rasul et al., 2023) among the Forecasting TSFMs with its leaderboard score under conventional scoring. We do not apply FDAS to it because it forecasts a single time step per forward pass and thus does not satisfy A2. $\mathcal{W}_1$-ACAS (Martinez Gil et al., 2026) reports results on a subset of the TSB-AD

[1]The leaderboard was accessed on 25 September 2026.

datasets. Appendix D.6 compares FDAS with this post-hoc scoring method on the univariate and multivariate datasets in that subset.

Evaluation metrics. Tables 1 and 2 report three range-wise and three point-wise metrics commonly used in recent work on TSB-AD (Park & Kang, 2026; Zhang et al., 2026). The range-wise metrics are VUS-PR, VUS-ROC (Paparrizos et al., 2022), and R-based-F1 (Tatbul et al., 2018). The point-wise metrics are AUC-PR, AUC-ROC, and Standard-F1. Following the benchmark paper (Liu & Paparrizos, 2024), we use VUS-PR as the primary metric. For our evaluations, we compute metrics per series under the benchmark settings and average them across series.

5.2 Results

Under conventional scoring, all evaluated backbones achieve mean VUS-PR similar to the model-free Last-3 baseline on each benchmark, despite differences in backbone size, architecture, and pre-training data (Figures 3(c) and 4(c)). This observation is consistent with Zhu et al. (2026). Applied to the same forecasts, FDAS increases mean VUS-PR by 0.15 to 0.21 over conventional scoring in all 13 configurations (Tables 1 and 2). Mean VUS-PR differs by at most 0.05 across backbones on the same benchmark under both conventional scoring and FDAS (Tables 1 and 2). FDAS outperforms conventional scoring on the five remaining evaluation metrics in every configuration. The per-series VUS-PR differences over conventional scoring and Last-3 are significant in every configuration under paired Wilcoxon signed-rank tests, and the bootstrap confidence intervals for the mean differences exclude zero (Appendix D.3).

The gains are also evident in the per-series comparisons. For Time-MoE base and TiRex-2, which yield the highest mean VUS-PR under FDAS on their benchmarks, FDAS outperforms conventional scoring and Last-3, each on a majority of series (Figures 3(a,b) and 4(a,b)). Appendices D.2, D.4, and D.5 provide dataset-level results, per-series results, and score traces on individual series.

With FDAS, frozen forecasting TSFMs become competitive with the leading AD TSFMs in mean VUS-PR on both benchmarks. On TSB-AD-U, the best FDAS configuration achieves 0.50, which is 0.02 below the 0.52 achieved by TimeRCD (Lan et al., 2026), the best-performing AD TSFM in this comparison (Table 1). On TSB-AD-M, three FDAS configurations achieve 0.40, which is 0.01 above CHARM (Dutta et al., 2025) at 0.39 (Table 2).

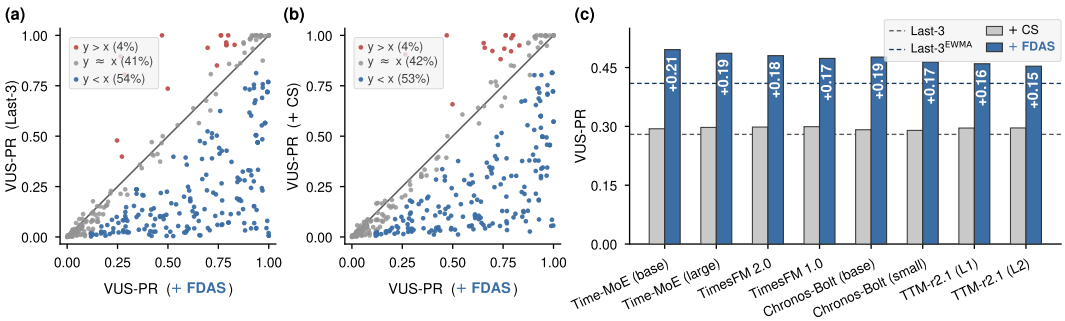


Figure 3: **TSB-AD-U** results. (a–b) Per-series VUS-PR of FDAS with Time-MoE base (Shi et al., 2025) versus Last-3 and conventional scoring. $y \approx x$ denotes differences within ± 0.1. (c) Mean VUS-PR across eight backbones. Numbers on the bars give the increase from CS to FDAS.

Ablation study. Without smoothing, adding the disagreement track to the forecasting-error track raises mean VUS-PR by 0.07–0.10 on univariate series and 0.03–0.07 on multivariate series (Table 3). On multivariate series, $\tilde{e}_1$ exceeds CS by 0.07–0.09 because it normalises each channel's error and takes the channel maximum instead of averaging across channels. However, without smoothing, FDAS does not improve on $\tilde{d}$ alone on multivariate series.

Smoothing improves both Last-3 and CS (Table 3). With smoothing, FDAS exceeds both $\tilde{E}_1$ and $\tilde{D}$ in all 13 configurations. Under paired Wilcoxon signed-rank tests, the gain is statistically significant over $\tilde{D}$ in all 13 configurations and over $\tilde{E}_1$ in eight (Appendix E.1). The tracks differ between

Table 1: Results on **TSB-AD-U**, averaged over 350 series. AD TSFMs directly support anomaly detection, and Base denotes the Last-3 baselines. The + CS rows report conventional scoring, and the + FDAS rows apply FDAS to the same forecasts. **Bold** and underline mark the best and second-best displayed values.

	Method		Range-Wise Metrics ↑			Point-Wise Metrics ↑		
			VUS-PR	VUS-ROC	R-based-F1	AUC-PR	AUC-ROC	Standard-F1
AD TSFMs	TimeRCD†		**0.52**	**0.87**	**0.55**	**0.50**	**0.84**	**0.53**
	CHARM†		0.50	–	–	–	–	–
	TSPulse†		0.48	–	–	–	–	–
	MOMENT†		0.38	0.75	0.36	0.30	0.68	0.35
Forecasting TSFMs	Time-MoE (base)	+ CS	0.29	0.74	0.34	0.29	0.67	0.34
		+ **FDAS**	0.50	0.82	0.45	0.39	0.76	0.44
	Time-MoE (large)	+ CS	0.30	0.74	0.35	0.29	0.67	0.34
		+ **FDAS**	0.49	0.83	0.45	0.38	0.77	0.44
	TimesFM 2.0	+ CS	0.30	0.75	0.35	0.29	0.68	0.35
		+ **FDAS**	0.48	0.81	0.44	0.38	0.76	0.44
	TimesFM 1.0	+ CS	0.30	0.75	0.35	0.29	0.69	0.34
		+ **FDAS**	0.47	0.82	0.44	0.36	0.77	0.43
	Chronos-Bolt (base)	+ CS	0.29	0.74	0.34	0.28	0.68	0.34
		+ **FDAS**	0.48	0.82	0.42	0.36	0.76	0.42
	Chronos-Bolt (small)	+ CS	0.29	0.74	0.35	0.28	0.67	0.34
		+ **FDAS**	0.46	0.81	0.42	0.35	0.76	0.41
	TTM-r2.1 (L1)	+ CS	0.30	0.74	0.34	0.28	0.67	0.33
		+ **FDAS**	0.46	0.80	0.42	0.34	0.73	0.40
	TTM-r2.1 (L2)	+ CS	0.30	0.74	0.34	0.28	0.67	0.33
		+ **FDAS**	0.45	0.80	0.41	0.34	0.73	0.39
	Lag-Llama†	+ CS	0.27	0.72	0.31	0.25	0.65	0.30
Base	Last-3		0.28	0.71	0.34	0.27	0.64	0.32
	Last-3$^{\text{EWMA}}$		0.41	0.76	0.36	0.29	0.69	0.35

† Taken from the TSB-AD leaderboard. Dashes mark unreported metrics.

Table 2: Results on **TSB-AD-M**, averaged over 180 series, in the same format as Table 1. **Bold** and underline mark the best and second-best displayed values.

	Method		Range-Wise Metrics ↑			Point-Wise Metrics ↑		
			VUS-PR	VUS-ROC	R-based-F1	AUC-PR	AUC-ROC	Standard-F1
AD TSFMs	CHARM†		0.39	–	–	–	–	–
	TSPulse†		0.36	–	–	–	–	–
	TimeRCD†		0.32	0.71	**0.43**	0.34	**0.72**	0.38
Forecasting TSFMs	TiRex-2	+ CS	0.21	0.67	0.25	0.20	0.62	0.24
		+ **FDAS**	**0.40**	0.75	0.32	0.34	0.69	0.39
	TTM-r2.1 (L2)¶	+ CS	0.24	0.70	0.25	0.23	0.65	0.27
		+ **FDAS**	**0.40**	**0.76**	0.35	**0.35**	**0.72**	**0.41**
	TTM-r2.1 (L1)¶	+ CS	0.24	0.67	0.25	0.23	0.62	0.27
		+ **FDAS**	**0.40**	0.75	0.35	**0.35**	0.71	**0.41**
	Chronos-2	+ CS	0.22	0.69	0.25	0.22	0.64	0.25
		+ **FDAS**	0.38	0.74	0.29	0.32	0.69	0.38
	Toto 2.0 (22m)	+ CS	0.21	0.67	0.25	0.21	0.62	0.24
		+ **FDAS**	0.37	0.74	0.28	0.32	0.70	0.37
Base	Last-3		0.21	0.63	0.25	0.21	0.58	0.25
	Last-3$^{\text{EWMA}}$		0.31	0.68	0.26	0.25	0.62	0.31

† Taken from the TSB-AD leaderboard. Dashes mark unreported metrics; ¶ TTM-r2.1 forecasts channels independently.

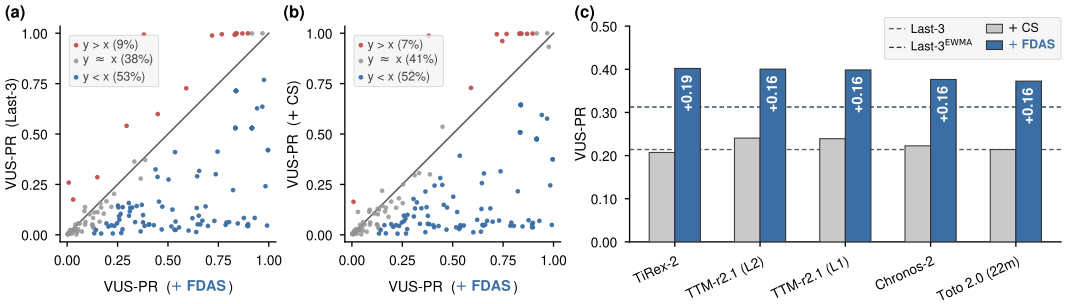


Figure 4: **TSB-AD-M** results in the same format as Figure 3, with TiRex-2 (Podest et al., 2026) in (a–b).

benchmarks: $\tilde{E}_1$ performs better than $\tilde{D}$ on univariate series, whereas $\tilde{D}$ performs at least as well as $\tilde{E}_1$ on multivariate series. The disagreement track alone performs worse than Last-3$^{\mathrm{EWMA}}$ on univariate series but better on multivariate series. The advantage of $\tilde{D}$ over $\tilde{E}_1$ increases with the average anomaly length of a series (Appendix E.4).

Table 3: Track and smoothing ablation. Mean VUS-PR on the evaluation splits. $\tilde{e}_1$ and $\tilde{d}$ are the normalised single tracks without smoothing, $\tilde{E}_1$ and $\tilde{D}$ with smoothing, and CS$^{\mathrm{EWMA}}$ is $\mathrm{EWMA}_\lambda(\mathrm{CS})$. The Last-3 rows report Last-3 and Last-3$^{\mathrm{EWMA}}$. **Bold** marks the maximum in each backbone row.

		Without smoothing ($\lambda = 0$)				EWMA ($\lambda = 0.95$)			
	Method	CS	$\tilde{e}_1$	$\tilde{d}$	FDAS	CS$^{\mathrm{EWMA}}$	$\tilde{E}_1$	$\tilde{D}$	FDAS
Univariate	Time-MoE (base)	0.29	0.29	0.32	0.38	0.47	0.47	0.34	**0.50**
	Time-MoE (large)	0.30	0.30	0.28	0.37	0.47	0.47	0.29	**0.49**
	TimesFM 2.0	0.30	0.30	0.33	0.39	0.46	0.46	0.34	**0.48**
	TimesFM 1.0	0.30	0.30	0.32	0.39	0.46	0.46	0.31	**0.47**
	Chronos-Bolt (base)	0.29	0.29	0.35	0.39	0.45	0.45	0.36	**0.48**
	Chronos-Bolt (small)	0.29	0.29	0.36	0.39	0.45	0.45	0.36	**0.46**
	TTM-r2.1 (L1)	0.30	0.30	0.33	0.39	0.43	0.43	0.33	**0.46**
	TTM-r2.1 (L2)	0.30	0.30	0.33	0.39	0.43	0.43	0.32	**0.45**
	Last-3		0.28				0.41		
Multivariate	TiRex-2	0.21	0.28	0.37	0.35	0.31	0.36	0.38	**0.40**
	TTM-r2.1 (L2)	0.24	0.32	0.39	0.38	0.33	0.36	0.39	**0.40**
	TTM-r2.1 (L1)	0.24	0.32	0.38	0.38	0.33	0.37	0.37	**0.40**
	Chronos-2	0.22	0.31	0.34	0.34	0.33	0.35	0.36	**0.38**
	Toto 2.0 (22m)	0.21	0.30	0.34	0.34	0.33	0.34	0.36	**0.37**
	Last-3		0.21				0.31		

6 Conclusion

Under conventional scoring, frozen forecasting TSFMs achieve mean VUS-PR similar to a model-free baseline. We introduced FDAS, a post-hoc scoring layer that aggregates the one-step forecasting error and the disagreement among a backbone's existing forecasts for the same target time step made at different forecast origins. FDAS outperformed conventional scoring in all 13 configurations without requiring model parameter updates or anomaly labels for the target series. It also made frozen forecasting TSFMs competitive with the leading AD TSFMs in mean VUS-PR. The disagreement track improved VUS-PR both without and with smoothing, whereas smoothing improved the model-free baseline as well as conventional scoring. These results suggest that a substantial part of the limitation lies in the scoring layer rather than in the forecasters, and that FDAS makes anomaly detection a viable downstream task for frozen forecasting TSFMs. Appendix F discusses its limitations.

AI USE STATEMENT

We used generative AI tools for the three purposes declared at submission. To aid writing, they assisted with drafting, translating, and revising the manuscript. For retrieval and discovery, they assisted with literature searches and reference formatting. For research ideation and execution, they provided feedback on our methodology and experimental design, critically reviewed our interpretations of the results, and assisted with implementing the method, editing code, and preparing figures and tables. We reviewed all AI-assisted work, tested the code, and verified references against the original sources. We take full responsibility for the content of this paper, including all AI-assisted text, code, claims, and artefacts.

ETHICS STATEMENT

We used only publicly available datasets distributed by the TSB-AD benchmark (Liu & Paparrizos, 2024) and publicly released model checkpoints, in accordance with their licences. Some of these datasets contain de-identified physiological recordings, including electrocardiograms and wearable sensor data from patients. We collected no new data and recruited no human participants.

REPRODUCIBILITY STATEMENT

Section 4 and Algorithm 1 specify the complete scoring procedure. Appendices B.3 and B.4 document the backbone checkpoints, preprocessing, point-forecast extraction, and the (L, H) search spaces and selected values, and Section 5.1 and Appendix B.1 describe the evaluation protocol. The supplementary material and the anonymised repository at `https://anonymous.4open.science/r/FDAS-TSBAD` contain the code for forecast generation, scoring, and table generation, pinned dependencies and checkpoint revisions for every backbone, and checksummed per-series results for every configuration. The reported tables are regenerated from these results without a GPU. Forecasts are not distributed because of their size, and the repository regenerates them from the official dataset archives.

REFERENCES

Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2021.

Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134–147, 2017.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=gerNCVqqtR`.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. URL `https://arxiv.org/abs/2510.15821`.

Marc Bächlin, Meir Plotnik, Daniel Roggen, Inbal Maidan, Jeffrey M. Hausdorff, Nir Giladi, and Gerhard Tröster. Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):436–446, 2010.

Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A. Lozano. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys*, 54(3):1–33, 2021. doi: 10.1145/3444690.

Paul Boniol, John Paparrizos, Themis Palpanas, and Michael J. Franklin. SAND: Streaming subsequence anomaly detection. *Proceedings of the VLDB Endowment*, 14(10):1717–1729, 2021.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning*, 2024.

Utsav Dutta, Sina Khoshfetrat Pakazad, and Henrik Ohlsson. Time to embed: Unlocking foundation models for time series with channel descriptions. *arXiv preprint arXiv:2505.14543*, 2025.

Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H. Nguyen, Wesley M. Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny time mixers (TTMs): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. In *Advances in Neural Information Processing Systems*, 2024.

Vijay Ekambaram, Subodh Kumar, Arindam Jati, Sumanta Mukherjee, Tomoya Sakai, Pankaj Dayama, Wesley M. Gifford, and Jayant Kalagnanam. TSPulse: Tiny pre-trained models with disentangled representations for rapid time-series analysis. In *International Conference on Learning Representations*, 2026.

Kamil Faber, Marcin Pietroń, Dominik Żurek, and Roberto Corizzo. xLSTMAD: A powerful xLSTM-based method for anomaly detection. In *IEEE International Conference on Data Mining (ICDM)*, pp. 247–256, 2025. doi: 10.1109/ICDM65498.2025.00032.

Pavel Filonov, Andrey Lavrentyev, and Artem Vorontsov. Multivariate industrial time series with cyber-attack simulation: Fault detection using an LSTM-based predictive data model. In *NIPS Time Series Workshop*, 2016.

Patrick Fleith. Controlled anomalies time series (CATS) dataset (version 2), September 2023. URL `https://zenodo.org/records/8338435`.

Ary L. Goldberger, Luis A. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Mononito Goswami, Cristian Ignacio Challu, Laurent Callot, Lenon Minorics, and Andrey Kan. Unsupervised model selection for time series anomaly detection. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=gOZ_pKANaPW`.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *International Conference on Machine Learning*, 2024.

Scott David Greenwald. *Improved Detection and Classification of Arrhythmias in Noise-Corrupted Electrocardiograms Using Contextual Information*. PhD thesis, Massachusetts Institute of Technology, 1990.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.

IOPS. IOPS. `http://iops.ai/dataset_detail/?id=10`, n.d.

Vincent Jacob, Fei Song, Arnaud Stiegler, Bijan Rad, Yanlei Diao, and Nesime Tatbul. Exathlon: A benchmark for explainable anomaly detection over time series. *Proceedings of the VLDB Endowment*, 14(11):2613–2626, 2021.

Sheo Yon Jhin and Noseong Park. Point-wise anomaly detection via Fold-Bifurcation ODE. In *International Conference on Learning Representations*, 2026.

Hyeongwon Kang, Jinwoo Park, Seunghun Han, and Pilsung Kang. Forecasting anomaly precursors via uncertainty-aware time-series ensembles. arXiv preprint arXiv:2602.17028, 2026.

Eamonn Keogh, Jessica Lin, Sang-Hee Lee, and Helga Van Herle. Finding the most unusual time series subsequence: Algorithms and applications. *Knowledge and Information Systems*, 11:1–27, 2007.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, Chenghao Liu, Ameet Talwalkar, and David Asker. Toto 2.0: Time series forecasting enters the scaling era. arXiv preprint arXiv:2605.20119, 2026.

Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Hu. Revisiting time series outlier detection: Definitions and benchmarks. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Tian Lan, Hao Duong Le, Jinbo Li, Wenjun He, Meng Wang, Chenghao Liu, and Chen Zhang. Towards foundation models for zero-shot time series anomaly detection: Leveraging synthetic data and relative context discrepancy. In *International Conference on Machine Learning*, 2026.

Nikolay Laptev, Saeed Amizadeh, and Youssef Billawala. S5 — a labeled anomaly detection dataset, version 1.0(16M), March 2015.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6555–6565, 2024.

Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In *European Symposium on Artificial Neural Networks (ESANN)*, 2015.

Natalia Martinez Gil, Fearghal O'Donncha, Wesley M. Gifford, Nianjun Zhou, Dhaval C. Patel, and Roman Vaculin. Adaptive conformal anomaly detection with time series foundation models for signal monitoring. In *International Conference on Learning Representations*, 2026.

Aditya P. Mathur and Nils Ole Tippenhauer. SWaT: A water treatment testbed for research and training on ICS security. In *International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater)*, pp. 31–36, 2016.

Steffen Moritz, Frederik Rehbach, Sowmya Chandrasekaran, Margarita Rebolledo, and Thomas Bartz-Beielstein. GECCO industrial challenge 2018 dataset: A water quality dataset for the "Internet of Things: Online anomaly detection for drinking water quality" competition at the Genetic and Evolutionary Computation Conference 2018, Kyoto, Japan, 2018. URL `https://zenodo.org/records/3884398`.

Mohsin Munir, Shoaib Ahmed Siddiqui, Andreas Dengel, and Sheraz Ahmed. DeepAnT: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*, 7:1991–2005, 2019. doi: 10.1109/ACCESS.2018.2886457.

John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S. Tsay, Aaron Elmore, and Michael J. Franklin. Volume under the surface: A new accuracy evaluation measure for time-series anomaly detection. *Proceedings of the VLDB Endowment*, 15(11):2774–2787, 2022.

Jinju Park and Seokho Kang. PaAno: Patch-based representation learning for time-series anomaly detection. In *International Conference on Learning Representations*, 2026.

Miseon Park and Kijung Yoon. A comparative study of adaptation strategies for time series foundation models in anomaly detection. In *International Conference on Pattern Recognition (ICPR)*, volume 16818 of *Lecture Notes in Computer Science*, pp. 636–650. Springer, 2026. doi: 10.1007/978-3-032-31397-3_42.

Patrick Podest, Marco Pichler, Elias Bürger, Levente Zólyomi, Bernhard Voggenberger, Wilhelm Berghammer, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex-2: Generalizing TiRex to multivariate data and streaming. arXiv preprint arXiv:2607.01204, 2026.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. Lag-Llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. Collecting complex activity datasets in highly rich networked sensor environments. In *International Conference on Networked Sensing Systems (INSS)*, pp. 233–240, 2010.

Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. Anomaly detection in time series: A comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15(9):1779–1797, 2022. doi: 10.14778/3538598.3538602.

Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *International Conference on Information Systems Security and Privacy (ICISSP)*, volume 1, pp. 108–116, 2018.

Qichao Shentu, Beibu Li, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Towards a general time series anomaly detector with adaptive bottlenecks and dual adversarial decoders. In *International Conference on Learning Representations*, 2025.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations*, 2025.

Haotian Si, Jianhui Li, Changhua Pei, Hang Cui, Jingwen Yang, Yongqian Sun, Shenglin Zhang, Jingjing Li, Haiming Zhang, Jing Han, Dan Pei, and Gaogang Xie. TimeSeriesBench: An industrial-grade benchmark for time series anomaly detection models. In *IEEE International Symposium on Software Reliability Engineering (ISSRE)*, 2024.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

TAO. Tropical Atmosphere Ocean (TAO) array. `https://tao.ndbc.noaa.gov/`, n.d.

Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. Precision and recall for time series. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Markus Thill, Wolfgang Konen, and Thomas Bäck. MarkusThill/MGAB: The Mackey-Glass anomaly benchmark. `https://doi.org/10.5281/zenodo.3762385`, April 2020.

Luan Tran, Liyue Fan, and Cyrus Shahabi. Distance-based outlier detection in data streams. *Proceedings of the VLDB Endowment*, 9(12):1089–1100, 2016.

Alexander von Birgelen and Oliver Niggemann. Anomaly detection and localization for cyber-physical production systems with self-organizing maps. In *IMPROVE — Innovative Modelling Approaches for Production Systems to Raise Validatable Efficiency*, pp. 55–71, 2018.

Renjie Wu and Eamonn J. Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Transactions on Knowledge and Data Engineering*, 35 (3):2421–2429, 2023.

Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1):1–42, 2024. doi: 10.1145/3691338.

Kang Zhang, Wei Jian Lau, Shoushou Ren, Dong Lin, Joon Son Chung, and Chuanhao Sun. PAI: Preserving amplitude information in representation-based time-series anomaly detection. arXiv preprint arXiv:2606.08935, 2026.

Shenglin Zhang, Zhenyu Zhong, Dongwen Li, Qiliang Fan, Yongqian Sun, Man Zhu, Yuzhi Zhang, Dan Pei, Jiyan Sun, Yinlong Liu, et al. Efficient KPI anomaly detection through transfer learning for large-scale web services. *IEEE Journal on Selected Areas in Communications*, 40(8):2440–2455, 2022.

Xiaokun Zhu, Louis Carpentier, and Mathias Verbeke. When foundation models are one-liners: Limitations and future directions for time series anomaly detection. In *International Conference on Learning Representations*, 2026.

A METHOD DETAILS

A.1 ALGORITHM

Algorithm 1 gives pseudocode for the complete procedure of Section 4, with the global hyperparameters fixed at $\lambda = 0.95$ and $a = 0.99$ in all 13 configurations.

Algorithm 1 FDAS

Input: series $Y = (y_1, \ldots, y_T) \in \mathbb{R}^{T \times C}$; train segment $Y_{\mathrm{tr}} = (y_1, \ldots, y_{t_{\mathrm{tr}}})$; frozen backbone f_θ with lookback length L and forecast horizon H satisfying A1 and A2 in Section 4.1; global hyperparameters λ, a; $q_{a,\mathcal{T}}(X)$ denotes the empirical a-quantile of X over $\mathcal{T}$ computed using linear interpolation.
Output: anomaly scores $\{S_t\}_{t=1}^{T}$, $S_t \in \mathbb{R}_{\geq 0}$

Forecast generation. The backbone is evaluated only in this loop, which uses past-only inputs and advances the lookback window by one time step, as in Section 3.

for $\tau = L$ **to** $T - 1$ **do** ▷ τ is an origin, so $\tau = t - i$
  $\hat{y}^{(\tau)}_{\tau+i} \leftarrow f_\theta\big(y_{\tau-L+1:\tau}\big)[\,i\,]$, $i = 1, \ldots, \min(H, T - \tau)$ ▷ Eq. (1), origin fixed
end for
$t_0 \leftarrow L + H$ ▷ scoring start
$\mathcal{T} \leftarrow \{\, t_0, \ldots, t_{\mathrm{tr}} \,\}$ ▷ Eq. (5)

Scoring. The following quantities are computed for $t_0 \leq t \leq T$. The two tracks are computed separately for each channel; superscripts are shown only where the channel index matters.

$\bar{y}_t \leftarrow \frac{1}{H}\sum_{i=1}^{H} \hat{y}_t^{(t-i)}$, $d_t \leftarrow \frac{1}{H}\sum_{i=1}^{H} \big(\hat{y}_t^{(t-i)} - \bar{y}_t\big)^2$ ▷ Eq. (2)
$e_{1,t} \leftarrow \big(\hat{y}_t^{(t-1)} - y_t\big)^2$ ▷ Eq. (3)
$D \leftarrow \mathrm{EWMA}_\lambda(d)$, $E_1 \leftarrow \mathrm{EWMA}_\lambda(e_1)$ ▷ Eq. (4)
$\mathcal{K} \leftarrow \big\{\, c \in \{1, \ldots, C\} : \mathcal{T} \neq \varnothing \wedge q_{a,\mathcal{T}}\big(D^{(c)}\big) > 0 \wedge q_{a,\mathcal{T}}\big(E_1^{(c)}\big) > 0 \,\big\}$ ▷ Eq. (7)
$\tilde{D}_t^{(c)} \leftarrow D_t^{(c)} / q_{a,\mathcal{T}}\big(D^{(c)}\big)$, $\tilde{E}_{1,t}^{(c)} \leftarrow E_{1,t}^{(c)} / q_{a,\mathcal{T}}\big(E_1^{(c)}\big)$ for $c \in \mathcal{K}$ ▷ Eq. (5)
$s_t^{(c)} \leftarrow \max\big(\tilde{D}_t^{(c)},\ \tilde{E}_{1,t}^{(c)}\big)$ for $c \in \mathcal{K}$ ▷ Eq. (6)
$s_t^{(c)} \leftarrow E_{1,t}^{(c)}$ for $c \notin \mathcal{K}$ ▷ Eq. (6)
if $\mathcal{K} \neq \varnothing$ **then**
  $S_t \leftarrow \max_{c \in \mathcal{K}}\ s_t^{(c)}$ ▷ Eq. (8)
else
  $S_t \leftarrow \max_{1 \leq c \leq C}\ s_t^{(c)}$ ▷ Eq. (8)
end if

$S_t \leftarrow S_{t_0}$ for $t < t_0$ ▷ Section 4.3
return $\{S_t\}_{t=1}^{T}$

A.2 COMPUTATIONAL COST

FDAS requires no additional backbone inference relative to conventional scoring under the protocol in Section 5.1. Algorithm 1 evaluates f_θ only during forecast generation, once at each of the $T - L$ forecast origins. Computing forecast disagreement takes $O(THC)$ time, while the one-step forecasting error, EWMA smoothing, and track and channel aggregation take $O(TC)$ time. The normal reference levels $q_{a,\mathcal{T}}$ are computed once per track and channel over the reference window $\mathcal{T}$.

A.3 ALIGNMENT OF THE TWO TRACKS AT t_0

The one-step forecasting error $e_{1,t}$ is defined for $t \geq L + 1$, whereas disagreement d_t requires all H forecasts of the same target time step and is defined only for $t \geq t_0 = L + H$. FDAS starts both tracks at t_0. Had the EWMA of $e_{1,t}$ started at $L + 1$, E_{1,t_0} would be a weighted average of H values, whereas $D_{t_0} = d_{t_0}$. Their relative magnitudes would then reflect initialisation as well

as the anomaly signal, and the same mismatch would affect the normal reference levels. As in Section 4.3, FDAS sets $s_t = s_{t_0}$ for $t < t_0$, and we likewise set conventional scoring and $\mathrm{CS}^{\mathrm{EWMA}}$ to their values at t_0. Last-3 is scored from the start of the series. For its first three time steps, which lack three preceding observations, we follow the reference implementation of Zhu et al. (2026) and backfill the rolling mean from the fourth time step.

B Experimental protocol

B.1 Experimental pipeline

Figure 5 summarises the experimental procedure, which we apply separately to the univariate and multivariate benchmarks. For each backbone and benchmark, we select an (L, H) pair on the tuning split. Appendix C describes the selection of the global hyperparameters λ and a. We compute metrics with the PyPI release 1.5 of TSB-AD, which the leaderboard uses. As in its evaluation scripts `Run_Detector_U.py` and `Run_Detector_M.py`, we score each series over its full length and set the sliding-window length of the VUS metrics for each series from the autocorrelation period of its first channel.

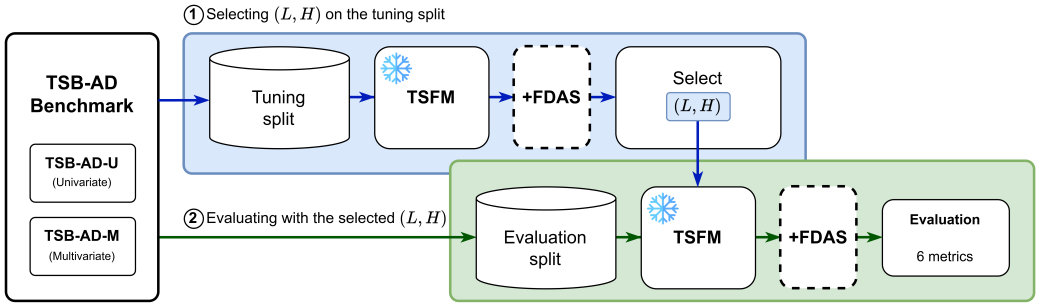


Figure 5: Experimental pipeline. Conventional scoring uses the same forecasts as FDAS.

B.2 Dataset composition

TSB-AD-U draws on 23 datasets and TSB-AD-M on 17, with 12 shared between the two benchmarks (Table 4). The tuning splits share no series with the evaluation splits.

B.3 Backbones and point-forecast extraction

Table 5 lists the checkpoint and point forecast of each configuration. Time-MoE, TimesFM, and TTM-r2.1 output point forecasts directly. For Chronos-Bolt, Chronos-2, TiRex-2, and Toto 2.0, which have a quantile head, we use the median $q_{0.5}$ to satisfy A1, as returned by the Chronos pipelines or extracted from the quantile outputs in the TiRex-2 and Toto 2.0 reference examples. We use each series as provided by the benchmark, follow each backbone's documented zero-shot inference procedure, including its input normalisation and forecast denormalisation, and apply no further processing to the forecasts.

B.4 Search space and selection of (L, H)

Table 6 lists the search space and selected (L, H) of each configuration, and Table 7 reports the tuning-split mean VUS-PR and fallback rate of every candidate pair. The search space contains the pairs from $L \in \{64, 128, 256, 512\}$ and $H \in \{32, 64, 96\}$ for which the backbone produces all H steps in a single forward pass, and TTM-r2.1 uses the pairs of its released checkpoints. For each backbone–benchmark pair, we select the (L, H) pair with the highest mean VUS-PR under FDAS on the tuning split, and conventional scoring uses the same pair. Appendix C.3 reports the fallback rates on the evaluation splits.

Table 4: Dataset composition of the two benchmarks. Each number is the count of series that a dataset contributes to the evaluation or tuning split. A dash marks no series in that split, and a blank marks a dataset absent from the benchmark.

			TSB-AD-U		TSB-AD-M	
	Dataset	Source	Evaluation	Tuning	Evaluation	Tuning
Shared	CATSv2	Fleith (2023)	1	–	5	1
	Daphnet	Bächlin et al. (2010)	1	–	1	–
	Exathlon	Jacob et al. (2021)	30	2	25	2
	LTDB	Goldberger et al. (2000)	8	1	4	1
	MITDB	Goldberger et al. (2000)	7	1	11	2
	MSL	Hundman et al. (2018)	7	2	14	2
	OPPORTUNITY	Roggen et al. (2010)	27	2	7	1
	SMAP	Hundman et al. (2018)	17	2	25	2
	SMD	Su et al. (2019)	33	5	20	2
	SVDB	Greenwald (1990)	18	2	28	3
	SWaT	Mathur & Tippenhauer (2016)	1	–	2	–
	TAO	TAO (n.d.)	2	1	11	2
Univariate	IOPS	IOPS (n.d.)	15	2		
	MGAB	Thill et al. (2020)	8	1		
	NAB	Ahmad et al. (2017)	23	5		
	NEK	Si et al. (2024)	8	1		
	Power	Keogh et al. (2007)	1	–		
	SED	Boniol et al. (2021)	2	1		
	Stock	Tran et al. (2016)	8	2		
	TODS	Lai et al. (2021)	13	2		
	UCR	Wu & Keogh (2023)	70	6		
	WSD	Zhang et al. (2022)	20	5		
	YAHOO	Laptev et al. (2015)	30	5		
Multivariate	CreditCard	Sharafaldin et al. (2018)			1	–
	GECCO	Moritz et al. (2018)			1	–
	Genesis	von Birgelen & Niggemann (2018)			1	–
	GHL	Filonov et al. (2016)			23	2
	PSM	Abdulaal et al. (2021)			1	–
Total			350	48	180	20

Table 5: Checkpoints and point-forecast extraction procedures for all 13 configurations. For TTM-r2.1, each released (L, H) pair uses a repository branch named `L-H-ft-r2.1` for L2 or `L-H-ft-l1-r2.1` for L1.

	TSFM	Source	Checkpoint	Point forecast
Univariate	Time-MoE (base)	Shi et al. (2025)	`Maple728/TimeMoE-50M`	point output
	Time-MoE (large)	Shi et al. (2025)	`Maple728/TimeMoE-200M`	point output
	TimesFM 2.0	Das et al. (2024)	`google/timesfm-2.0-500m-pytorch`	point output
	TimesFM 1.0	Das et al. (2024)	`google/timesfm-1.0-200m-pytorch`	point output
	Chronos-Bolt (base)	Ansari et al. (2024)	`amazon/chronos-bolt-base`	median $q_{0.5}$
	Chronos-Bolt (small)	Ansari et al. (2024)	`amazon/chronos-bolt-small`	median $q_{0.5}$
	TTM-r2.1 (L1)	Ekambaram et al. (2024)	`ibm-granite/granite-timeseries-ttm-r2`	point output
	TTM-r2.1 (L2)	Ekambaram et al. (2024)	`ibm-granite/granite-timeseries-ttm-r2`	point output
Multivariate	TiRex-2	Podest et al. (2026)	`NX-AI/TiRex-2`	median $q_{0.5}$
	Chronos-2	Ansari et al. (2025)	`amazon/chronos-2`	median $q_{0.5}$
	Toto 2.0 (22m)	Khwaja et al. (2026)	`Datadog/Toto-2.0-22m`	median $q_{0.5}$
	TTM-r2.1 (L1)	Ekambaram et al. (2024)	`ibm-granite/granite-timeseries-ttm-r2`	point output
	TTM-r2.1 (L2)	Ekambaram et al. (2024)	`ibm-granite/granite-timeseries-ttm-r2`	point output

Table 6: Search spaces and selected (L, H) pairs. For each backbone, the selected pair maximises tuning-split mean VUS-PR under FDAS. Candidates counts the pairs for which the backbone produces all H steps in a single forward pass, as documented under Horizon output. TTM-r2.1 uses its released checkpoint pairs.

	TSFM (+FDAS)	(L, H)	Horizon output	Candidates	Selected
Univariate	Time-MoE (base)	$\{64, 128, 256, 512\} \times \{32, 64\}$	heads 1, 8, 32, 64	8	**(512, 32)**
	Time-MoE (large)	$\{64, 128, 256, 512\} \times \{32, 64\}$	heads 1, 8, 32, 64	8	**(512, 64)**
	TimesFM 2.0	$\{64, 128, 256, 512\} \times \{32, 64, 96\}$	output patch 128	12	**(512, 32)**
	TimesFM 1.0	$\{64, 128, 256, 512\} \times \{32, 64, 96\}$	output patch 128	12	**(512, 64)**
	Chronos-Bolt (base)	$\{64, 128, 256, 512\} \times \{32, 64\}$	native horizon 64	8	**(512, 32)**
	Chronos-Bolt (small)	$\{64, 128, 256, 512\} \times \{32, 64\}$	native horizon 64	8	**(512, 32)**
	TTM-r2.1 (L1)	$(52, 16)$ $(90, 30)$ $(180, 60)$ $(360, 60)$ $(512, 48)$ $(512, 96)$	one checkpoint per pair	6	**(512, 48)**
	TTM-r2.1 (L2)	$(52, 16)$ $(90, 30)$ $(512, 48)$ $(512, 96)$	one checkpoint per pair	4	**(512, 48)**
Multivariate	TiRex-2	$\{64, 128, 256, 512\} \times \{32, 64, 96\}$	output patch 32	12	**(64, 64)**
	Chronos-2	$\{64, 128, 256, 512\} \times \{32, 64, 96\}$	output patch 16	12	**(256, 64)**
	Toto 2.0 (22m)	$\{64, 128, 256, 512\} \times \{32, 64, 96\}$	output patch 32	12	**(256, 64)**
	TTM-r2.1 (L1)	$(52, 16)$ $(90, 30)$ $(180, 60)$ $(360, 60)$ $(512, 48)$ $(512, 96)$	one checkpoint per pair	6	**(512, 96)**
	TTM-r2.1 (L2)	$(52, 16)$ $(90, 30)$ $(512, 48)$ $(512, 96)$	one checkpoint per pair	4	**(512, 48)**

Table 7: FDAS tuning-split mean VUS-PR over the search spaces of Table 6. **Bold** and underline mark the selected and runner-up pairs, and dashes mark pairs outside the backbone's search space. Shading is scaled within each row, darker for higher scores. Fallback is the percentage of tuning series scored through the fallback branch of Eq. (6). It depends on $t_0 = L + H$ alone and is therefore identical for every backbone in a block. The TTM-r2.1 block is indexed by its released pairs, not by the L and H headings. Its fallback rates are 29% on TSB-AD-U and 15% on TSB-AD-M at $L = 512$, and zero otherwise.

		$L = 64$			$L = 128$			$L = 256$			$L = 512$		
	TSFM (+FDAS)	32	64	96	32	64	96	32	64	96	32	64	96
Univariate	Time-MoE (base)	0.409	0.413	–	0.419	0.405	–	0.431	0.417	–	**0.440**	0.438	–
	Time-MoE (large)	0.408	0.411	–	0.422	0.422	–	0.429	0.405	–	0.441	**0.443**	–
	TimesFM 2.0	0.425	0.423	0.421	0.430	0.415	0.405	0.407	0.386	0.381	**0.444**	0.428	0.416
	TimesFM 1.0	0.407	0.411	0.385	0.403	0.408	0.384	0.394	0.382	0.384	0.428	**0.428**	0.410
	Chronos-Bolt (base)	0.404	0.399	–	0.421	0.410	–	0.423	0.406	–	**0.450**	0.438	–
	Chronos-Bolt (small)	0.408	0.416	–	0.414	0.409	–	0.422	0.414	–	**0.446**	0.439	–
	Fallback	0%	0%	0%	0%	0%	0%	0%	0%	0%	27%	29%	29%
Multivariate	TiRex-2	0.502	**0.514**	0.493	0.493	0.504	0.495	0.487	0.503	0.493	0.465	0.486	0.468
	Chronos-2	0.409	0.454	0.451	0.405	0.429	0.428	0.445	**0.464**	0.437	0.438	0.442	0.427
	Toto 2.0 (22m)	0.389	0.409	0.413	0.378	0.384	0.381	0.410	**0.433**	0.424	0.415	0.422	0.413
	Fallback	0%	0%	0%	0%	0%	0%	0%	0%	0%	15%	15%	15%
TTM-r2.1	(L, H)	$(52, 16)$		$(90, 30)$	$(180, 60)$		$(360, 60)$		$(512, 48)$		$(512, 96)$		
	(L1), univariate	0.403		0.415	0.393		0.391		**0.423**		0.402		
	(L2), univariate	0.405		0.410	–		–		**0.422**		0.413		
	(L1), multivariate	0.401		0.456	0.464		0.437		0.468		**0.471**		
	(L2), multivariate	0.440		0.450	–		–		**0.468**		0.462		

C SENSITIVITY ANALYSES

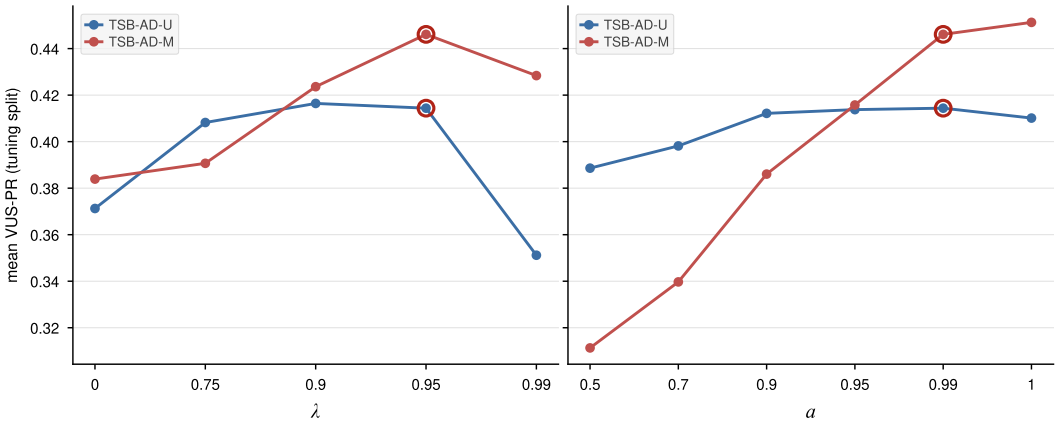


Figure 6: Sensitivity to the two global hyperparameters. The left panel varies λ with $a = 0.99$ fixed and the right panel varies a with $\lambda = 0.95$ fixed. Lines show tuning-split mean VUS-PR averaged over backbone–(L, H) combinations: 66 for TSB-AD-U and 46 for TSB-AD-M. Rings mark the values used in the reported experiments.

C.1 EWMA λ

We use a single decay factor, $\lambda = 0.95$, in every configuration. Among the grid values $\{0, 0.75, 0.90, 0.95, 0.99\}$, where $\lambda = 0$ corresponds to no smoothing, $\lambda = 0.95$ maximises mean VUS-PR when averaged over all backbone–(L, H) combinations from both tuning splits. The left panel of Figure 6 shows the benchmark-level sensitivity curves, and Figures 7 and 8 show the backbone-specific curves for both hyperparameters. On each benchmark, mean VUS-PR is lower at $\lambda = 0$ and $\lambda = 0.99$ than at $\lambda = 0.95$. Relative to $\lambda = 0.90$, $\lambda = 0.95$ lowers mean VUS-PR by 0.002 on TSB-AD-U and increases it by 0.022 on TSB-AD-M. Every backbone peaks at 0.90 or 0.95.

C.2 NORMALISATION QUANTILE a

We set $a = 0.99$ in every configuration, making $q_{a,\mathcal{T}}$ in Eq. (5) the empirical 0.99-quantile. Among the grid values $\{0.50, 0.70, 0.90, 0.95, 0.99, 1.00\}$, $a = 0.99$ maximises mean VUS-PR when averaged over all backbone–(L, H) combinations from both tuning splits. The right panel of Figure 6 shows the benchmark-level sensitivity curves. On each benchmark, mean VUS-PR is lower at smaller values of a than at $a = 0.99$. Relative to $a = 0.99$, $a = 1$ lowers mean VUS-PR by 0.004 on TSB-AD-U and increases it by 0.005 on TSB-AD-M. At $a = 1$, each normal reference level is the maximum over $\mathcal{T}$, so we retain $a = 0.99$ to reduce sensitivity to outliers.

C.3 FALLBACK

A series is scored through the fallback of Eq. (6) when no channel satisfies the three conditions of Section 4.3, which in almost every case is because the train segment ends before $t_0 = L + H$ and leaves the reference window $\mathcal{T}$ empty. Table 8 reports the fallback rate in each configuration and its effect on FDAS's gain over $\tilde{E}_1$. Fallback occurs only in configurations with $L = 512$. On TSB-AD-U, all eight configurations use the fallback on the same 78 series with an empty reference window, and each Chronos-Bolt configuration has three additional fallback series with zero normal reference levels. On TSB-AD-M, the two TTM-r2.1 configurations differ by two fallback series because their H and hence t_0 differ. FDAS and $\tilde{E}_1$ both use the unnormalised forecasting-error track E_1 on fallback series and therefore coincide. The mean gain on the remaining series is positive in every configuration, so including the fallback series moves the overall mean gain toward zero. The main text reports the gain including them, so the fallback does not inflate the reported improvement.

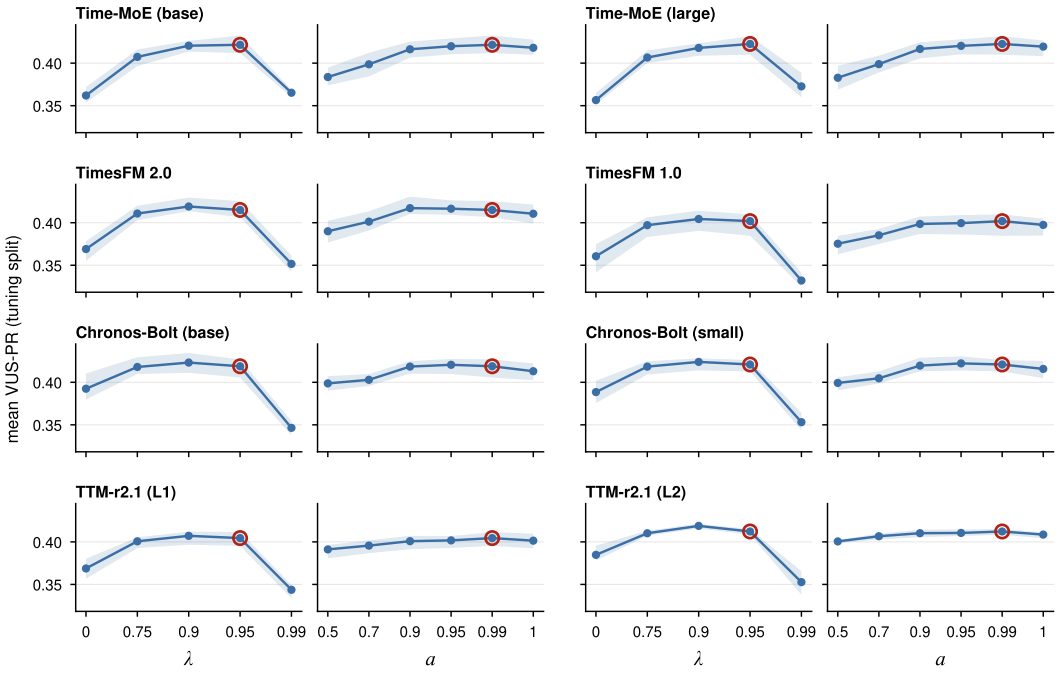


Figure 7: Sensitivity by backbone on **TSB-AD-U**. For each backbone, the left panel varies λ with $a = 0.99$ fixed and the right panel varies a with $\lambda = 0.95$ fixed, and all panels share the vertical axis. Lines and bands show the mean and interquartile range of tuning-split mean VUS-PR across that backbone's (L, H) combinations. Rings mark the values used in the reported experiments.

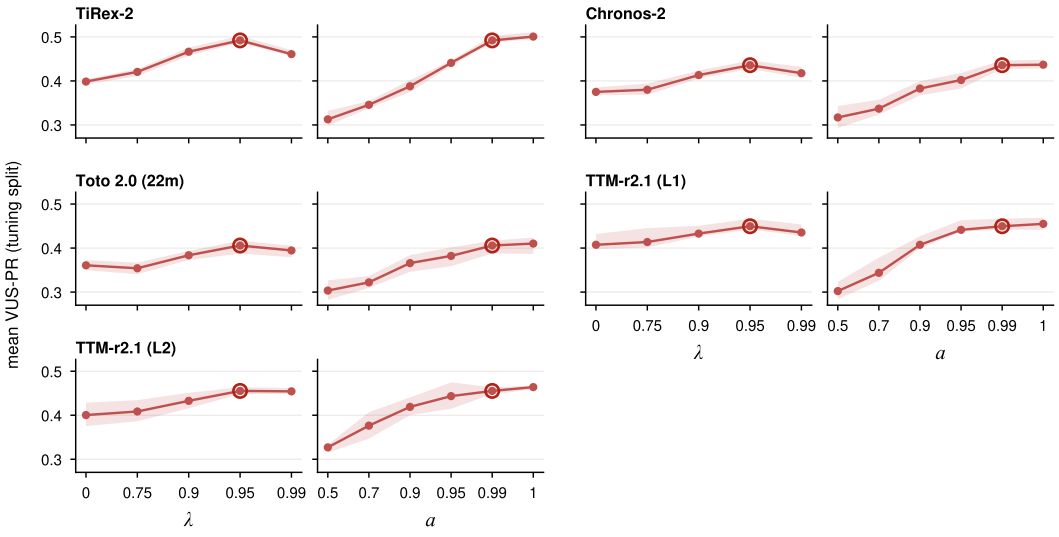


Figure 8: Sensitivity by backbone on **TSB-AD-M**. For each backbone, the left panel varies λ with $a = 0.99$ fixed and the right panel varies a with $\lambda = 0.95$ fixed, and all panels share the vertical axis. Lines and bands show the mean and interquartile range of tuning-split mean VUS-PR across that backbone's (L, H) combinations. Rings mark the values used in the reported experiments.

Table 8: Fallback on the evaluation split and its effect on FDAS's VUS-PR gain over $\tilde{E}_1$. Count and Share are the number and percentage of evaluation series scored through the fallback branch of Eq. (6). Gains are mean per-series VUS-PR differences, FDAS minus $\tilde{E}_1$, computed over all series and over the series without fallback. **Bold** marks the configurations where excluding the fallback series gives a larger gain.

		Fallback series		Gain (FDAS $-$ $\tilde{E}_1$)	
	TSFM (+FDAS)	Count	Share	All series	Excluding fallback series
Univariate	Time-MoE (base)	78	22%	0.0255	**0.0329**
	Time-MoE (large)	78	22%	0.0169	**0.0217**
	TimesFM 2.0	78	22%	0.0195	**0.0251**
	TimesFM 1.0	78	22%	0.0173	**0.0223**
	Chronos-Bolt (base)	81	23%	0.0249	**0.0323**
	Chronos-Bolt (small)	81	23%	0.0177	**0.0231**
	TTM-r2.1 (L1)	78	22%	0.0277	**0.0357**
	TTM-r2.1 (L2)	78	22%	0.0251	**0.0323**
Multivariate	TiRex-2	0	0%	0.0455	0.0455
	Chronos-2	0	0%	0.0296	0.0296
	Toto 2.0 (22m)	0	0%	0.0326	0.0326
	TTM-r2.1 (L1)	30	17%	0.0250	**0.0300**
	TTM-r2.1 (L2)	28	16%	0.0374	**0.0443**

D ADDITIONAL RESULTS

D.1 EXTENDED COMPARISON WITH PUBLISHED METHODS

Tables 9 and 10 repeat Tables 1 and 2 and add, for reference, the methods reported in the benchmark paper (Liu & Paparrizos, 2024), which fall outside the comparison criteria of Section 5.1. As in Tables 1 and 2, **bold** and underline mark the best and second-best displayed values.

Table 9: Extended results on **TSB-AD-U**, in the format of Table 1 with the Published reference block added. **Bold** and underline mark the best and second-best displayed values.

	Method		Range-Wise Metrics ↑			Point-Wise Metrics ↑		
			VUS-PR	VUS-ROC	R-based-F1	AUC-PR	AUC-ROC	Standard-F1
AD TSFMs	TimeRCD†		**0.52**	**0.87**	**0.55**	**0.50**	**0.84**	**0.53**
	CHARM†		0.50	–	–	–	–	–
	TSPulse†		0.48	–	–	–	–	–
	MOMENT†		0.38	0.75	0.36	0.30	0.68	0.35
Forecasting TSFMs	Time-MoE (base)	+ CS	0.29	0.74	0.34	0.29	0.67	0.34
		+ **FDAS**	0.50	0.82	0.45	0.39	0.76	0.44
	Time-MoE (large)	+ CS	0.30	0.74	0.35	0.29	0.67	0.34
		+ **FDAS**	0.49	0.83	0.45	0.38	0.77	0.44
	TimesFM 2.0	+ CS	0.30	0.75	0.35	0.29	0.68	0.35
		+ **FDAS**	0.48	0.81	0.44	0.38	0.76	0.44
	TimesFM 1.0	+ CS	0.30	0.75	0.35	0.29	0.69	0.34
		+ **FDAS**	0.47	0.82	0.44	0.36	0.77	0.43
	Chronos-Bolt (base)	+ CS	0.29	0.74	0.34	0.28	0.68	0.34
		+ **FDAS**	0.48	0.82	0.42	0.36	0.76	0.42
	Chronos-Bolt (small)	+ CS	0.29	0.74	0.35	0.28	0.67	0.34
		+ **FDAS**	0.46	0.81	0.42	0.35	0.76	0.41
	TTM-r2.1 (L1)	+ CS	0.30	0.74	0.34	0.28	0.67	0.33
		+ **FDAS**	0.46	0.80	0.42	0.34	0.73	0.40
	TTM-r2.1 (L2)	+ CS	0.30	0.74	0.34	0.28	0.67	0.33
		+ **FDAS**	0.45	0.80	0.41	0.34	0.73	0.39
	Lag-Llama†	+ CS	0.27	0.72	0.31	0.25	0.65	0.30
Published reference‡	Sub-PCA		0.42	0.76	0.41	0.37	0.71	0.42
	KShapeAD		0.40	0.76	0.40	0.35	0.74	0.39
	POLY		0.39	0.76	0.35	0.31	0.73	0.37
	Series2Graph		0.39	0.80	0.35	0.33	0.76	0.38
	MOMENT (FT)		0.39	0.76	0.35	0.30	0.69	0.35
	KMeansAD		0.37	0.76	0.38	0.32	0.74	0.37
	USAD		0.36	0.71	0.40	0.32	0.66	0.37
	Sub-KNN		0.35	0.79	0.32	0.27	0.76	0.34
	MatrixProfile		0.35	0.76	0.32	0.26	0.73	0.33
	SAND		0.34	0.76	0.36	0.29	0.73	0.35
	CNN		0.34	0.79	0.35	0.33	0.71	0.38
	LSTMAD		0.33	0.76	0.34	0.31	0.68	0.37
	SR		0.32	0.81	0.35	0.32	0.74	0.38
	IForest		0.30	0.78	0.30	0.29	0.71	0.35
	OmniAnomaly		0.29	0.72	0.29	0.27	0.65	0.31
	TimesNet		0.26	0.72	0.21	0.18	0.61	0.24
	AutoEncoder		0.26	0.69	0.28	0.19	0.63	0.25
	TranAD		0.26	0.68	0.25	0.20	0.57	0.25
	FITS		0.26	0.73	0.20	0.17	0.61	0.23
	Sub-LOF		0.25	0.73	0.25	0.16	0.68	0.24
	OFA		0.24	0.71	0.20	0.16	0.59	0.22
	Sub-MCD		0.24	0.72	0.24	0.15	0.67	0.23
	Sub-HBOS		0.23	0.67	0.27	0.18	0.61	0.23
	Sub-OCSVM		0.23	0.73	0.23	0.16	0.65	0.22
	Sub-IForest		0.22	0.72	0.23	0.16	0.63	0.22
	Donut		0.20	0.68	0.20	0.14	0.56	0.20
	LOF		0.17	0.68	0.22	0.14	0.58	0.21
	AnomalyTransformer		0.12	0.56	0.14	0.08	0.50	0.12
Base	Last-3		0.28	0.71	0.34	0.27	0.64	0.32
	Last-3^{EWMA}		0.41	0.76	0.36	0.29	0.69	0.35

† Taken from the TSB-AD leaderboard. Dashes mark unreported metrics.
‡ Taken from the benchmark paper (Liu & Paparrizos, 2024).

Table 10: Extended results on **TSB-AD-M**, in the format of Table 2 with the Published reference block added. **Bold** and underline mark the best and second-best displayed values.

	Method		Range-Wise Metrics ↑			Point-Wise Metrics ↑		
			VUS-PR	VUS-ROC	R-based-F1	AUC-PR	AUC-ROC	Standard-F1
AD TSFMs	CHARM†		0.39	–	–	–	–	–
	TSPulse†		0.36	–	–	–	–	–
	TimeRCD†		0.32	0.71	**0.43**	0.34	0.72	0.38
Forecasting TSFMs	TiRex-2	+ CS	0.21	0.67	0.25	0.20	0.62	0.24
		+ **FDAS**	**0.40**	0.75	0.32	0.34	0.69	0.39
	TTM-r2.1 (L2)¶	+ CS	0.24	0.70	0.25	0.23	0.65	0.27
		+ **FDAS**	**0.40**	**0.76**	0.35	**0.35**	0.72	**0.41**
	TTM-r2.1 (L1)¶	+ CS	0.24	0.67	0.25	0.23	0.62	0.27
		+ **FDAS**	**0.40**	0.75	0.35	**0.35**	0.71	**0.41**
	Chronos-2	+ CS	0.22	0.69	0.25	0.22	0.64	0.25
		+ **FDAS**	0.38	0.74	0.29	0.32	0.69	0.38
	Toto 2.0 (22m)	+ CS	0.21	0.67	0.25	0.21	0.62	0.24
		+ **FDAS**	0.37	0.74	0.28	0.32	0.70	0.37
Published reference‡‡	PCA		0.31	0.74	0.29	0.31	0.70	0.37
	OmniAnomaly		0.31	0.69	0.37	0.27	0.65	0.32
	CNN		0.31	**0.76**	0.37	0.32	**0.73**	0.37
	LSTMAD		0.31	0.74	0.38	0.31	0.70	0.36
	USAD		0.30	0.68	0.37	0.26	0.64	0.31
	AutoEncoder		0.30	0.69	0.28	0.30	0.67	0.34
	KMeansAD		0.29	0.73	0.33	0.25	0.69	0.31
	MCD		0.27	0.69	0.20	0.27	0.65	0.33
	CBLOF		0.27	0.70	0.31	0.28	0.67	0.32
	Donut		0.26	0.71	0.21	0.20	0.64	0.28
	OCSVM		0.26	0.67	0.30	0.23	0.61	0.28
	RobustPCA		0.24	0.61	0.33	0.24	0.58	0.29
	FITS		0.21	0.66	0.16	0.15	0.58	0.22
	OFA		0.21	0.63	0.17	0.15	0.55	0.21
	EIF		0.21	0.71	0.26	0.19	0.67	0.26
	IForest		0.20	0.69	0.24	0.19	0.66	0.26
	COPOD		0.20	0.69	0.24	0.20	0.65	0.27
	TimesNet		0.19	0.64	0.17	0.13	0.56	0.20
	HBOS		0.19	0.67	0.24	0.16	0.63	0.24
	TranAD		0.18	0.65	0.21	0.14	0.59	0.21
	KNN		0.18	0.59	0.21	0.14	0.51	0.19
	LOF		0.14	0.60	0.14	0.10	0.53	0.15
	AnomalyTransformer		0.12	0.57	0.14	0.07	0.52	0.12
Base	Last-3		0.21	0.63	0.25	0.21	0.58	0.25
	Last-3^{EWMA}		0.31	0.68	0.26	0.25	0.62	0.31

† Taken from the TSB-AD leaderboard. Dashes mark unreported metrics.
‡ Taken from the benchmark paper (Liu & Paparrizos, 2024).
¶ TTM-r2.1 forecasts channels independently.

D.2 RESULTS BY DATASET

Tables 11 and 12 report dataset-level mean VUS-PR for the methods in Tables 9 and 10. The TSPulse rows are taken from Ekambaram et al. (2026), and the TimeRCD rows are re-aggregated from the per-series results released with the leaderboard. CHARM releases only benchmark-level VUS-PR, so its rows show dashes.

Table 11: Mean VUS-PR for each dataset on **TSB-AD-U**, with the blocks of Table 9. **Bold** and underline mark the best and second-best displayed values in each column.

	Method		CATSv2	Daphnet	Exathlon	IOPS	LTDB	MGAB	MITDB	MSL	NAB	NEK	OPPORTUNITY	Power	SED	SMAP	SMD	SVDB	SWaT	Stock	TAO	TODS	UCR	WSD	YAHOO
AD TSFMs	TimeRCD†		0.13	0.19	0.89	0.46	0.34	0.05	0.33	0.44	0.45	0.77	0.52	0.26	0.52	0.45	0.62	0.21	0.23	0.77	0.92	**0.91**	0.36	0.25	**0.88**
	CHARM†		–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	TSPulse†		0.35	**0.55**	0.83	0.42	0.41	0.00	0.10	**0.64**	**0.50**	0.58	0.06	0.07	0.06	**0.71**	0.78	0.36	0.10	0.98	**1.00**	0.68	0.18	0.42	0.83
	MOMENT†		0.30	0.52	0.81	0.37	0.44	0.00	0.14	0.53	0.39	0.73	0.07	0.08	0.04	0.62	0.74	0.27	0.07	0.81	0.94	0.58	0.07	0.49	0.23
Forecasting TSFMs	Time-MoE (base)	+ CS	0.31	0.34	0.52	0.19	0.28	0.00	0.06	0.24	0.19	0.28	0.05	0.08	0.07	0.24	0.38	0.06	0.17	0.98	0.97	0.73	0.07	0.20	0.85
		+ **FDAS**	0.34	0.45	0.82	0.47	0.42	0.00	0.12	0.60	0.45	0.66	0.11	0.13	0.05	0.64	0.78	0.22	0.12	0.84	0.94	0.72	0.29	**0.55**	0.77
	Time-MoE (large)	+ CS	0.31	0.34	0.51	0.21	0.28	0.01	0.06	0.20	0.19	0.29	0.05	0.08	0.08	0.26	0.38	0.07	0.23	0.98	0.99	0.72	0.07	0.20	0.86
		+ **FDAS**	0.33	0.43	0.79	0.46	0.43	0.01	0.16	0.53	0.41	0.68	0.15	0.14	0.05	0.65	0.76	0.21	0.12	0.82	0.95	0.71	0.28	**0.55**	0.76
	TimesFM 2.0	+ CS	0.20	0.33	0.50	0.20	0.27	0.00	0.06	0.25	0.19	0.29	0.05	0.09	0.06	0.30	0.39	0.06	0.10	0.98	0.97	0.74	0.07	0.20	0.85
		+ **FDAS**	0.31	0.42	0.84	**0.48**	0.35	0.01	0.11	0.54	0.42	0.65	0.08	0.10	0.05	0.60	**0.80**	0.23	0.10	0.83	0.94	0.71	0.27	0.49	0.77
	TimesFM 1.0	+ CS	0.23	0.34	0.51	0.21	0.28	0.00	0.07	0.26	0.20	0.29	0.05	0.08	0.06	0.28	0.38	0.07	0.26	0.98	0.99	0.72	0.07	0.21	0.86
		+ **FDAS**	0.30	0.44	0.83	0.46	0.41	0.01	0.17	0.53	0.40	0.67	0.11	0.08	0.06	0.55	0.78	0.18	0.11	0.83	0.95	0.69	0.26	0.54	0.76
	Chronos-Bolt (base)	+ CS	0.16	0.33	0.51	0.20	0.27	0.00	0.07	0.28	0.19	0.29	0.05	0.08	0.07	0.26	0.38	0.06	0.09	0.98	0.97	0.68	0.07	0.20	0.82
		+ **FDAS**	0.34	0.45	0.86	0.47	0.44	0.00	0.09	0.56	0.42	0.66	0.07	0.14	0.05	0.63	0.79	0.24	0.10	0.83	0.94	0.69	0.25	0.46	0.77
	Chronos-Bolt (small)	+ CS	0.20	0.33	0.51	0.20	0.28	0.00	0.07	0.24	0.19	0.32	0.05	0.08	0.08	0.27	0.40	0.06	0.09	0.98	0.97	0.61	0.07	0.20	0.81
		+ **FDAS**	0.31	0.42	0.85	0.46	0.46	0.00	0.09	0.54	0.42	0.67	0.07	0.11	0.06	0.62	**0.80**	0.26	0.10	0.83	0.94	0.67	0.19	0.47	0.76
	TTM-r2.1 (L1)	+ CS	0.24	0.38	0.51	0.21	0.27	0.00	0.06	0.35	0.18	0.29	0.05	0.08	0.06	0.34	0.39	0.06	0.10	0.97	0.99	0.72	0.06	0.20	0.81
		+ **FDAS**	0.31	0.43	0.86	0.45	0.45	0.00	0.17	0.54	0.41	0.65	0.07	0.08	0.05	0.57	0.78	0.28	0.10	0.83	0.95	0.71	0.17	0.50	0.74
	TTM-r2.1 (L2)	+ CS	0.26	0.41	0.54	0.21	0.28	0.00	0.07	0.27	0.20	0.30	0.06	0.08	0.06	0.30	0.40	0.06	0.09	0.98	**1.00**	0.70	0.04	0.20	0.83
		+ **FDAS**	0.26	0.47	0.82	0.44	0.43	0.00	0.16	0.54	0.43	0.68	0.08	0.08	0.04	0.57	0.76	0.27	0.10	0.84	0.95	0.67	0.17	0.50	0.73
	Lag-Llama†	+ CS	0.21	0.39	0.53	0.22	0.29	0.00	0.08	0.31	0.18	0.38	0.05	0.08	0.07	0.28	0.36	0.08	0.09	0.97	0.99	0.61	0.02	0.22	0.68
Published reference‡	Sub-PCA		0.26	0.42	**0.93**	0.23	0.56	0.01	0.36	0.51	0.44	**0.91**	**0.91**	0.08	0.03	0.52	0.45	0.52	0.39	0.84	0.93	0.54	0.12	0.09	0.14
	KShapeAD		0.25	0.04	0.33	0.09	**0.83**	0.02	**0.69**	0.55	0.37	0.24	0.33	0.19	**0.89**	0.58	0.13	**0.82**	0.43	0.75	0.91	0.75	**0.38**	0.10	0.55
	POLY		0.23	0.51	0.74	0.31	0.51	0.01	0.34	0.54	0.48	0.61	0.10	0.09	0.04	0.64	0.61	0.44	0.10	0.82	0.92	0.57	0.13	0.41	0.25
	Series2Graph		0.21	0.19	0.60	0.22	0.79	0.00	0.61	0.25	0.44	0.67	0.11	0.07	0.15	0.55	0.46	0.55	0.22	0.79	0.91	0.73	0.25	0.27	0.28
	MOMENT (FT)		0.38	0.51	0.83	0.38	0.45	0.00	0.13	0.53	0.39	0.73	0.07	0.07	0.04	0.63	0.75	0.23	0.07	0.81	0.94	0.58	0.08	0.50	0.25
	KMeansAD		0.23	0.04	0.41	0.06	0.49	0.01	0.27	0.48	0.33	0.20	0.30	**0.39**	0.87	0.63	0.18	0.44	0.10	0.76	0.92	0.65	**0.38**	0.10	0.56
	USAD		**0.40**	0.12	0.89	0.13	0.55	0.00	0.18	0.27	0.28	0.73	0.67	0.06	0.03	0.27	0.66	0.43	0.37	0.75	0.93	0.52	0.08	0.04	0.10
	Sub-KNN		0.29	0.04	0.47	0.10	0.58	0.24	0.36	0.33	0.29	0.23	0.30	0.21	0.87	0.51	0.14	0.56	0.10	0.75	0.92	0.65	0.37	0.10	0.31
	MatrixProfile		0.36	0.04	0.56	0.10	0.58	0.29	0.39	0.48	0.32	0.13	0.25	0.15	0.72	0.47	0.13	0.36	0.11	0.72	0.92	0.76	0.34	0.02	0.43
	SAND		0.27	0.04	0.25	0.06	0.79	0.01	0.67	0.30	0.38	0.32	0.18	0.16	0.75	0.56	0.11	0.72	0.21	0.74	0.91	0.70	0.34	0.08	0.41
	CNN		0.32	0.40	0.60	0.26	0.42	0.01	0.15	0.33	0.19	0.73	0.40	0.08	0.06	0.34	0.55	0.21	**0.68**	0.92	**1.00**	0.54	0.05	0.24	0.53
	LSTMAD		0.33	0.13	0.73	0.20	0.36	0.03	0.12	0.32	0.18	0.73	0.58	0.07	0.06	0.26	0.49	0.13	0.67	0.85	**1.00**	0.47	0.02	0.13	0.45
	SR		0.28	0.20	0.73	0.24	0.29	0.01	0.07	0.22	0.20	0.50	0.33	0.10	0.07	0.29	0.36	0.08	0.35	**1.00**	**1.00**	0.64	0.07	0.22	0.61
	IForest		0.08	0.36	0.67	0.28	0.34	0.00	0.10	0.29	0.22	0.59	0.43	0.08	0.36	0.25	0.34	0.09	0.50	0.99	0.99	0.52	0.02	0.14	0.44
	OmniAnomaly		0.12	0.16	0.83	0.20	0.32	0.00	0.10	0.25	0.19	0.85	0.60	0.07	0.06	0.15	0.36	0.09	0.44	0.82	0.98	0.44	0.03	0.14	0.19
	TimesNet		0.10	0.39	0.53	0.22	0.29	0.00	0.08	0.31	0.20	0.37	0.05	0.08	0.05	0.38	0.54	0.09	0.11	0.79	0.91	0.59	0.02	0.27	0.29
	AutoEncoder		0.18	0.09	0.36	0.25	0.69	0.01	0.07	0.28	0.32	0.51	0.12	0.09	0.41	0.49	0.14	0.32	0.38	0.72	0.93	0.65	0.09	0.14	0.29
	TranAD		0.08	0.13	0.72	0.18	0.31	0.00	0.09	0.18	0.18	0.72	0.58	0.07	0.05	0.13	0.16	0.09	0.46	0.79	0.94	0.45	0.02	0.11	0.28
	FITS		0.17	0.43	0.55	0.17	0.34	0.00	0.09	0.36	0.24	0.49	0.07	0.07	0.05	0.42	0.54	0.10	0.10	0.76	0.91	0.58	0.02	0.14	0.18
	Sub-LOF		0.31	0.04	0.25	0.11	0.34	**0.44**	0.26	0.35	0.32	0.25	0.12	0.14	0.22	0.40	0.04	0.18	0.11	0.76	0.92	0.53	0.29	0.03	0.27
	OFA		0.16	0.36	0.55	0.20	0.30	0.00	0.07	0.29	0.21	0.37	0.05	0.08	0.06	0.33	0.45	0.07	0.11	0.76	0.91	0.54	0.02	0.16	0.24
	Sub-MCD		0.37	0.04	0.23	0.13	0.24	0.01	0.11	0.16	0.19	0.11	0.32	0.30	0.12	0.30	0.08	0.07	0.09	0.75	0.90	0.64	0.26	0.15	0.28
	Sub-HBOS		0.04	0.05	0.45	0.05	0.69	0.00	0.17	0.25	0.30	0.23	0.08	0.12	0.88	0.55	0.10	0.24	0.12	0.70	0.93	0.64	0.14	0.01	0.06
	Sub-OCSVM		0.26	0.06	0.29	0.07	0.33	0.01	0.14	0.28	0.26	0.26	0.11	0.16	0.06	0.51	0.08	0.20	0.09	0.73	0.92	0.65	0.18	0.03	0.23
	Sub-IForest		0.05	0.07	0.49	0.04	0.66	0.00	0.24	0.36	0.30	0.22	0.07	0.12	0.79	0.47	0.09	0.27	0.13	0.69	0.90	0.66	0.10	0.01	0.06
	Donut		0.08	0.06	0.45	0.10	0.31	0.00	0.10	0.20	0.18	0.47	0.18	0.09	0.14	0.31	0.28	0.08	0.47	0.78	0.91	0.48	0.01	0.06	0.12
	LOF		0.06	0.13	0.20	0.12	0.26	0.00	0.06	0.15	0.17	0.38	0.14	0.09	0.11	0.15	0.13	0.05	0.12	0.75	0.91	0.49	0.02	0.09	0.37
	AnomalyTransformer		0.05	0.07	0.13	0.06	0.27	0.00	0.09	0.14	0.14	0.23	0.07	0.09	0.09	0.09	0.18	0.07	0.10	0.75	0.90	0.46	0.01	0.02	0.07
Base	Last-3		0.10	0.41	0.51	0.25	0.27	0.00	0.06	0.17	0.16	0.36	0.05	0.08	0.08	0.12	0.39	0.06	0.12	0.97	**1.00**	0.77	0.03	0.26	0.77
	Last-3^{EWMA}		0.33	0.50	0.80	0.47	0.26	0.00	0.07	0.49	0.32	0.76	0.05	0.07	0.06	0.57	0.79	0.07	0.08	0.83	0.95	0.76	0.10	0.52	0.60

† Taken from the TSB-AD leaderboard, except that the TSPulse rows are from Ekambaram et al. (2026). Dashes mark unavailable dataset results.
‡ Taken from the benchmark paper (Liu & Paparrizos, 2024).

Table 12: Mean VUS-PR for each dataset on **TSB-AD-M**, with the blocks of Table 10. **Bold** and underline mark the best and second-best displayed values in each column.

	Method		CATSv2	CreditCard	Daphnet	Exathlon	GECCO	GHL	Genesis	LTDB	MITDB	MSL	OPPORTUNITY	PSM	SMAP	SMD	SVDB	SWaT	TAO
AD TSFMs	CHARM†		–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	TSPulse†		0.05	0.00	0.35	0.89	0.17	0.01	0.01	0.36	0.07	0.20	0.07	0.14	0.30	0.35	**0.38**	0.13	0.93
	TimeRCD†		0.23	0.04	0.15	0.93	0.08	0.02	0.00	0.21	**0.24**	0.23	0.28	0.22	0.28	0.46	0.15	0.17	0.19
Forecasting TSFMs	TiRex-2	+ CS	0.05	0.02	0.31	0.51	0.03	0.01	0.05	0.22	0.04	0.13	0.05	0.13	0.12	0.18	0.08	0.13	**1.00**
		+ **FDAS**	0.32	0.04	0.33	0.82	0.47	0.01	0.65	0.37	0.21	0.29	0.15	0.17	0.55	0.39	0.31	0.23	0.81
	TTM-r2.1 (L2)¶	+ CS	0.09	0.02	0.35	0.67	0.03	0.01	0.07	0.23	0.04	0.19	0.08	0.14	0.12	0.21	0.09	0.11	0.98
		+ **FDAS**	0.23	0.02	0.31	0.80	0.47	0.01	0.63	0.38	0.17	0.30	0.15	0.16	**0.59**	0.37	0.29	0.25	0.88
	TTM-r2.1 (L1)¶	+ CS	0.05	0.02	0.35	0.59	0.03	0.01	0.07	0.23	0.05	0.24	0.06	0.14	0.18	0.21	0.09	0.14	0.97
		+ **FDAS**	**0.42**	0.02	0.32	0.81	**0.49**	0.01	0.62	0.36	0.23	0.25	0.15	0.15	0.55	0.36	0.28	0.24	0.88
	Chronos-2	+ CS	0.06	0.02	0.30	0.55	0.03	0.01	0.08	0.24	0.04	0.19	0.06	0.13	0.13	0.19	0.10	0.12	0.99
		+ **FDAS**	0.39	0.03	0.28	0.91	0.47	0.01	0.62	0.38	0.15	0.21	0.15	0.18	0.46	0.27	0.29	0.21	0.79
	Toto 2.0 (22m)	+ CS	0.05	0.02	0.30	0.51	0.03	0.01	0.05	0.23	0.04	0.17	0.06	0.13	0.13	0.18	0.09	0.13	0.99
		+ **FDAS**	0.39	0.02	0.28	0.85	0.45	0.01	0.51	0.40	0.12	0.21	0.15	0.18	0.47	0.27	0.32	0.21	0.80
Published reference‡	PCA		0.12	**0.10**	0.13	**0.95**	0.20	0.01	0.02	0.24	0.07	0.15	**0.30**	0.16	0.09	0.36	0.11	0.45	**1.00**
	OmniAnomaly		0.04	0.02	0.34	0.84	0.02	**0.07**	0.00	**0.44**	0.11	0.22	0.18	0.16	0.12	0.17	0.35	0.15	0.81
	CNN		0.08	0.02	0.21	0.68	0.03	0.02	0.10	0.33	0.14	0.35	0.16	0.22	0.19	0.35	0.19	0.41	**1.00**
	LSTMAD		0.04	0.02	0.31	0.82	0.02	0.06	0.04	0.30	0.09	0.22	0.17	0.24	0.16	0.33	0.15	0.16	0.99
	USAD		0.04	0.02	0.34	0.84	0.02	**0.07**	0.00	0.41	0.12	0.23	0.18	0.19	0.11	0.16	0.32	0.15	0.81
	AutoEncoder		0.06	0.03	0.13	0.91	0.05	0.05	0.01	0.21	0.04	0.22	0.14	**0.28**	0.13	0.30	0.07	**0.58**	**1.00**
	KMeansAD		0.12	0.02	0.30	0.37	0.06	0.03	**0.89**	0.41	0.06	**0.44**	0.06	0.21	0.38	0.36	0.20	0.16	0.86
	MCD		0.13	0.06	0.14	0.80	0.03	0.01	0.06	0.21	0.04	0.23	0.17	0.26	0.10	0.26	0.07	0.54	**1.00**
	CBLOF		0.06	0.03	0.10	0.86	0.03	0.02	0.02	0.20	0.04	0.21	0.14	0.19	0.14	0.22	0.07	0.29	**1.00**
	Donut		0.07	0.02	0.17	0.66	0.03	0.05	0.18	0.26	0.12	0.30	0.15	0.20	0.18	0.19	0.11	0.44	0.75
	OCSVM		0.08	0.02	0.06	0.83	0.04	0.04	0.08	0.20	0.04	0.22	0.12	0.19	0.12	0.28	0.07	0.44	0.81
	RobustPCA		0.04	0.02	0.06	0.77	0.02	0.03	0.00	0.23	0.04	0.22	0.13	0.12	0.07	0.10	0.08	0.12	**1.00**
	FITS		0.13	0.02	0.33	0.63	0.03	0.01	0.10	0.23	0.05	0.17	0.05	0.13	0.08	0.17	0.10	0.15	0.78
	OFA		0.13	0.02	0.31	0.58	0.04	0.01	0.22	0.29	0.06	0.14	0.05	0.17	0.08	0.17	0.12	0.12	0.78
	EIF		0.06	0.02	0.15	0.41	0.04	0.02	0.06	0.19	0.04	0.18	0.10	0.18	0.13	0.32	0.07	0.32	0.89
	IForest		0.06	0.03	0.13	0.35	0.04	0.05	0.08	0.21	0.04	0.21	0.18	0.19	0.09	0.26	0.07	0.39	0.93
	COPOD		0.05	0.05	0.11	0.40	0.04	0.03	0.08	0.21	0.04	0.21	0.17	0.20	0.10	0.19	0.07	0.31	0.99
	TimesNet		0.07	0.02	0.27	0.42	0.04	0.01	0.02	0.27	0.07	0.17	0.06	0.14	0.09	0.14	0.11	0.14	0.79
	HBOS		0.05	0.04	0.15	0.32	0.04	0.04	0.08	0.21	0.04	0.23	0.17	0.17	0.09	0.25	0.07	0.30	0.83
	TranAD		0.04	0.02	0.31	0.10	0.02	0.06	0.04	0.26	0.07	0.24	0.16	0.23	0.09	0.30	0.12	0.15	0.81
	KNN		0.07	0.02	0.25	0.33	0.11	0.01	0.04	0.19	0.04	0.18	0.06	0.12	0.12	0.30	0.06	0.11	0.78
	LOF		0.05	0.02	0.11	0.16	0.13	0.01	0.08	0.19	0.04	0.14	0.10	0.15	0.09	0.16	0.06	0.15	0.79
	AnomalyTransformer		0.03	0.02	0.07	0.10	0.02	0.03	0.01	0.21	0.05	0.12	0.07	0.21	0.06	0.07	0.08	0.18	0.77
Base	Last-3		0.07	0.02	0.36	0.55	0.03	0.01	0.09	0.22	0.04	0.13	0.06	0.14	0.08	0.23	0.08	0.11	**1.00**
	Last-3^{EWMA}		0.14	0.02	**0.37**	0.80	0.04	0.01	0.64	0.20	0.04	0.34	0.06	0.15	0.20	**0.49**	0.11	0.15	0.88

† Taken from the TSB-AD leaderboard, except that the TSPulse rows are from Ekambaram et al. (2026). Dashes mark unavailable dataset results.
‡ Taken from the benchmark paper (Liu & Paparrizos, 2024).
¶ TTM-r2.1 forecasts channels independently.

D.3 PAIRED SIGNIFICANCE TESTS

Figure 9 compares FDAS with conventional scoring, Last-3, and Last-3$^{\mathrm{EWMA}}$ in each of the 13 configurations, using per-series VUS-PR differences of FDAS minus the comparator. We report their mean with a 95% confidence interval from 10,000 paired bootstrap resamples of series, together with a paired Wilcoxon signed-rank p-value. All 13 configurations give intervals above zero and p-values below 0.05 against all three comparators. The largest of these 39 p-values is 0.0006, below the Bonferroni threshold of $0.05/39 \approx 0.0013$. Thus, the gains are supported by paired per-series comparisons, not only by the benchmark means of Section 5.2.

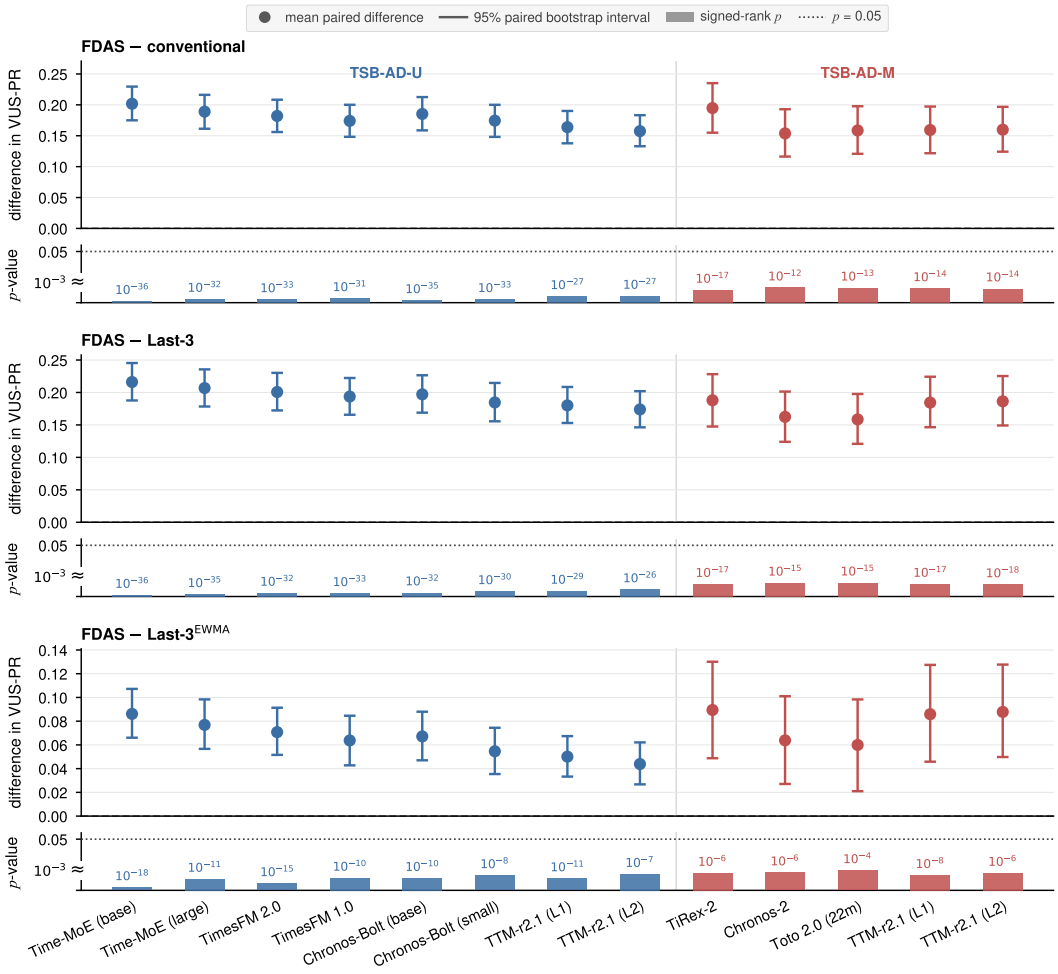


Figure 9: Paired comparisons of VUS-PR across the 13 configurations. Each row shows the mean per-series difference, FDAS minus the comparator, with its 95% paired bootstrap confidence interval, on a vertical axis scaled to that row. The strip below gives the unadjusted paired Wilcoxon signed-rank p-value of each configuration with a reference line at $p = 0.05$. The p-value axis is linear above 10^{-3} and logarithmic below it, and each bar is labelled with the order of magnitude of its p-value.

D.4 PER-SERIES COMPARISON FOR ALL CONFIGURATIONS

Figures 10 and 11 extend the per-series comparisons of Figures 3(a–b) and 4(a–b) to all 13 configurations. Each configuration occupies two panels and each point is one evaluation series, with VUS-PR under FDAS on the horizontal axis and under Last-3 or conventional scoring on the vertical axis. Differences within ± 0.1 are treated as ties (Zhu et al., 2026), and each legend gives the percentage of series in the resulting three groups. Ties account for 38–49% of series against Last-3 and 41–53% against conventional scoring. In all 26 panels, more series favour FDAS than the comparator.

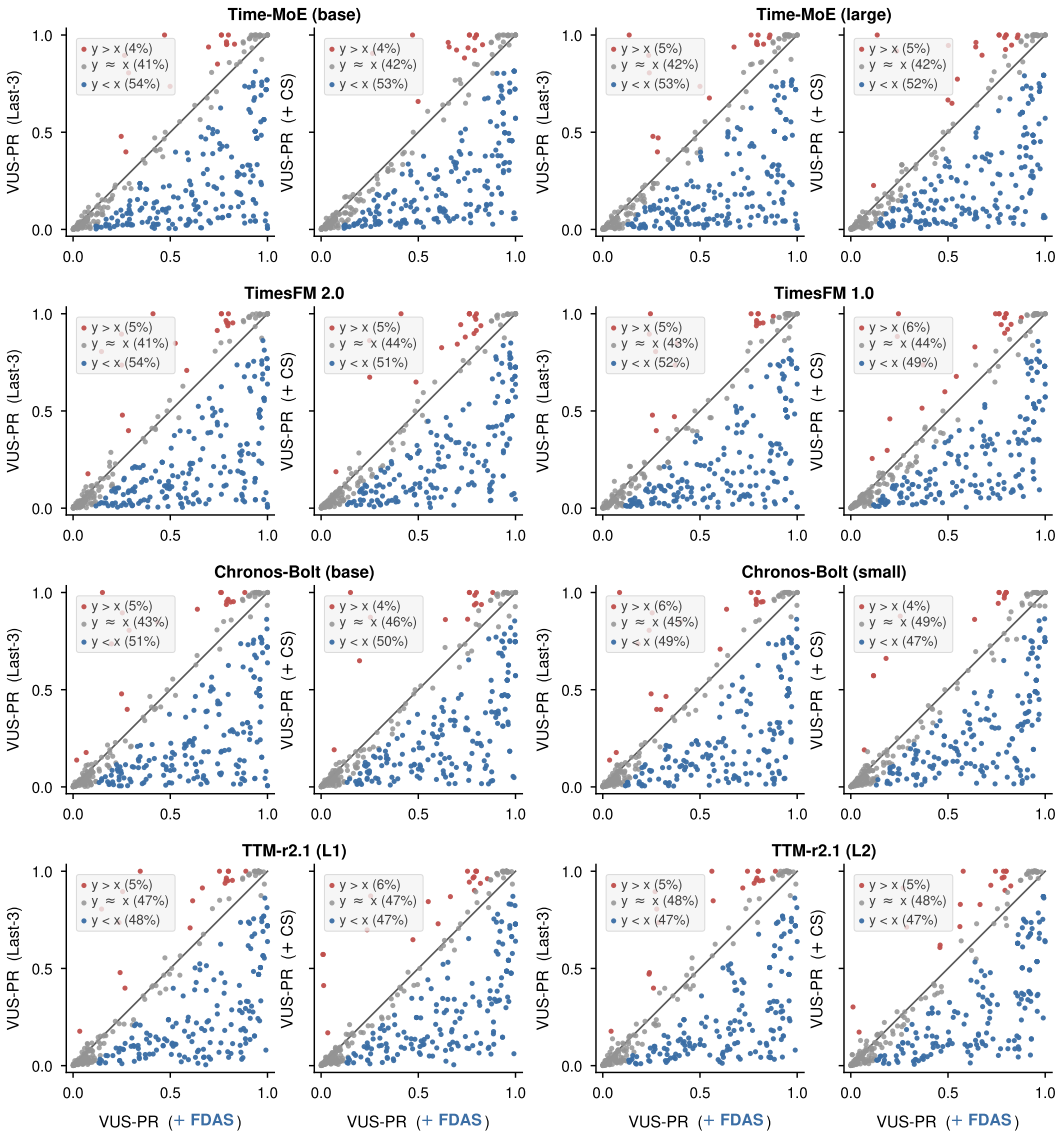


Figure 10: Per-series VUS-PR on **TSB-AD-U**, with FDAS on the horizontal axis. In each pair of panels the vertical axis is Last-3 on the left and conventional scoring on the right, so points below the diagonal favour FDAS. Ties are differences within ± 0.1.

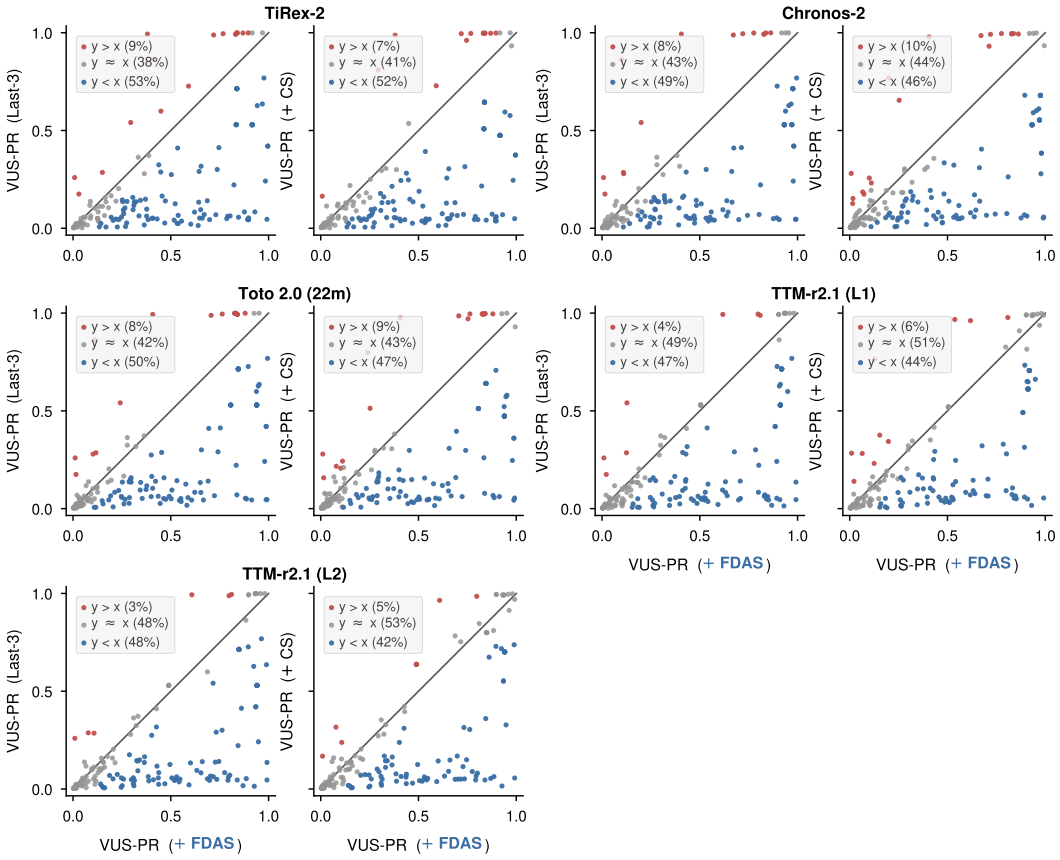


Figure 11: Per-series VUS-PR on **TSB-AD-M**, with FDAS on the horizontal axis. In each pair of panels the vertical axis is Last-3 on the left and conventional scoring on the right, so points below the diagonal favour FDAS. Ties are differences within ± 0.1.

D.5 SCORE VISUALISATION

Figures 12 and 13 show score traces on twelve series from each benchmark.

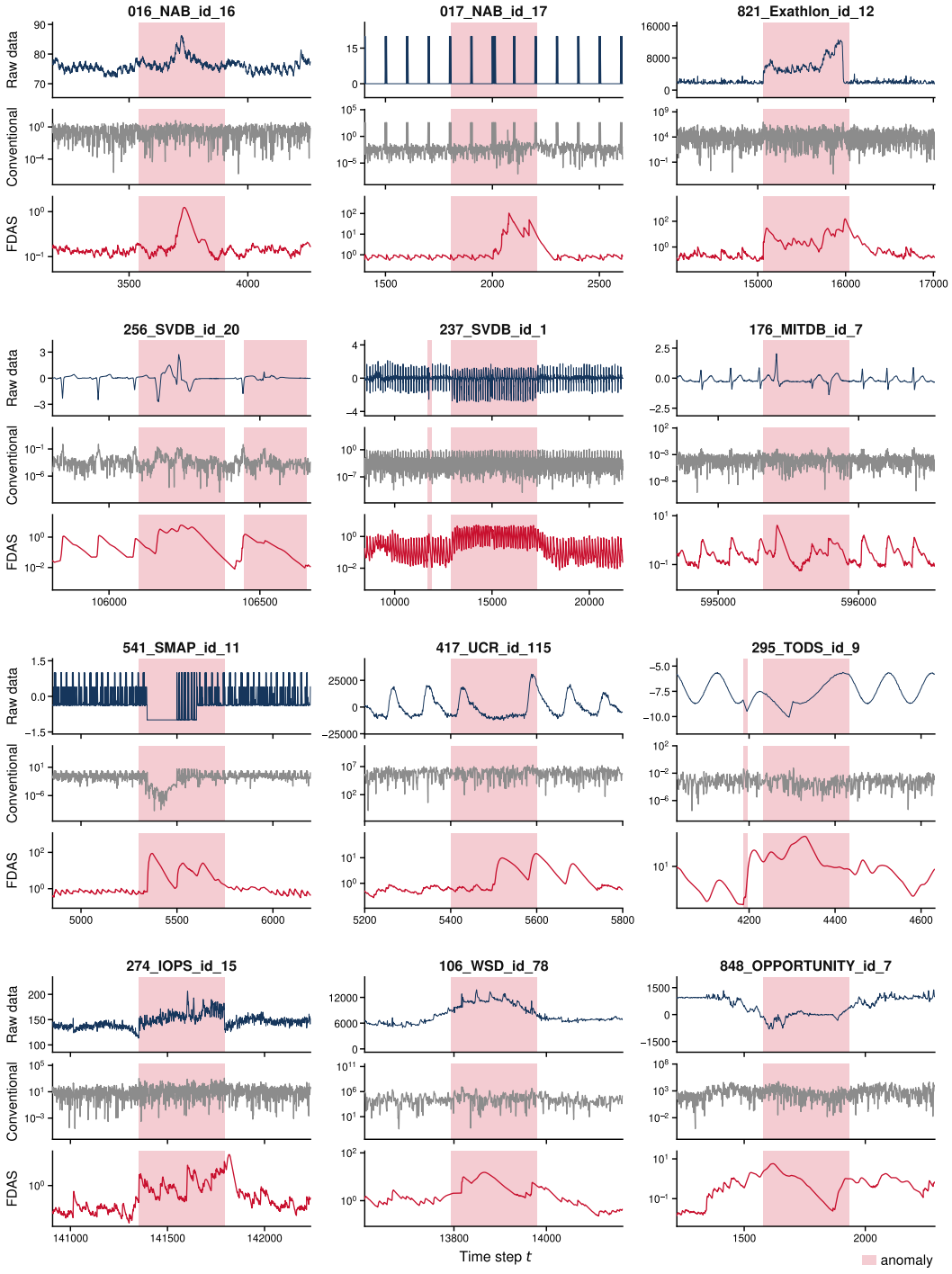


Figure 12: Score traces on twelve **TSB-AD-U** series with Time-MoE base. Each block stacks the observed series as provided by the benchmark, the conventional score, and FDAS on one time axis, with labelled anomalies shaded. The two score rows use a logarithmic vertical axis.

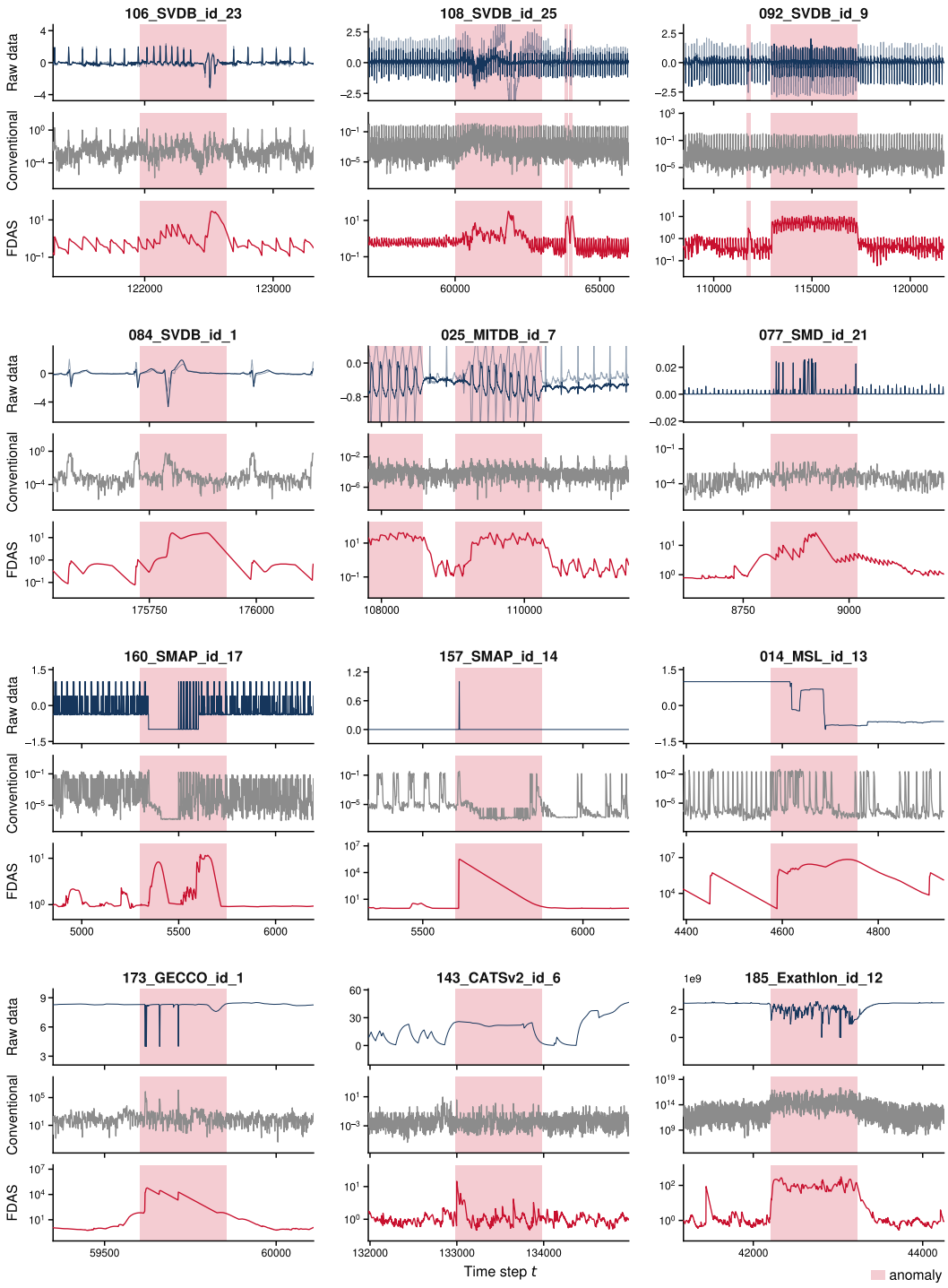


Figure 13: Score traces on twelve **TSB-AD-M** series with TiRex-2, in the layout of Figure 12. The observed row highlights a channel contributing to the FDAS score during the anomaly and uses its scale. For series with at most three channels, it also shows the others as translucent lines. The FDAS row is the reported score after the channel maximum.

D.6 COMPARISON WITH $\mathcal{W}_1$-ACAS

$\mathcal{W}_1$-ACAS reports dataset-level averages for seven TSB-AD-U datasets and four TSB-AD-M datasets rather than the full benchmarks (Martinez Gil et al., 2026). Tables 13 and 14 compare these published results with FDAS's mean VUS-PR and AUC-PR on the same series, for each dataset. For the multivariate comparison, we use the FDAS configurations of the same backbone families, Chronos-2 and TiRex-2.

Table 13: Comparison with $\mathcal{W}_1$-ACAS on seven **TSB-AD-U** datasets. Mean denotes the equally weighted average across datasets. $\mathcal{W}_1$-ACAS uses its published configuration, with lookback length 52, horizon 15, and its own scoring. Shaded FDAS rows use this paper's configurations on the same series.

	Method	Forecasting TSFM	YAHOO	NEK	NAB	MSL	IOPS	Stock	WSD	Mean
VUS-PR	$\mathcal{W}_1$-ACAS†	Chronos-Bolt (small)	0.68	0.45	0.22	0.40	0.29	1.00	0.23	0.47
	$\mathcal{W}_1$-ACAS†	TTM	0.68	0.43	0.23	0.40	0.30	0.99	0.24	0.46
	FDAS	Chronos-Bolt (small)	0.76	0.67	0.42	0.54	0.46	0.83	0.47	0.59
	FDAS	TTM-r2.1 (L1)	0.74	0.65	0.41	0.54	0.45	0.83	0.50	0.59
	FDAS	TTM-r2.1 (L2)	0.73	0.68	0.43	0.54	0.44	0.84	0.50	0.59
AUC-PR	$\mathcal{W}_1$-ACAS†	Chronos-Bolt (small)	0.33	0.41	0.20	0.28	0.16	0.97	0.22	0.37
	$\mathcal{W}_1$-ACAS†	TTM	0.28	0.38	0.22	0.23	0.17	0.96	0.23	0.35
	FDAS	Chronos-Bolt (small)	0.32	0.63	0.37	0.45	0.25	0.20	0.32	0.36
	FDAS	TTM-r2.1 (L1)	0.16	0.62	0.38	0.43	0.24	0.17	0.34	0.33
	FDAS	TTM-r2.1 (L2)	0.15	0.63	0.41	0.43	0.27	0.17	0.34	0.34

† Quoted from Martinez Gil et al. (2026), Tables 2–4. TTM there is the release its authors used.

Table 14: Comparison with $\mathcal{W}_1$-ACAS on four **TSB-AD-M** datasets. Mean denotes the equally weighted average across datasets. $\mathcal{W}_1$-ACAS-F and $\mathcal{W}_1$-ACAS-H are its two multivariate variants, with lookback length 52, horizon 15, and its own scoring. Shaded FDAS rows use this paper's configurations on the same series, which include the tuning-split series of TAO and LTDB.

	Method	Forecasting TSFM	TAO	GECCO	LTDB	Genesis	Mean
VUS-PR	$\mathcal{W}_1$-ACAS-F†	Chronos	1.00	0.24	0.27	0.07	0.39
	$\mathcal{W}_1$-ACAS-H†	Chronos	1.00	0.13	0.26	0.07	0.36
	$\mathcal{W}_1$-ACAS-F†	TiRex	1.00	0.24	0.26	0.03	0.38
	$\mathcal{W}_1$-ACAS-H†	TiRex	1.00	0.14	0.26	0.05	0.36
	FDAS	Chronos-2	0.82	0.47	0.44	0.62	0.59
	FDAS	TiRex-2	0.83	0.47	0.45	0.65	0.60
AUC-PR	$\mathcal{W}_1$-ACAS-F†	Chronos	0.90	0.23	0.23	0.08	0.36
	$\mathcal{W}_1$-ACAS-H†	Chronos	0.95	0.11	0.17	0.08	0.33
	$\mathcal{W}_1$-ACAS-F†	TiRex	0.91	0.24	0.23	0.04	0.35
	$\mathcal{W}_1$-ACAS-H†	TiRex	0.94	0.12	0.17	0.06	0.32
	FDAS	Chronos-2	0.11	0.49	0.38	0.10	0.27
	FDAS	TiRex-2	0.12	0.49	0.36	0.16	0.28

† Quoted from Martinez Gil et al. (2026), Table 6. Chronos and TiRex there are the releases its authors used.

E ABLATIONS AND ADDITIONAL ANALYSES

E.1 TRACK ABLATION

Figure 14 reports the mean per-series VUS-PR difference between FDAS and each individual track, with paired bootstrap confidence intervals and Wilcoxon signed-rank tests. Since removing one track leaves the other, the difference from $\tilde{E}_1$ is the contribution of the disagreement track and the difference from $\tilde{D}$ that of the forecasting-error track.

The forecasting-error track contributes more in every univariate configuration, and the disagreement track contributes more in every multivariate configuration except TTM-r2.1 (L1). At the unadjusted 0.05 level, the contribution of the forecasting-error track is significant in all 13 configurations, and that of the disagreement track in four of the eight univariate configurations and four of the five multivariate configurations. Appendix E.4 examines this difference in terms of anomaly length.

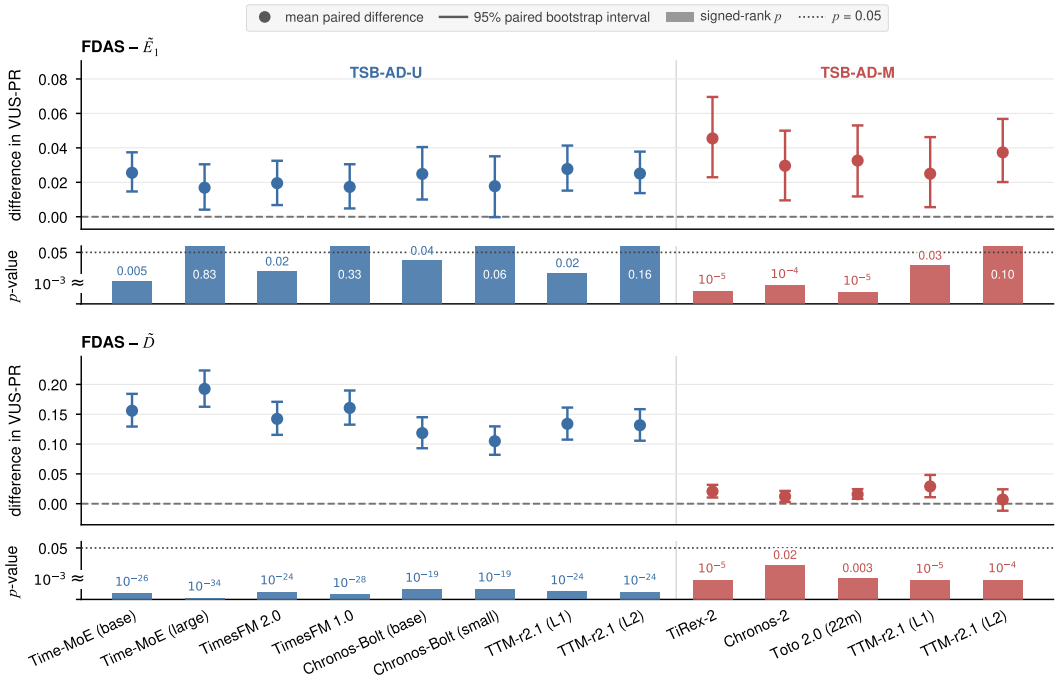


Figure 14: Effect of track removal. Mean per-series VUS-PR differences, FDAS minus each individual track of Table 3, $\tilde{E}_1$ and $\tilde{D}$, in the layout of Figure 9, except that bars are labelled with the p-value itself above 10^{-3}. Bars for p-values above 0.06 are clipped at 0.06 and labelled inside the bar.

E.2 TRACK AGGREGATION ANALYSIS

Table 15 compares the maximum in Eq. (6) with three alternatives applied to the same two normalised tracks: their arithmetic mean, their product, and their minimum. Section 4.3 takes the maximum so that a high score in either track is preserved even when the other is low. The minimum requires both tracks to be high at the same time and is the weakest rule in every configuration. The product, which also lets a low track pull the score down, sits between the minimum and the mean. Averaged across configurations, the arithmetic mean is within 0.003 of the maximum on each benchmark. Appendix E.1 shows that the track with the larger contribution differs between the two benchmarks, and the maximum follows whichever track responds.

Table 15: Mean VUS-PR on the evaluation splits for four ways of combining the two normalised tracks. FDAS uses the maximum, as in Eq. (6).

	TSFM	maximum	mean	product	minimum
Univariate	Time-MoE (base)	0.50	0.49	0.47	0.40
	Time-MoE (large)	0.49	0.49	0.45	0.38
	TimesFM 2.0	0.48	0.47	0.45	0.40
	TimesFM 1.0	0.47	0.47	0.44	0.38
	Chronos-Bolt (base)	0.48	0.47	0.46	0.41
	Chronos-Bolt (small)	0.46	0.46	0.45	0.40
	TTM-r2.1 (L1)	0.46	0.46	0.44	0.38
	TTM-r2.1 (L2)	0.45	0.45	0.43	0.38
Multivariate	TiRex-2	0.40	0.40	0.38	0.35
	TTM-r2.1 (L2)	0.40	0.40	0.39	0.36
	TTM-r2.1 (L1)	0.40	0.40	0.40	0.35
	Chronos-2	0.38	0.38	0.37	0.34
	Toto 2.0 (22m)	0.37	0.37	0.36	0.33

E.3 CHANNEL AGGREGATION ANALYSIS

Table 16 compares the channel maximum in Eq. (8) with two alternatives on TSB-AD-M: the mean and the median across channels. The set of alternatives differs from Appendix E.2 because the aggregation is over a variable number of channels rather than two tracks. The median coincides with the mean for two values, while a product over channels depends on how many there are and a minimum requires every channel to respond.

Section 4.4 takes the maximum across channels because it preserves high scores that occur in only a subset of channels. The median gives the lowest mean VUS-PR in all five configurations. The channel mean exceeds the maximum by 0.007 on average. A high score confined to one channel is preserved by the maximum, diluted by the mean, and suppressed by the median when there are at least three channels.

Table 16: Mean VUS-PR on the **TSB-AD-M** evaluation split for three ways of aggregating the per-channel scores. FDAS uses the maximum, as in Eq. (8).

TSFM	maximum	mean	median
TiRex-2	0.40	0.41	0.32
TTM-r2.1 (L2)	0.40	0.41	0.30
TTM-r2.1 (L1)	0.40	0.41	0.30
Chronos-2	0.38	0.38	0.32
Toto 2.0 (22m)	0.37	0.38	0.31

E.4 Anomaly length and the role of each track

Figure 15 relates the VUS-PR difference between the two tracks, $\tilde{D}$ minus $\tilde{E}_1$, to average anomaly length per series. On both benchmarks, the difference tends to increase with average anomaly length. Spearman correlations are positive in all 13 configurations, ranging from 0.44 to 0.53 on TSB-AD-U and from 0.13 to 0.47 on TSB-AD-M. An anomalous observation enters the lookback windows only for later forecasts, so any resulting rise in disagreement may fall outside a short labelled anomaly. In contrast, $e_{1,t}$ uses the current observation and may rise at anomaly onset. The median average anomaly length is 116 time steps on TSB-AD-U and 241 on TSB-AD-M, which is consistent with the larger contribution of the disagreement track on the latter in Appendix E.1.

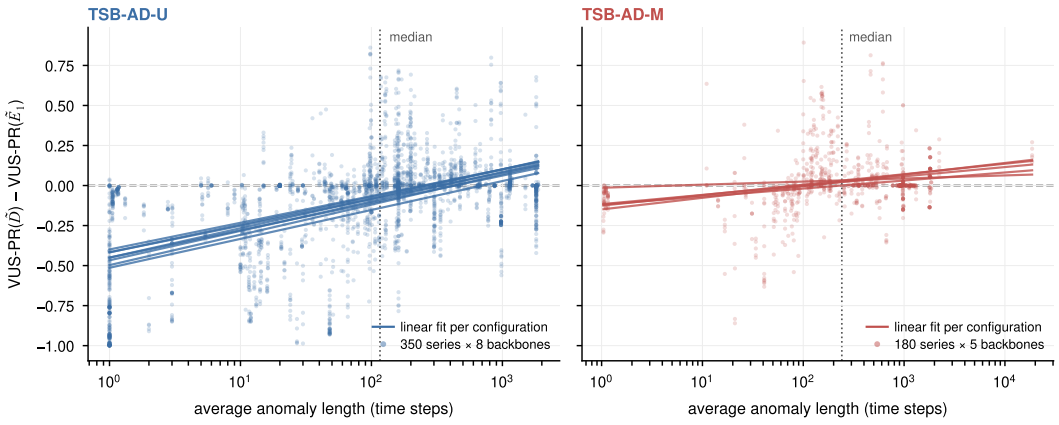


Figure 15: Anomaly length and relative track performance. VUS-PR difference between the two smoothed tracks, $\tilde{D}$ minus $\tilde{E}_1$, plotted against average anomaly length per series for every evaluation series of all 13 configurations, with a linear fit per configuration in $\log_{10}$ length. Dotted lines mark the median anomaly length of each benchmark.

F Limitations

Normalisation. Section 3 assumes that the train segment contains no anomalies. Anomalies in the reference window may increase a track's normal reference level and thereby reduce its normalised scores. We do not measure how detection performance changes with contamination of the train segment.

Train-segment causality. Following the benchmark protocol, we evaluate scores over the entire series, including the train segment. When $\mathcal{T}$ is nonempty, scores for $t_0 \leq t < t_{\mathrm{tr}}$ use normal reference levels computed from later train-segment observations. Scores at $t < t_0$ are set to s_{t_0} and therefore also depend on later observations. Scoring is causal only for $t \geq \max(t_0, t_{\mathrm{tr}})$, as stated in Section 4.4.

Short train segments. When the train segment ends before $t_0 = L + H$, the reference window $\mathcal{T}$ is empty and FDAS falls back to the smoothed but unnormalised forecasting-error track, so the disagreement track does not contribute for these series. This affects 78 of the 350 univariate series in every configuration and 30 and 28 of the 180 multivariate series in the two TTM-r2.1 configurations, all of which use $L = 512$. Appendix C.3 reports the fallback rates and their effect on the reported gains.

Channel interaction. FDAS scores each channel separately, and taking the maximum across channels does not model their dependencies. An anomaly that changes the dependencies between channels while preserving each channel's temporal pattern may therefore go undetected when the backbone forecasts channels independently, as TTM-r2.1 does when used zero-shot.