WHEN EARLY IS NOT EARLY: THE PRECURSOR ILLUSION IN TIME-SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Early anomaly warning requires both detectable evidence before event onset and an alarm that exploits that evidence. Yet common evaluations define pre-onset targets by temporal proximity and apply point adjustment, allowing reactive detections to receive credit as apparent predictions. We term this failure mode the *precursor illusion*. We introduce a model-agnostic, placebo-calibrated framework that first tests whether pre-onset intervals are distinguishable from matched normal operation, then tests whether frozen causal detector scores exploit the certified evidence, and finally isolates evaluation effects while preserving scores and thresholds. The test detects evidence in 43 of 47 independently annotated transitions in the 3W dataset, yet across five benchmarks it identifies pre-onset deviations in only 70 of 500 evaluable events, revealing sparse and dataset-dependent evidence within the specified descriptor family. Mean detector scores achieve above-chance discrimination between precursor and matched normal windows in none of the 35 model–dataset combinations after multiple-testing correction, and only the near-onset score peaks of LSTM-VAE separate SMD precursor windows beyond a selection control, although all seven evaluated models distinguish the 3W transitions from normal operation. Point adjustment raises mean pre-onset Shifted-warning F1 from 0.036 to 0.177 without changing any alarm time, and input-blind alarms at matched rates obtain higher adjusted F1 than every detector. These results show that benchmark warnability must be established independently before early-warning models are ranked.

1 INTRODUCTION

Time-series anomaly detectors typically operate reactively and score deviations that are already observable at or after event onset (Xu et al., 2022; Tuli et al., 2022; Song et al., 2023; Wang et al., 2023; Feng et al., 2024; Wang et al., 2024; Wu et al., 2025; Zamanzadeh Darban et al., 2024). Early-warning approaches instead aim to raise an alarm before the onset time t_e, using only the observations available at the alarm time, so that operators have time to intervene (You et al., 2024; Hu et al., 2024; Park et al., 2025; Lee et al., 2023; Chen et al., 2025). Such an alarm is possible only if the pre-onset interval contains observable evidence that distinguishes an approaching event from normal operation. Following work on early-warning signals for critical transitions, we refer to this evidence as a precursor (Scheffer et al., 2009; Bury et al., 2021; 2023; Huang et al., 2024).

However, a common evaluation protocol labels the time steps just before each event onset as positive (Zhao et al., 2024; Kang et al., 2026). An alarm raised in this period then counts as an early warning, whether or not the data in the period differ from normal operation. The protocol thus assumes that every event is preceded by a precursor, without testing this assumption.

Another common evaluation practice further compounds this problem. Point adjustment counts an entire positive segment as detected if an alarm occurs anywhere within it (Su et al., 2019; Zhao et al., 2024; Lee et al., 2023; Kim et al., 2022; Wu & Keogh, 2023; Sarfraz et al., 2024; Wagner et al., 2023). When labels extend into the pre-onset period, a single post-onset hit can credit the entire extended segment. Figure 1 illustrates the resulting precursor illusion: a reactive detection can receive prediction credit without establishing both detectable precursor evidence and a pre-onset alarm.

These observations motivate three research questions:

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.

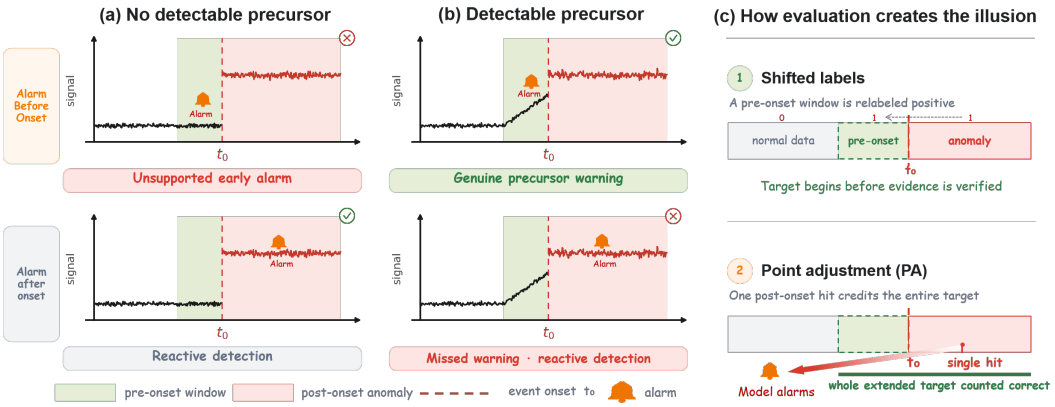

Figure 1: The precursor illusion. Genuine warning requires detectable evidence and a pre-onset alarm. A late alarm is reactive even when a precursor exists. Under an extended-label protocol, point adjustment (PA) can credit pre-onset points from a post-onset hit.

RQ1. Benchmark warnability. Do widely used benchmarks contain detectable precursor evidence, and how far before event onset can such evidence be detected?

RQ2. Model use of precursor evidence. For events identified as precursor-positive by the test, can frozen causal detector scores distinguish precursor windows from matched normal windows?

RQ3. Evaluation-protocol effects. With detector scores and thresholds held fixed, to what extent do shifting the evaluation target from the post-onset interval to a disjoint pre-onset interval and applying point adjustment alter reported F1?

The main contributions are summarized as follows:

- We formalize the distinction between precursor existence and alarm timing, and show why extending anomaly labels into the pre-onset period and applying point adjustment are insufficient to establish genuine early warning.
- We propose a model-agnostic, placebo-calibrated test for precursor detectability that uses a prespecified descriptor family to evaluate whether pre-onset intervals at different lead times contain evidence distinguishable from same-entity normal operation.
- We introduce matched-window precursor AUC (MPA) to assess whether frozen causal detector scores exploit certified precursor evidence. We further develop a trace- and threshold-preserving comparison that isolates the effects of evaluation targets and detection-credit rules.
- We conduct the first systematic characterization of the warnability limits of existing early-warning benchmarks, revealing sparse and dataset-dependent precursor evidence, score separation confined to near-onset peaks and a separate per-machine setting, and systematic F1 inflation under point adjustment.

2 Related Work

2.1 Reactive Time-Series Anomaly Detection

Time-series anomaly detection identifies observations or intervals that deviate from normal behavior. Established families include isolation-based methods (Liu et al., 2008), reconstruction models (Park et al., 2018; Audibert et al., 2020), density-estimation approaches (Zong et al., 2018; Su et al., 2019), and temporal-association models (Xu et al., 2022). Recent work further develops recurrent and attention-based memory, temporal decomposition, cross-channel dependency modeling, and frequency-domain representations (Tuli et al., 2022; Song et al., 2023; Wang et al., 2023; Feng et al., 2024; Wang et al., 2024; Wu et al., 2025; Zamanzadeh Darban et al., 2024).

Despite architectural differences, most detectors remain reactive under standard online evaluation: the score at time t reflects observations available only up to that time. They can establish detectability once anomalous manifestations emerge, but performance near or after onset does not show that a usable warning signal existed beforehand. Forecast-error detectors face the same limitation because prediction errors can be computed only after the target observation arrives. Conventional detection accuracy therefore does not establish pre-onset warnability.

2.2 Early Anomaly Warning and Precursors

Early anomaly-warning methods instead infer future abnormality from pre-onset observations. Existing approaches include precursor detection, modeling of warning delays and horizons, forecasting future observations or anomaly states, and learning pre-onset representations (You et al., 2024; Zhao et al., 2024; Hu et al., 2024; Kang et al., 2026; Park et al., 2025; Lee et al., 2023; Chen et al., 2025). Some methods derive warnings from forecasts, whereas others construct event-relative targets by extending anomaly labels before onset or labeling sampled pre-onset windows as positive.

Research on critical transitions shows that rising autocorrelation, increasing variance, and learned dynamical signatures may precede certain state changes (Scheffer et al., 2009; Bury et al., 2021; 2023; Huang et al., 2024). However, their presence, form, and detectable horizon are system-dependent. Early-warning targets are often defined by temporal position, while their distinguishability from normal operation is not independently verified. Performance on such targets therefore does not necessarily demonstrate detectable precursor evidence.

2.3 Evaluation Metrics and Benchmark Validity

Time-series anomaly-detection evaluation has received increasing scrutiny. Pointwise F1 poorly represents interval-valued anomalies, whereas point adjustment may substantially inflate performance by counting an entire segment as detected after a single hit (Kim et al., 2022; Sarfraz et al., 2024; Wagner et al., 2023). Range-, affiliation-, volume-, soft-matching-, and proximity-based metrics have been proposed to capture overlap, fragmentation, tolerance, and detection delay (Huet et al., 2022; Paparrizos et al., 2022a; Ghorbani et al., 2024; Salles et al., 2024; Sørbø & Ruocco, 2024). These metrics refine how predictions are counted once a target interval is accepted, but do not establish whether that interval contains detectable information.

Benchmark studies have also identified duplicated or trivial segments, contaminated training data, inconsistent annotations, and unstable model rankings (Wu & Keogh, 2023; Liu & Paparrizos, 2024; Paparrizos et al., 2022b; Wagner et al., 2023). Their focus, however, remains benchmark reliability after abnormal behavior becomes observable. Whether designated pre-onset intervals contain detectable evidence before the annotated event begins remains insufficiently examined, limiting the interpretation of both model performance and evaluation metrics in early anomaly warning.

3 Methodology

Our evaluation follows three stages aligned with RQ1–RQ3. Benchmark evidence is tested first, frozen scores are then evaluated on the resulting event strata, and evaluation targets and scoring rules are finally varied while scores, normalization, and thresholds remain fixed. This ordering keeps the evidence for benchmark warnability independent of model performance.

3.1 Problem Setting

For a multivariate series $x^{(i)}_{1:T_i} \in \mathbb{R}^{T_i \times C_i}$ from entity i, event e is a maximal labeled interval $A_e = [t_e, u_e]$ with onset t_e. Its width-W candidate precursor interval is $P_e(W) = [t_e - W, t_e)$. Unless stated otherwise, $P_e = P_e(100)$ for the five benchmarks, and 3W uses its annotated transition stage.

The normal pool $\mathcal{N}_i$ contains only labeled-normal points at least H_{guard} steps from every event and is partitioned as $\mathcal{N}_i = \mathcal{N}_i^{\text{fit}} \dot\cup \mathcal{N}_i^{\text{null}} \dot\cup \mathcal{N}_i^{\text{ver}}$. The three subsets estimate normal descriptor distributions, construct the placebo null, and verify its held-out rejection rate, respectively. Candidate intervals overlapping an annotated anomaly are excluded and reported as NA.

At scale $w \in \mathcal{S} = \{25, 50\}$ and lead $\ell \geq 0$, the precursor detectability test compares the candidate interval $Q_e(w, \ell) = [t_e - \ell - w, t_e - \ell)$ with its preceding reference $R_e(w, \ell) = [t_e - \ell - 2w, t_e -$

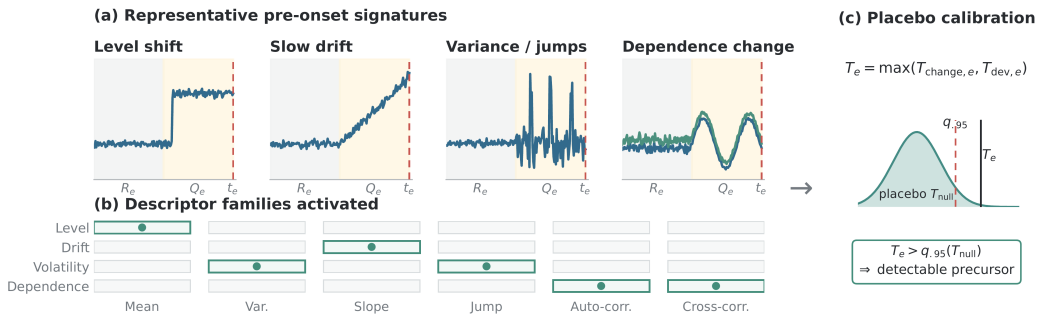

Figure 2: Precursor detectability test. Descriptor changes and deviations from the normal regime are compared with a same-entity placebo null. Candidate and reference windows use no post-onset samples; the evaluated detectors do not enter the test.

$\ell - w$). The primary event taxonomy is defined at $\ell = 0$. Since $w_{\max} = 50$, the largest pair spans the same 100-step interval used in matched-window model evaluation.

A detector produces the causal score $s_i(t) = f(x_{1:t}^{(i)})$ and alarm $a_i(t;\tau_i) = \mathbf{1}[s_i(t) > \tau_i]$. The threshold τ_i and all normalization statistics are fitted using normal-only data. Alarms in P_e are warnings, whereas alarms in A_e are post-onset detections.

3.2 RQ1: Precursor Detectability

An early-warning benchmark rewards alarms raised within its pre-onset intervals, so each such interval should contain evidence that separates it from normal operation. The precursor detectability test checks this requirement before any detector is evaluated, using only the benchmark data, the annotated onsets, and matched normal windows from the same entity. At each lead, the test compares the candidate with matched normal windows independently of all evaluated detectors. The descriptor vector $\boldsymbol{\phi}(I) \in \mathbb{R}^{12}$ stacks mean level, dispersion, slope, jump energy, autocorrelation, and cross-channel correlation at the two scales, each reduced to a root-mean-square over channels. Rising dispersion and lag-1 autocorrelation are the classical early-warning indicators of critical slowing down (Scheffer et al., 2009), and the remaining descriptors capture shifts in level, trend, abrupt jumps, and cross-channel coupling. The change $\Delta_j(Q_e, R_e)$ is the root-mean-square over channels of the change in descriptor j from R_e to Q_e. Normal-only fit data estimate its median m_j and spread $\widehat{\sigma}_j^{(0)}$ (95th percentile minus median), the mean $\boldsymbol{\mu}_i$ of $\boldsymbol{\phi}$, and a shrinkage-regularized covariance $\widetilde{\boldsymbol{\Sigma}}_i$. With $\varepsilon = 10^{-8}$, the composite statistic is

$$
\begin{aligned}
T_{\text{change},e} &= \max_j \frac{\Delta_j(Q_e, R_e) - m_j}{\max\left(\widehat{\sigma}_j^{(0)}, \varepsilon\right)}, \\
T_{\text{dev},e} &= \left[\left(\boldsymbol{\phi}(Q_e) - \boldsymbol{\mu}_i\right)^\top \widetilde{\boldsymbol{\Sigma}}_i^{-1} \left(\boldsymbol{\phi}(Q_e) - \boldsymbol{\mu}_i\right) \right]^{1/2}, \\
T_e &= \max\left(T_{\text{change},e}, T_{\text{dev},e}\right).
\end{aligned}
\tag{1}
$$

The two branches capture local changes and departures from the normal regime. An identical multi-scale calculation at M_0 pseudo-onsets in $\mathcal{N}_i^{\text{null}}$ yields placebo statistics $T_{i,m}^{(0)}$. Under event–placebo exchangeability, the empirical p-value and verdict are

$$
p_{e,\ell} = \frac{1 + \sum_{m=1}^{M_0} \mathbf{1}\left[T_{i,m}^{(0)} \geq T_e\right]}{M_0 + 1}, \qquad D_e(\ell;\alpha) = \mathbf{1}\left[p_{e,\ell} \leq \alpha\right]. \tag{2}
$$

Under event–placebo exchangeability, the null rejection probability is at most α; overlapping windows and operating-mode drift may violate this assumption. The held-out pool $\mathcal{N}_i^{\text{ver}}$ estimates the empirical placebo rejection rate $\widehat{r}_d^{(0)}$. For each dataset and lead, the test reports the event rejection rate $\widehat{r}_{d,\ell}$ and its excess $\widehat{\eta}_{d,\ell} = \widehat{r}_{d,\ell} - \widehat{r}_d^{(0)}$. Events of one entity share a calibration and neighbouring verification windows overlap, so the dataset-level test resamples entities for multi-entity datasets and events otherwise, propagates placebo-rate uncertainty through an effective pool size

approximated from the stride and window length, and computes one-sided bootstrap p-values with Benjamini–Hochberg correction over the dataset–lead grid. For PSM and SWaT, these p-values rely on an independent-event approximation; sensitivity to the effective pool size is reported in Appendix B.2.2.

At $\ell = 0$, events with detectable pre-onset evidence are assigned to P. Remaining events are assigned to I if equal-length post-onset support is detectable, or O otherwise. Invalid intervals are NA. RQ2 contrasts P with $\mathrm{NoP} = \mathrm{I} \cup \mathrm{O}$. Placebo-adjusted P counts subtract expected null discoveries from raw counts and estimate at the dataset level how many events carry detectable evidence.

3.3 RQ2: Matched-Window Precursor Discrimination

For each event e, frozen causal scores are averaged over the candidate P_e and M_c equal-length same-entity normal controls $C_{e,m}$, giving $\overline{s}_e^P$ and $\overline{s}_{e,m}^C$. MPA is the probability that a precursor window scores above a matched normal window, with ties counted as one half, which makes it the paired analogue of AUC:

$$\mathrm{MPA}(\mathcal{E}) = \frac{1}{|\mathcal{E}| M_c} \sum_{e \in \mathcal{E}} \sum_{m=1}^{M_c} \left(\mathbf{1}\left[\overline{s}_e^P > \overline{s}_{e,m}^C\right] + \frac{1}{2}\mathbf{1}\left[\overline{s}_e^P = \overline{s}_{e,m}^C\right] \right). \tag{3}$$

Paired-window exchangeability implies $\mathbb{E}[\mathrm{MPA}] = 0.5$, including for constant scores. For test-selected P events, MPA^P measures score separation from matched normal windows; $\mathrm{MPA}^{\mathrm{NoP}}$ and $\Delta_{\mathrm{MPA}} = \mathrm{MPA}^P - \mathrm{MPA}^{\mathrm{NoP}}$ provide comparisons with NoP events. Tests randomize at the event level, and confidence intervals bootstrap events and matched controls. Because the test can locate evidence within the final steps of P_e, MPA is also reported with window maxima over the final $L \in \{10, 25, 50\}$ steps of P_e and of each control. Because the test selects windows that deviate from normal operation, a detector that scores such deviations can separate the selected windows without using a precursor. Test-positive held-out normal windows therefore serve as pseudo-events, and their MPA is compared with that of P events to assess whether the observed separation can be explained by test selection alone. MPA measures discrimination, while RQ1 determines the evidence horizon.

3.4 RQ3: Trace-Preserving Protocol Comparison

In post hoc detection, point adjustment tolerates small misalignments between an annotated anomaly segment and the alarms raised inside it (Su et al., 2019; Kim et al., 2022). Within a post-onset segment, a single alarm shows that the detector responded to the anomaly, which makes this tolerance appropriate for detection. Our concern is the transfer of credit across the onset. A test-supported warning requires both $D_e = 1$ and an alarm in P_e; extending a target to $P_e \cup A_e$ alone establishes neither condition. Under PA on the merged target $P_e \cup A_e$, an alarm raised after the anomaly has begun marks the preceding interval as warned. RQ3 evaluates PA only as a credit rule for early warning.

To isolate these effects, RQ3 freezes scores, normalization, and normal-only thresholds. The Detection target is A_e, the Shifted-warning target is the disjoint P_e, and the Merged target is $P_e \cup A_e$. Each PA variant expands a hit only within its own target, and every protocol is compared with input-blind alarms drawn at each detector's own alarm rate. Detection and Shifted-warning PA measure within-segment inflation; Merged PA directly tests whether post-onset hits receive pre-onset credit. Thus target choice and credit assignment change, but detector outputs do not. Appendix A gives the full definitions and failure-mode analysis.

4 Experiments

Experiment overview Section 4.1 validates the framework on known ground truth, Sections 4.2–4.3 evaluate benchmark warnability and model use of test-identified evidence, and Section 4.4 isolates the effects of evaluation targets and point adjustment (PA).

Datasets We use SMD (Su et al., 2019), MSL and SMAP (Hundman et al., 2018), SWaT (Goh et al., 2016), and PSM (Abdulaal et al., 2021). These five benchmarks contain 537 annotated anomaly events. They define event onset and duration but not whether the preceding normal-labeled

interval contains a precursor, which is therefore evaluated using the test in Section 3.2. The 3W dataset (Vargas et al., 2019) is used only as an external positive control because it explicitly labels a transitional stage associated with fault development.

Baselines Seven representative methods span isolation, reconstruction, density estimation, association, and forecasting: iForest (Liu et al., 2008), LSTM-VAE (Park et al., 2018), DAGMM (Zong et al., 2018), USAD (Audibert et al., 2020), Anomaly Transformer (Xu et al., 2022), Future Context Modeling (FCM) (Zhao et al., 2024), and Anomaly to Prompt (A2P) (Park et al., 2025). The first five assign abnormality scores to causal input windows, whereas FCM and A2P forecast future abnormality. A constant scorer verifies the input-blind chance guarantee of MPA.

Implementation details Scores are causal and aligned to the end of each input window. Normalization and threshold selection use normal-only data. Table 1 summarizes the prespecified settings affecting the test verdict. Every protocol reuses the same saved score trace, and a cell is reported only when its trace, configuration, and dataset manifest are available.

Table 1: Prespecified precursor-test and evaluation configuration.

Component	Primary setting
Features	level, spread, slope, jumps, lag-1 autocorrelation, cross-channel correlation
Windows	$w = \{25, 50\}$; ten leads from 0 to 200
Controls	same entity; cap 800; stride 5; lookback 100; guard 50; minimum 120 windows
Calibration	disjoint fit/null/verification; $\alpha = 0.05$; entity bootstrap; BH over 50 lead cells
Covariance	shrinkage 0.10; ridge 10^{-6}
MPA/F1	20 controls; mean-score MPA; normal-score 95th percentile

The threshold is fixed at the 95th normal-score percentile, a nominal 5% false-alarm rate. One sample is one minute in SMD, MSL, SMAP, and PSM, but one second in SWaT (Su et al., 2019; Goh et al., 2016; Abdulaal et al., 2021). Results report leads in each dataset's time unit.

4.1 Protocol Validity

Setup The test uses 500 null trials and 500 trials for each of eight precursor morphologies at four strengths. MPA uses 500 simulated events per cell with 20 matched controls and compares constant, random, time-only, and post-onset-only scores with a precursor-aware score that receives increasing fractions of the injected signal. No learned model or benchmark label enters this validation.

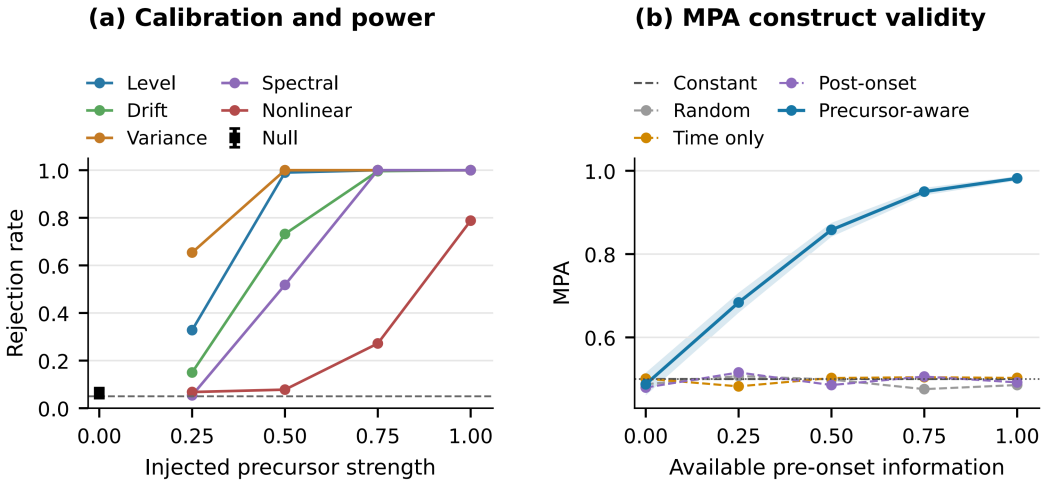

Figure 3: Controlled protocol validity. (a) Calibration and power of the precursor detectability test for five of eight tested morphologies. (b) Input-blind scores remain near MPA 0.5, whereas precursor-aware MPA increases with available evidence. Shading and bars denote event-level 95% intervals.

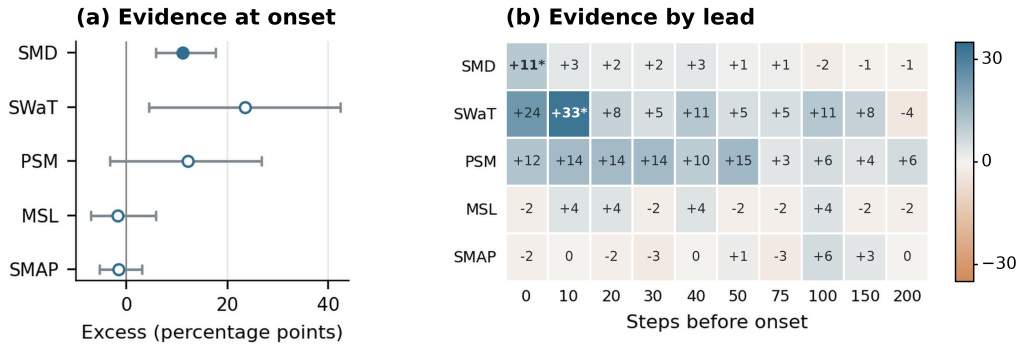

Figure 4: Benchmark warnability. (a) Excess over the held-out placebo rate at onset. (b) Excess at earlier leads. Horizontal bars are 95% bootstrap intervals, resampling entities for SMD, MSL, and SMAP and events for PSM and SWaT. Filled markers and asterisks denote BH $q < 0.05$. Heatmap entries are excess percentage points.

Results Figure 3 shows that the test is calibrated and that MPA orders scores from chance to full information. The null rejection rate is 0.060 $[0.042, 0.084]$, and at full strength power exceeds 0.95 for seven of the eight morphologies, with nonlinear dependence the hardest case at 0.79. Every input-blind score stays near 0.5, and precursor-aware MPA rises from 0.487 to 0.982. The test also recovers real precursors, detecting evidence in 43 of 47 annotated 3W transitions at a held-out placebo rate of 0.050. The verdict depends only on the observed history, since pairing identical pre-onset histories with a benign or a faulty future leaves every p-value unchanged. An onset continuity check identifies 94.5% of onsets annotated five steps late, the case that turns 93% of precursor-free synthetic events into apparent precursors, whereas an early annotation leaves the rate at the placebo level. Detailed results are provided in Appendix B.1.

4.2 Benchmark Warnability

Setup Results are reported at three levels, event verdicts, placebo-adjusted counts, and corrected dataset tests. Of the 537 annotated events, 500 are eligible and are assigned to P, I, or O, and the multiscale pair moves by $\ell \in \{0, 10, 20, 30, 40, 50, 75, 100, 150, 200\}$ steps before onset. Within each entity, disjoint normal windows fit the descriptors, build the split-conformal null, and verify the held-out false-positive rate. A bootstrap that resamples entities for SMD, MSL, and SMAP and events for PSM and SWaT, with Benjamini–Hochberg correction across 50 dataset–lead cells, tests the excess. Robustness runs add spectral entropy, lagged cross-correlation, and quadratic dependence, and vary scale, pool size, shrinkage, guard size, and channel selection one at a time.

Results Figure 4 and Table 2 show that the test detects pre-onset deviations in a small fraction of events. It identifies 70 P events among 500, about 49 above the placebo expectation; under the primary bootstrap procedure, only SMD at onset and SWaT at 10 seconds remain significant after correction. Of the 293 events with a test-detected change, 223 are positive only on the post-onset support.

Table 2: Event taxonomy at onset. Common P events have scores from all seven detectors.

Dataset	Eligible	P	I	Rejection at onset	Placebo rate	Common P
SMD	320	51	156	0.159	0.047	49
MSL	35	1	16	0.029	0.046	1
SMAP	67	2	26	0.030	0.045	2
PSM	45	7	20	0.156	0.034	6
SWaT	33	9	5	0.273	0.037	9
Total	500	70	223			67

The evidence is also short-lived. On SMD, the rejection rate halves within ten minutes of onset and approaches the placebo level by 50 minutes. On 3W, it stays above 0.91 at every lead up to 200 seconds, so the test recovers evidence far from onset when it exists. SWaT is the only benchmark

Table 3: Mean-score MPA (chance: 0.5) and F1 with fixed scores and thresholds. For dataset-wise MPA, bold marks BH-adjusted $q < 0.05$; MSL/SMAP are omitted ($n < 5$). Benchmark MPA^P weights common P events from all five benchmarks equally, excluding 3W.

	Mean-score MPA					F1 protocols			
Model	3W ($n = 48$)	SMD ($n = 49$)	PSM ($n = 6$)	SWaT ($n = 9$)	Benchmark MPA^P	Det.	Det.+PA	Shift.	Shift.+PA
iForest	**0.906**	0.553	0.675	0.467	0.551	0.283	0.676	0.047	0.193
DAGMM	**0.942**	0.544	0.525	0.467	0.547	0.241	0.652	0.040	0.235
USAD	**0.996**	0.471	0.417	0.378	0.459	0.114	0.440	0.038	0.203
LSTM-VAE	**0.991**	0.534	0.567	0.367	0.524	0.270	0.719	0.045	0.264
AnomTrans	**0.681**	0.500	0.500	0.389	0.475	0.010	0.452	0.007	0.207
FCM	**1.000**	0.431	0.550	0.328	0.447	0.213	0.503	0.037	0.075
A2P	**0.943**	0.573	0.175	0.350	0.507	0.123	0.415	0.040	0.060

with more P events than I events, but the significance of its 10-second result depends on the effective placebo-pool size (Appendix B.2.2).

These patterns hold under checks on the onset and on the test. SMD evidence remains significant at the unadjusted level after the final five minutes are removed, 74.5% of SMD P events show a distinct jump at the annotated onset, and an expanded descriptor set leaves the P events of four datasets unchanged. Detailed results are provided in Appendices B.1 and B.2.

4.3 Model Precursor Discrimination

Setup Each frozen score is evaluated on the P interval against equal-length normal windows from the same entity, and 3W uses its annotated transition. Of the 70 P events, 67 have scores from all seven detectors and form the common P set. MPA is computed separately for P and NoP events with event-level randomization and BH correction across all model–dataset cells. Cells with fewer than five P events are not interpreted. FCM and A2P are trained specifically for early warning.

Results Table 3 shows that none of the 35 benchmark mean-score comparisons is significant after correction, and on SMD the contrasts between P and NoP events also remain nonsignificant after correction. All seven detectors distinguish the annotated 3W transitions from normal controls, with MPA of at least 0.68. The minimum MPA detectable with 80% power on SMD is about 0.60 without correction and 0.66 with Bonferroni correction, and the upper 95% bound of mean-score MPA on SMD P events is at most 0.656 for every detector.

Additional score separation appears in two SMD settings. Normal windows flagged by the test already raise MPA, so separation specific to precursor windows must exceed this selection level. LSTM-VAE does so with near-onset score peaks, 0.743 against 0.630, consistent with evidence confined to the final minutes before onset. In a separate per-machine experiment with a train-only normal reference, iForest reaches 0.861 on its 92 P events against 0.655 on test-positive normal windows. This experiment also changes the normal reference and the selected events, so it does not isolate the effect of training. FCM and A2P, trained for early warning, reach 1.000 and 0.943 on 3W but stay near chance on benchmark P events. Detailed results are provided in Appendix B.3.

4.4 Protocol Comparison

Setup A threshold at the 95th or 99th percentile of normal scores is reused across targets, and a warning counts toward recall only when a new alarm episode starts before onset. For the seven pooled-trained detectors, F1 is pooled within each dataset and macro-averaged across datasets.

Results Figure 5 and Table 3 show that PA inflates warning F1 without producing earlier alarms. For per-machine iForest traces, merged F1 more than doubles under PA while recall of newly issued alarms stays unchanged, and about a quarter of the credited events contain no alarm before onset. Across the seven pooled-trained detectors, PA raises mean Shifted-warning F1 from 0.036 to 0.177, and 20% to 46% of merged-target hits receive warning credit solely from post-onset alarms. Early recall shows the same effect, reaching 98.8% for a random score against 62.8% for iForest at a similar alarm-episode rate, so early recall alone also rewards frequent alarms.

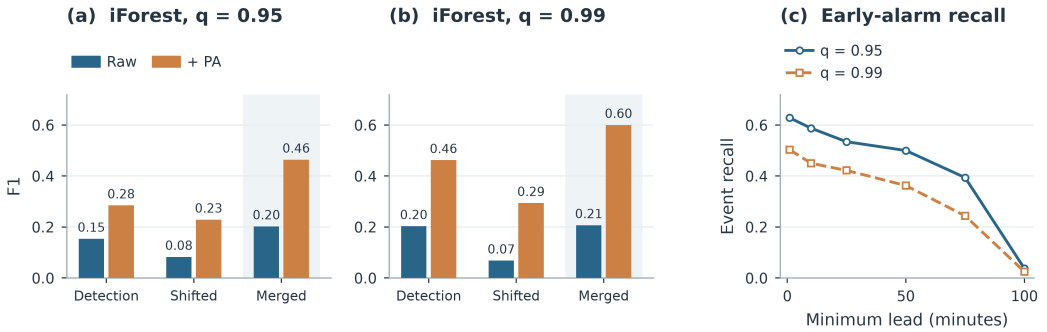

Figure 5: Credit transfer across onset on SMD with per-machine iForest traces. (a, b) F1 under the Detection, Shifted, and Merged targets with and without PA at two normal-score thresholds. (c) Recall of newly issued alarm episodes that start at least the given lead before onset.

Table 4: Mean F1 under PA for detectors and input-blind alarms at matched rates.

Target	Detectors	Random alarms	Circularly shifted alarms
Shifted warning	0.177	0.431	0.167
Merged	0.570	0.788	0.497

The inflation depends on the alarm rate. At a 5% pointwise rate, independent alarms hit a 100-step interval with probability about 0.99, and Table 4 shows that such input-blind alarms reach an adjusted Shifted-warning F1 of 0.431, above every detector. Circularly shifted alarms, which keep each detector's alarm runs, obtain adjusted F1 close to the detectors on the Shifted target and lower on the Merged target, and without PA detectors match random alarms at the same rate to within 0.002 across the tested settings. On 3W, where precursors are present, detectors already alarm before onset, and PA changes Shifted-warning F1 by less than 0.03 for six of the seven detectors, whereas on the benchmarks it raises this F1 about fivefold. Detailed results are provided in Appendix B.4.

5 Limitations and Future Work

A negative result does not rule out predictive information outside the specified descriptor family; a positive result indicates a deviation from the normal reference, not necessarily a fault-specific precursor. The test applies to entities whose normal operation is stationary. Entities that switch between operating modes require placebo windows matched to the current mode, as the drift of normal operation before the annotated 3W transitions in Appendix B.1 illustrates. The seven detectors were trained on pooled entities, and the per-machine experiment for iForest on SMD changes training, the normal reference, and the selected events together. Future work will separate these three factors, extend the descriptor set, develop mode-aware placebo matching, evaluate additional benchmarks with explicit transition annotations, combine MPA with event-level timeliness measures, and extend the test from onset-aligned windows to continuous scanning of all endpoints.

6 Conclusion

Early warning requires detectable pre-onset evidence and a pre-onset alarm. The placebo-calibrated test identifies deviations in 70 of 500 eligible events across five benchmarks and 43 of 47 independently annotated 3W transitions; 223 of 293 benchmark events with detected changes are positive only on post-onset support. Near-onset peaks and a separate per-machine SMD setting reveal separation missed by the primary mean-score comparison. With scores and thresholds fixed, PA raises warning F1 without improving alarm timing or precursor discrimination; it adds little on 3W, where detectors already alarm before onset. Evaluation should characterize benchmark warnability under an explicit evidence definition before model comparison and jointly assess precursor discrimination and alarm timeliness. The framework separates precursor existence, detector response, and credit for pre-onset alarms.

Reproducibility Statement

The paper specifies the event construction, placebo calibration, statistical tests, correction families, score alignment, thresholds, and all primary evaluation settings. Appendix A provides the complete estimands and derivations. The supplementary material contains the code used for the reported experiments and diagnostic analyses.

AI Use Statement

Generative AI was used for language editing and presentation assistance. All scientific content, analyses, results, and conclusions were produced and verified by the authors.

References

Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *KDD*, 2021.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In *KDD*, 2020.

Thomas M. Bury, R. I. Sujith, Induja Pavithran, Marten Scheffer, Timothy M. Lenton, Madhur Anand, and Chris T. Bauch. Deep learning for early warning signals of tipping points. *Proceedings of the National Academy of Sciences*, 118(39), 2021.

Thomas M. Bury, Daniel Dylewsky, Chris T. Bauch, Madhur Anand, Leon Glass, Alvin Shrier, and Gil Bub. Predicting discrete-time bifurcations with deep learning. *Nature Communications*, 14: 6331, 2023.

PengYu Chen, Xiaohou Shi, Yuan Chang, Yan Sun, and Sajal K. Das. RED-F: Reconstruction-elimination based dual-stream contrastive forecasting for multivariate time series anomaly prediction. *arXiv preprint arXiv:2511.20044*, 2025.

Yuye Feng, Wei Zhang, Yao Fu, Weihao Jiang, Jiang Zhu, and Wenqi Ren. Sensitivehue: Multivariate time series anomaly detection by enhancing the sensitivity to normal patterns. In *KDD*, pp. 782–793, 2024.

Ramin Ghorbani, Marcel J. T. Reinders, and David M. J. Tax. Pate: Proximity-aware time series anomaly evaluation. In *KDD*, pp. 872–883, 2024.

Jonathan Goh, Sridhar Adepu, Khurum Nazir Junejo, and Aditya Mathur. A dataset to support research in the design of secure water treatment systems. In *International Conference on Critical Information Infrastructures Security*, 2016.

Shiyan Hu, Kai Zhao, Xiangfei Qiu, Yang Shu, Jilin Hu, Bin Yang, and Chenjuan Guo. Multirc: Joint learning for time series anomaly prediction and detection with multi-scale reconstructive contrast. *arXiv preprint arXiv:2410.15997*, 2024.

Yu Huang, Sebastian Bathiany, Peter Ashwin, and Niklas Boers. Deep learning for predicting rate-induced tipping. *Nature Machine Intelligence*, 6(12):1556–1565, 2024.

Alexis Huet, José Manuel Navarro, and Dario Rossi. Local evaluation of time series anomaly detection algorithms. In *KDD*, pp. 635–645, 2022.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *KDD*, 2018.

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

Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous evaluation of time-series anomaly detection. In *AAAI*, 2022.

Cheryl Lee, Tianyi Yang, Zhuangbin Chen, Yuxin Su, and Michael R. Lyu. Maat: Performance metric anomaly anticipation for cloud services with conditional diffusion. In *ASE*, 2023.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *ICDM*, 2008.

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

John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S. Tsay, Aaron J. 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, 2022a.

John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S. Tsay, Themis Palpanas, and Michael J. Franklin. Tsb-uad: An end-to-end benchmark suite for univariate time-series anomaly detection. *Proceedings of the VLDB Endowment*, 15(8):1697–1711, 2022b.

Daehyung Park, Yuuna Hoshi, and Charles C. Kemp. A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. *IEEE Robotics and Automation Letters*, 3(3):1544–1551, 2018.

Min-Yeong Park, Won-Jeong Lee, Seong Tae Kim, and Gyeong-Moon Park. When will it fail?: Anomaly to prompt for forecasting future anomalies in time series. In *ICML*, 2025.

Rebecca Salles, Janio Lima, Michel Reis, Rafaelli Coutinho, Esther Pacitti, Florent Masseglia, Reza Akbarinia, Chao Chen, Jonathan Garibaldi, Fabio Porto, and Eduardo Ogasawara. SoftED: Metrics for soft evaluation of time series event detection. *Computers & Industrial Engineering*, 198: 110728, 2024. doi: 10.1016/j.cie.2024.110728.

M. Saquib Sarfraz, Mei-Yen Chen, Lukas Layer, Kunyu Peng, and Marios Koulakis. Position: Quo vadis, unsupervised time series anomaly detection? In *ICML*, pp. 43461–43476, 2024.

Marten Scheffer, Jordi Bascompte, William A. Brock, Victor Brovkin, Stephen R. Carpenter, Vasilis Dakos, Hermann Held, Egbert H. van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461:53–59, 2009.

Junho Song, Keonwoo Kim, Jeonglyul Oh, and Sungzoon Cho. Memto: Memory-guided transformer for multivariate time series anomaly detection. In *NeurIPS*, 2023.

Sondre Sørbø and Massimiliano Ruocco. Navigating the metric maze: A taxonomy of evaluation metrics for anomaly detection in time series. *Data Mining and Knowledge Discovery*, 38(3): 1027–1068, 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 *KDD*, 2019.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 15(6): 1201–1214, 2022.

Ricardo Emanuel Vaz Vargas, Celso José Munaro, Patrick Marques Ciarelli, André Gonçalves Medeiros, Bruno Guberfain do Amaral, Daniel Centurion Barrionuevo, Jean Carlos Dias de Araújo, Jorge Lins Ribeiro, and Lucas Pierezan Magalhães. A realistic and public dataset with rare undesirable real events in oil wells. *Journal of Petroleum Science and Engineering*, 181, 2019.

Dennis Wagner, Tobias Michels, Florian C. F. Schulz, Arjun Nair, Maja Rudolph, and Marius Kloft. Timesead: Benchmarking deep multivariate time-series anomaly detection. *Transactions on Machine Learning Research*, 2023.

Chengsen Wang, Zirui Zhuang, Qi Qi, Jingyu Wang, Xingyu Wang, Haifeng Sun, and Jianxin Liao. Drift doesn't matter: Dynamic decomposition with diffusion reconstruction for unstable multivariate time series anomaly detection. In *NeurIPS*, 2023.

Rui Wang, Xudong Mou, Renyu Yang, Kai Gao, Pin Liu, Chongwei Liu, Tianyu Wo, and Xudong Liu. Cutaddpaste: Time series anomaly detection by exploiting abnormal knowledge. In *KDD*, pp. 3176–3187, 2024.

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

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. Catch: Channel-aware multivariate time series anomaly detection via frequency patching. In *ICLR*, 2025.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *ICLR*, 2022.

Jiang You, Arben Çela, René Natowicz, Jacob Ouanounou, and Patrick Siarry. Anomaly prediction: A novel approach with explicit delay and horizon. In *2024 IEEE 20th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 1–8, 2024. doi: 10.1109/ICCP63557.2024.10793012.

Zahra Zamanzadeh Darban, Geoffrey I. Webb, Shirui Pan, Charu C. Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1): 15:1–15:42, 2024. doi: 10.1145/3691338.

Sinong Zhao, Wenrui Wang, Hongzuo Xu, Zhaoyang Yu, Qingsong Wen, Gang Wang, Xiaoguang Liu, and Guansong Pang. Abnormality forecasting: Time series anomaly prediction via future context modeling. *arXiv preprint arXiv:2410.12206*, 2024.

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR*, 2018.

APPENDIX

A DETAILED METHODOLOGY

A.1 FAILURE MODES AND DESIGN REQUIREMENTS

A common shifted-label protocol extends each anomaly interval into the preceding W steps:

$$\widetilde{y}_W^{\text{ext}}(t) = \mathbf{1}\left[t \in \bigcup_e (P_e(W) \cup A_e)\right]. \tag{4}$$

This target depends on temporal position relative to t_e, not on whether P_e contains detectable evidence. For any positive segment B_e, point adjustment (PA) replaces every prediction within that segment by the same segment-level hit:

$$a_i^{\text{PA},B}(t) = \begin{cases} \max_{v \in B_e} a_i(v;\tau_i), & t \in B_e, \\ a_i(t;\tau_i), & t \notin B_e. \end{cases} \tag{5}$$

Let $D_e \in \{0,1\}$ denote the verdict of the precursor detectability test for event e. A test-supported warning requires a positive verdict and an alarm within the designated pre-onset interval:

$$G_e = D_e\, \mathbf{1}\left[\exists\, t \in P_e(W) : a_i(t;\tau_i) = 1\right]. \tag{6}$$

An early alarm in a test-negative interval is unsupported by this test, whereas a test-positive interval followed only by a post-onset alarm represents a missed warning followed by reactive detection.

Shifted labels. Equation 4 makes the target recoverable from event-relative position. A diagnostic oracle can set $\widehat{y}_{\text{pos}}(t) = \widetilde{y}_W^{\text{ext}}(t)$ and classify perfectly without reading x_t. Although not deployable, this oracle shows that high shifted-label performance need not imply input-dependent precursor evidence.

Point adjustment. For an input-blind alarm firing independently with probability p per point, the probability of a PA hit on B_e is $1-(1-p)^{|B_e|}$, which increases with segment length. If PA is applied to the extended segment $P_e \cup A_e$, a post-onset hit can propagate to its pre-onset portion.

A.2 RQ1: PLACEBO-CALIBRATED PRECURSOR DETECTABILITY TEST

At each lead, RQ1 compares the event pair with an identically constructed multiscale pair $\mathcal{C}_{i,m}(\ell)$ at pseudo-onset m in $\mathcal{N}_i^{\text{null}}$. Under the null, the event and placebo pairs follow the same distribution.

The test uses six descriptors of a window, namely mean level, log dispersion, slope between the first and last quarter, log jump energy, lag-1 autocorrelation, and cross-channel correlation. The first five are computed per channel on the standardized series and reduced to a root-mean-square over channels. Cross-channel correlation is the root-mean-square of the off-diagonal correlations among the 16 channels with the largest first-difference variability. Stacking the six descriptors at the two scales gives $\boldsymbol{\phi}(I) \in \mathbb{R}^d$ with $d = 12$ for every dataset. The change $\Delta_j(Q_e, R_e)$ is the root-mean-square of the change in descriptor j from R_e to Q_e, computed per channel or, for correlation, per channel pair, and m_j and $\widehat{\sigma}_j^{(0)}$ are the median of Δ_j and its 95th percentile minus the median over reference–candidate pairs in $\mathcal{N}_i^{\text{fit}}$. The entity-specific normal mean $\boldsymbol{\mu}_i$ and covariance $\widehat{\boldsymbol{\Sigma}}_i$ are estimated from $\mathcal{N}_i^{\text{fit}}$. For numerical stability, the covariance is regularized as $\widetilde{\boldsymbol{\Sigma}}_i = (1-\lambda)\widehat{\boldsymbol{\Sigma}}_i + \lambda\,\text{diag}(\widehat{\boldsymbol{\Sigma}}_i) + \rho\mathbf{I}$, where λ and ρ are fixed before benchmark evaluation. Suppressing lead and scale indices, the test statistics are

$$\begin{aligned} T_{\text{change},e} &= \max_{j=1,\ldots,d} \frac{\Delta_j(Q_e,R_e) - m_j}{\max\left(\widehat{\sigma}_j^{(0)}, \varepsilon\right)}, \qquad \varepsilon = 10^{-8}, \\ T_{\text{dev},e} &= \left[\left(\boldsymbol{\phi}(Q_e) - \boldsymbol{\mu}_i\right)^\top \widetilde{\boldsymbol{\Sigma}}_i^{-1} \left(\boldsymbol{\phi}(Q_e) - \boldsymbol{\mu}_i\right)\right]^{1/2}, \\ T_e &= \max\left(T_{\text{change},e}, T_{\text{dev},e}\right). \end{aligned} \tag{7}$$

The change branch captures local changes, whereas the deviation branch captures slower departures from the normal regime. Applying the same composite statistic to event and placebo pairs preserves branch scales and dependence in the empirical null.

Placebo calibration. The same statistic, evaluated at M_0 pseudo-onsets in $\mathcal{N}_i^{\text{null}}$, gives placebo values $T_{i,1}^{(0)}, \ldots, T_{i,M_0}^{(0)}$. The event p-value and verdict are

$$p_{e,\ell} = \frac{1 + \sum_{m=1}^{M_0} \mathbf{1}\left[T_{i,m}^{(0)} \geq T_e\right]}{M_0 + 1}, \qquad D_e(\ell;\alpha) = \mathbf{1}\left[p_{e,\ell} \leq \alpha\right]. \tag{8}$$

A precursor is declared detectable only when its statistic exceeds the variation produced by the same procedure on matched normal data. Evaluated detectors do not enter this verdict.

For dataset d, RQ1 reports the rejection rate and its excess over the held-out placebo rate estimated from $\mathcal{N}_i^{\text{ver}}$:

$$\widehat{r}_{d,\ell} = \frac{1}{|\mathcal{E}_{d,\ell}|} \sum_{e \in \mathcal{E}_{d,\ell}} D_e(\ell;\alpha), \qquad \widehat{\eta}_{d,\ell} = \widehat{r}_{d,\ell} - \widehat{r}_d^{(0)}. \tag{9}$$

An entity-clustered bootstrap compares $\widehat{r}_{d,\ell}$ with $\widehat{r}_d^{(0)}$ (Appendix B.2.2), with multiplicity controlled over the prespecified dataset–lead grid. A significant $\widehat{\eta}_{d,\ell} > 0$ indicates excess pre-onset rejection at lead ℓ under the stated bootstrap assumptions; the resulting profile describes how detectability changes with lead.

A.3 Evidence-Based Event Stratification

Let $D_e^{\text{pre}} = D_e(0;\alpha)$ be the verdict for support ending at onset, and let D_e^{on} be the verdict for equal-length support beginning at t_e. Each evaluable event is assigned to

$$\text{group}(e) = \begin{cases} \text{P}, & D_e^{\text{pre}} = 1, \\ \text{I}, & D_e^{\text{pre}} = 0 \text{ and } D_e^{\text{on}} = 1, \\ \text{O}, & D_e^{\text{pre}} = 0 \text{ and } D_e^{\text{on}} = 0. \end{cases} \tag{10}$$

P denotes leading evidence, I an onset-only change, and O neither. RQ2 contrasts P with $\text{NoP} = \text{I} \cup \text{O}$. In addition to raw P counts, placebo-adjusted counts are reported as $\max(0, |\text{P}_d| - \widehat{r}_d^{(0)} N_d)$, where N_d is the number of evaluable events in dataset d.

A.4 RQ2: Evidence-Conditioned Model Evaluation

The precursor detectability test asks whether evidence exists; matched-window precursor discrimination (MPA) asks whether frozen detector scores discriminate it. Each event is paired with M_c equal-length, same-entity controls $C_{e,m} \subset \mathcal{N}_i$. Their mean scores are

$$\overline{s}_e^P = \frac{1}{|P_e|} \sum_{t \in P_e} s_i(t), \qquad \overline{s}_{e,m}^C = \frac{1}{|C_{e,m}|} \sum_{t \in C_{e,m}} s_i(t). \tag{11}$$

A precursor win receives one and a tie one half. For event set $\mathcal{E}$,

$$w_{e,m} = \mathbf{1}\left[\overline{s}_e^P > \overline{s}_{e,m}^C\right] + \frac{1}{2}\mathbf{1}\left[\overline{s}_e^P = \overline{s}_{e,m}^C\right], \qquad \text{MPA} = \frac{1}{|\mathcal{E}| M_c} \sum_{e \in \mathcal{E}} \sum_{m=1}^{M_c} w_{e,m}. \tag{12}$$

The primary RQ2 estimand is threshold-free, although a thresholded variant may replace mean scores with alarm fractions.

If paired interval scores are exchangeable under the null, precursor and control wins are equiprobable and half credit splits ties. Hence $\mathbb{E}[\text{MPA}] = 0.5$, including for a constant scorer. Statistical tests use events as the randomization unit, and confidence intervals use a two-stage bootstrap over events and matched controls.

For each model m and dataset d, P and $\mathrm{NoP} = \mathrm{I} \cup \mathrm{O}$ are evaluated separately. Their evidence-conditioned contrast is

$$\Delta_{\mathrm{MPA},m,d} = \mathrm{MPA}^{P}_{m,d} - \mathrm{MPA}^{\mathrm{NoP}}_{m,d} . \tag{13}$$

$\mathrm{MPA}^{P} > 0.5$ indicates that P windows tend to score above matched normal windows, but does not by itself establish precursor-specific discrimination. $\mathrm{MPA}^{\mathrm{NoP}}$ and Δ_{MPA} compare discrimination between P and NoP events, while test-positive normal pseudo-events help assess the effect of test selection. MPA measures discrimination rather than earliness: RQ1 estimates the evidence horizon, whereas RQ2 evaluates detector scores on the resulting strata.

B Diagnostic Analyses

This appendix reports analyses that test whether the conclusions depend on specific design choices of the measurement framework. Each subsection states the purpose, the analysis, and the result. All analyses reuse the frozen event taxonomy and the frozen causal score traces, and no detector is retrained. Decision criteria were fixed before the corresponding results were computed. Analyses labelled post hoc were added after inspecting results and serve as sensitivity checks.

B.1 Protocol Validity

B.1.1 Power by Precursor Morphology

Table 5 reports the rejection rate of the precursor detectability test for all eight morphologies injected in Section 4.1. Each cell is estimated from 500 trials, and the rejection rate under the null is 0.060 $[0.042, 0.084]$.

Table 5: Rejection rate of the precursor detectability test by injected morphology and strength.

Morphology	0.25	0.50	0.75	1.00
Level	0.328	0.990	1.000	1.000
Slow drift	0.150	0.732	0.996	1.000
Variance	0.654	1.000	1.000	1.000
Spectral	0.054	0.518	1.000	1.000
Nonlinear dependence	0.068	0.078	0.272	0.788
Jump	0.092	0.974	1.000	1.000
Autocorrelation	0.144	0.878	1.000	1.000
Linear dependence	0.144	0.290	0.620	0.954

B.1.2 Recovery of Annotated 3W Transitions

The 3W positive control in Section 4.3 evaluates detector scores. This analysis applies the precursor detectability test itself to the 3W transitions with the benchmark settings unchanged. For each of the 48 instances, the normal pool is the operation before the annotated transition, the event point is the start of the fault steady state, and candidate and reference windows at every lead lie inside the annotated transition. Channels with more than 1% non-finite values or zero variance during normal operation are removed, and one instance with fewer than two remaining channels is excluded. Well-clustered bootstrap inference follows Section B.2.2.

The held-out placebo rate is 0.050 with a Wilson 95% interval of $[0.047, 0.054]$ over 12,538 windows. The test detects evidence in 43 of 47 transitions at $\ell = 0$ ($p < 10^{-4}$), and the detection rate stays between 0.915 and 0.936 for every lead from 0 to 200 seconds. All detections come from the deviation branch. For comparison, the rates at $\ell = 0$ are 0.159 for SMD, 0.156 for PSM, 0.273 for SWaT, 0.029 for MSL, and 0.030 for SMAP. As a negative control, windows ending at the annotated start of the transition were also tested. Of these, 11 of 47 are rejected ($p = 0.002$ against the placebo rate).

A second design splits the leading normal operation of each record chronologically into fit, null, and verification thirds, so that the verification windows lie closest to the transition. On a deterministic subset of 16 records from nine wells, this design rejects 84.8% of transition windows (well-bootstrap interval $[59.4, 98.4]\%$), 29.2% of verification windows ($[16.2, 41.7]\%$), and 37.5% of windows ending at the transition start. Both designs therefore find elevated rejection in the normal operation that

immediately precedes an annotated 3W transition, and the random split used for the benchmarks keeps the held-out rate at its nominal level.

B.1.3 Sensitivity to the Annotated Onset

The test takes the benchmark onset t_e as given. If an annotated onset is later than the true onset, the candidate window contains part of the anomaly. We examine this in three steps.

Synthetic onset shift Using the generator of Section 4.1, each synthetic series contains an anomaly, a level shift of $+3$ on four of eight channels starting at a true onset T, either alone or preceded by one of the eight injected precursor morphologies at strength 0.5 or 1.0. The test is applied at an annotated onset $T + \delta$ with 500 trials per cell. Table 6 reports the rate at which the test declares a precursor.

Table 6: Precursor verdict rate under annotated-onset shifts δ (positive means annotated later than the true onset).

Condition	-20	-10	-5	0	$+5$	$+10$	$+20$
Anomaly only	0.060	0.072	0.056	0.058	0.930	1.000	1.000
Precursor and anomaly	0.662	0.756	0.798	0.836	0.991	1.000	1.000

An onset annotated five steps late turns 93% of precursor-free events into apparent precursors, whereas an early annotation leaves the rate at the placebo level.

Known transition boundaries A second simulation uses six-channel AR(1) series with a true onset at step 300, an optional 80-step precursor (level, ramp, or variance) before it, and a common post-onset fault, with 300 trials per condition. Figure 6(a) shows that the precursor-free rejection rate is 15 of 300 at the true boundary and 300 of 300 once the annotation is 25 steps late. Figure 6(b) shows that all three precursors are detected at lead zero, while the ramp is detected in 68.3% of trials at lead 50 and 3.7% at lead 75, so the detectable horizon of a given statistic can be shorter than the true transition. Table 7 follows the same trials through the full framework at the true onset. A causal 25-step energy score separates every precursor condition from its controls and issues a new alarm before onset in every precursor trial, whereas precursor-free trials stay at chance.

Table 7: End-to-end outcome at the true onset in the known-boundary simulation, 300 trials per condition. Early recall requires a newly issued alarm episode before onset, and lead is the median first-alarm lead in steps.

Condition	Test positive	MPA	Early recall	Lead
No precursor	15	0.479	0.103	57
Level	300	1.000	1.000	75
Ramp	300	1.000	1.000	51
Variance	300	1.000	1.000	76

Continuity at the onset A late annotation places t_e inside an ongoing change, so the signal shows no additional jump at t_e and the largest change point lies earlier. For every event we computed the five-versus-five-step jump $J(s) = \|\bar{z}_{[s,s+5)} - \bar{z}_{[s-5,s)}\|_2$ on the standardized channels and compared $J(t_e)$ with the 95th percentile of J over fully normal neighbourhoods. An event is classified as onset jump if $J(t_e)$ exceeds this percentile, as boundary shift if it does not while the largest jump in $[t_e - 45, t_e + 5]$ lies at least three steps before t_e, exceeds the percentile, and points in the same direction as the post-onset deviation, and as undetermined otherwise. On the synthetic series, the check classifies 94.5% of the apparent precursors created by late annotations as boundary shift and 84.5% of correctly annotated precursors as onset jump, and it assigns 9.8% of correctly annotated precursors to boundary shift. Table 8 applies the check to SMD.

Table 8: Continuity at the annotated onset for SMD events.

Group	Onset jump	Boundary shift	Undetermined	Events
P	38 (74.5%)	6 (11.8%)	7 (13.7%)	51
I	109 (69.9%)	10 (6.4%)	37 (23.7%)	156

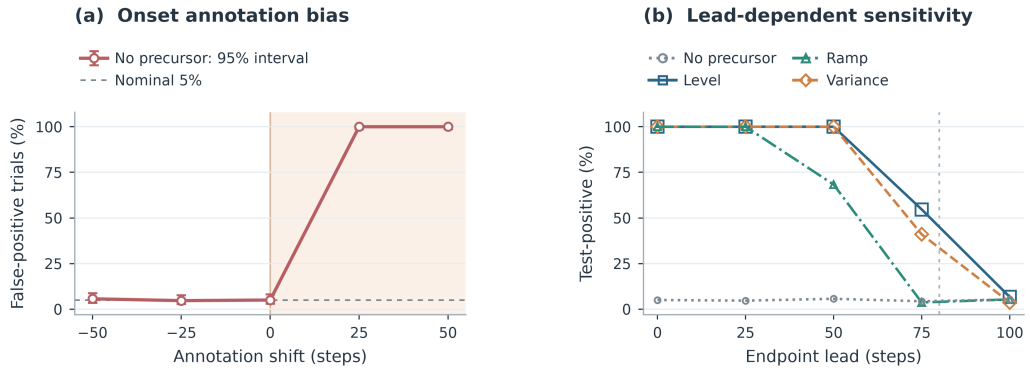

Figure 6: Known-boundary simulation, 300 trials per condition. (a) Rejection rate of precursor-free trials as the annotated onset moves relative to the true onset, with 95% Wilson intervals; shading marks late annotations that admit post-onset observations. (b) Rejection rate at the true onset across endpoint leads; the vertical line marks the 80-step precursor boundary.

The share of SMD P events classified as boundary shift, 11.8%, is close to the 9.8% that the check assigns to correctly annotated synthetic precursors.

Near-onset leads The candidate window at lead ℓ equals the lead-0 window under an onset moved ℓ steps earlier. Adding $\ell = 5$ to the prespecified grid reproduces every verdict of the original leads exactly. Table 9 reports the uncorrected primary bootstrap p-values.

Table 9: Event rejection rates near the annotated onset, with uncorrected bootstrap p-values.

Dataset	Placebo rate	$\ell = 0$	$\ell = 5$	$\ell = 10$
SMD	0.047	51/320 (0.001)	30/319 (0.015)	24/318 (0.072)
PSM	0.034	7/45 (0.051)	6/45 (0.089)	8/45 (0.031)
SWaT	0.037	9/33 (0.005)	11/33 (0.001)	12/33 ($<$ 0.001)

B.1.4 Identical Histories with Different Futures

Each synthetic history ending at the candidate endpoint is paired with a benign and a faulty continuation, 1,000 pairs per history type. The test reads only observations before the endpoint, so all 4,000 paired p-values are identical. The rejection rate is 0.055 for precursor-free histories, 0.982 for slow drift, and 1.000 for level and variance precursors, whichever future follows. In these paired trials, identical histories receive identical verdicts whichever future follows.

B.2 Benchmark Warnability

B.2.1 Corrected Entity Boundaries

SMD, MSL, and SMAP are distributed as concatenated arrays without entity boundaries. A verification against the official per-entity label files, matching both segment length and anomaly count, showed that the boundary table used in an earlier version assigned some SMD and SMAP segments to the wrong entity, while MSL was unaffected. All three research questions were rerun with the corrected table and every other setting unchanged, and all reported results use the corrected table. Table 10 lists every quantity that changed.

The one-event difference in the annotated total arises because a boundary in the original table split one SMD event into two adjacent pieces. Under the corrected table the SMD label array contains 327 contiguous events whether it is counted as a whole or entity by entity.

B.2.2 Dependence-Aware Dataset-Level Inference

Events of the same entity share one fitted calibration, control windows at stride 5 overlap, and the placebo rate is estimated from a finite verification pool. A binomial test that treats events as independent and the placebo rate as known therefore overstates significance. The primary test resamples

Table 10: Quantities that change under the corrected entity boundaries.

Quantity	Original table	Corrected table
Annotated events (five benchmarks)	538	537
Evaluable events	496	500
P events (all benchmarks)	69	70
SMD P events / evaluable	49 / 316	51 / 320
SMAP P events	3	2
SMD common P events in RQ2	45	49
Mean Detection F1, without / with PA	0.186 / 0.549	0.179 / 0.551
Mean Shifted-warning F1, without / with PA	0.037 / 0.173	0.036 / 0.177
Significant benchmark MPA cells after BH	0 of 35	0 of 35
3W MPA cells with $q < 0.05$	7 of 7	7 of 7

whole entities with replacement for SMD, MSL, and SMAP and resamples events for the single-entity PSM and SWaT, draws the placebo rate from a binomial distribution over the effective pool size $n_{\text{ver}} \times 5/100$, an effective-sample-size approximation based on stride 5 and the 100-step support, and computes a centred bootstrap p-value from 20,000 draws. Table 11 compares this test with alternatives. BH correction over the 50 dataset–lead cells is applied within every row.

Table 11: Dataset–lead cells with BH $q < 0.05$ under alternative dataset-level tests. MSL and SMAP have no significant cell under any test.

Test	SMD	PSM	SWaT
Binomial, placebo rate known	$\ell = 0$	$\ell = 0$–50	$\ell = 0, 10, 40, 100$
Fisher exact, nominal pool (post hoc)	$\ell = 0$	$\ell = 0, 10, 20, 30, 50$	$\ell = 0, 10$
Bootstrap, nominal pool	$\ell = 0$	none	$\ell = 0, 10$
Bootstrap, effective pool (primary)	$\ell = 0$	none	$\ell = 10$
Fisher exact, effective pool (post hoc)	$\ell = 0$	none	none

SMD at $\ell = 0$ is significant under every test, and no SMD lead beyond onset is significant under any test. The PSM and SWaT profiles depend on the effective size of the placebo pool, so no lead beyond onset is interpreted as an established warning horizon.

B.2.3 Placebo Rejection Rate by Distance to the Next Event

Candidate windows end at an annotated onset, whereas placebo windows lie at least 50 steps before the next event. Table 12 stratifies the held-out placebo windows by their distance to the next annotated onset.

Table 12: Held-out placebo rejection rate by distance to the next annotated onset in samples, with Wilson 95% intervals. PSM and SWaT have fewer than 25 windows in each band below 200 and are omitted.

Dataset	50–100	100–200	200–500	$\geq$500
SMD	0.053 [0.029, 0.095]	0.042 [0.026, 0.067]	0.068 [0.054, 0.086]	0.041 [0.036, 0.047]
MSL	0.037 [0.014, 0.091]	0.062 [0.038, 0.100]	0.031 [0.020, 0.048]	0.046 [0.037, 0.057]
SMAP	0.047 [0.023, 0.094]	0.076 [0.050, 0.115]	0.030 [0.020, 0.044]	0.041 [0.036, 0.047]

No dataset shows a rejection rate that increases as the next event approaches, and the band closest to an event lies within the interval of the most distant band in all three datasets.

B.3 Model Precursor Discrimination

B.3.1 Statistical Power

Of the 70 P events, 67 have common support across the seven detectors in RQ2 (SMD 51 to 49, PSM 7 to 6, and no loss for MSL, SMAP, and SWaT). Table 13 traces each dataset from labeled events to the common support used for MPA.

Table 14 reports the smallest MPA that a one-sided event-level test at $\alpha = 0.05$ detects with power 0.8, using the event-level standard deviation implied by the two-stage bootstrap intervals and a normal approximation. The Bonferroni column corresponds to correction over the 35 benchmark cells.

Table 13: Event flow from labeled events to the common support of the seven detectors. Eligible events have uncontaminated pre-onset support within their entity.

Dataset	Labeled	Eligible	P	Common P	Common NoP
SMD	327	320	51	49	249
MSL	36	35	1	1	33
SMAP	67	67	2	2	64
PSM	72	45	7	6	38
SWaT	35	33	9	9	24
Total	537	500	70	67	408

Table 14: Minimum detectable MPA on the common P support.

Dataset	Common P events	MDE ($\alpha = 0.05$)	MDE (Bonferroni)
SMD	49	0.602	0.657
SWaT	9	0.729	0.853
PSM	6	0.791	0.948
MSL	1	–	–
SMAP	2	–	–

The upper 95% bootstrap bound of mean-score MPA^P on SMD is at most 0.656 across the seven detectors. Table 15 reports the evidence-conditioned contrast of Equation (13) for SMD.

Table 15: Evidence-conditioned contrast on SMD, $\Delta_{\mathrm{MPA}} = \mathrm{MPA}^P - \mathrm{MPA}^{\mathrm{NoP}}$, with 95% bootstrap intervals and BH-adjusted q over 35 benchmark cells.

Model	MPA^P	$\mathrm{MPA}^{\mathrm{NoP}}$	Δ_{MPA} [95% CI]	q
iForest	0.553	0.506	0.047 [−0.042, 0.137]	0.61
DAGMM	0.544	0.513	0.031 [−0.050, 0.115]	0.61
USAD	0.471	0.524	−0.053 [−0.157, 0.051]	1.00
LSTM-VAE	0.534	0.490	0.044 [−0.047, 0.135]	0.61
AnomTrans	0.500	0.500	0.000 [0.000, 0.000]	1.00
FCM	0.431	0.449	−0.018 [−0.111, 0.073]	1.00
A2P	0.573	0.469	0.105 [0.013, 0.194]	0.43

Table 16 reports the same contrast for the other four benchmarks. For MSL and SMAP, with one and two P events, the intervals reflect variation across controls only.

For Anomaly Transformer on SMD and PSM, every event-level outcome equals 0.5, so these comparisons show no ranking advantage.

B.3.2 Window Alignment

The test locates SMD evidence within the final minutes before onset, whereas the primary MPA averages scores over the 100-step interval. Table 17 repeats the SMD analysis with the window mean or maximum over the final L steps of the candidate and of each control. Each variant draws its own 20 controls per event, and event-level sign-flip randomization with 10,000 draws is BH-corrected over the 35 benchmark cells within each variant. No cell outside SMD is significant in any variant.

Across the seven variants, the maximum-over-50 cells of LSTM-VAE and DAGMM and the maximum-over-25 cell of LSTM-VAE remain below 0.05 after an additional Bonferroni factor of six for the number of alternative variants.

Selection control The test selects windows that deviate from normal operation, and a detector that scores such deviations can separate selected windows from controls without any relation to a later event. We therefore apply the same computation to held-out normal placebo windows of SMD, treating each window end as a pseudo-onset. Of these pseudo-events, 355 are test-positive and 1,000 randomly drawn test-negative ones serve as a reference. Criteria were fixed before the analysis. A difference in MPA between the 49 P events and the 355 test-positive pseudo-events is considered statistically significant when its bootstrap interval excludes zero.

Table 16: Evidence-conditioned contrast on PSM, SWaT, SMAP, and MSL, with common P and NoP counts in parentheses, 95% bootstrap intervals, and BH-adjusted q over 35 benchmark cells.

Dataset	Model	MPA^{P}	$\mathrm{MPA}^{\mathrm{NoP}}$	Δ_{MPA} [95% CI]	q
PSM (6/38)	iForest	0.675	0.507	0.168 [−0.131, 0.438]	0.61
	DAGMM	0.525	0.467	0.058 [−0.178, 0.270]	0.77
	USAD	0.417	0.476	−0.060 [−0.298, 0.203]	1.00
	LSTM-VAE	0.567	0.449	0.118 [−0.161, 0.365]	0.61
	AnomTrans	0.500	0.500	0.000 [0.000, 0.000]	1.00
	FCM	0.550	0.451	0.099 [−0.187, 0.358]	0.61
	A2P	0.175	0.489	−0.314 [−0.486, −0.122]	1.00
SWaT (9/24)	iForest	0.467	0.504	−0.037 [−0.283, 0.209]	1.00
	DAGMM	0.467	0.452	0.015 [−0.231, 0.259]	0.91
	USAD	0.378	0.444	−0.066 [−0.290, 0.178]	1.00
	LSTM-VAE	0.367	0.460	−0.094 [−0.322, 0.147]	1.00
	AnomTrans	0.389	0.447	−0.058 [−0.230, 0.126]	1.00
	FCM	0.328	0.438	−0.110 [−0.323, 0.134]	1.00
	A2P	0.350	0.406	−0.056 [−0.263, 0.177]	1.00
SMAP (2/64)	iForest	0.775	0.550	0.225 [0.143, 0.305]	0.61
	DAGMM	0.825	0.529	0.296 [0.173, 0.418]	0.61
	USAD	0.875	0.534	0.341 [0.261, 0.421]	0.44
	LSTM-VAE	0.725	0.527	0.198 [0.023, 0.375]	0.61
	AnomTrans	0.362	0.534	−0.172 [−0.222, −0.124]	1.00
	FCM	0.900	0.516	0.384 [0.278, 0.490]	0.52
	A2P	0.850	0.483	0.367 [0.174, 0.561]	0.43
MSL (1/33)	iForest	0.000	0.555	−0.555 [−0.650, −0.456]	1.00
	DAGMM	1.000	0.517	0.483 [0.383, 0.585]	0.43
	USAD	0.000	0.514	−0.514 [−0.621, −0.409]	1.00
	LSTM-VAE	0.800	0.626	0.174 [0.068, 0.283]	0.91
	AnomTrans	0.100	0.595	−0.495 [−0.614, −0.376]	1.00
	FCM	0.800	0.470	0.330 [0.227, 0.433]	0.61
	A2P	0.000	0.350	−0.350 [−0.453, −0.252]	1.00

Table 17: SMD MPA^{P} under aligned windows. Bold marks BH-adjusted $q < 0.05$. The 100-step mean variant uses an independent control draw and differs from Table 3 by at most 0.03.

	Window mean, final L steps				Window maximum, final L steps		
Model	100	50	25	10	50	25	10
iForest	0.553	0.564	0.584	0.607	0.596	0.594	0.591
DAGMM	0.552	0.578	0.588	0.596	**0.662**	**0.635**	0.612
USAD	0.467	0.479	0.471	0.461	0.496	0.506	0.491
LSTM-VAE	0.530	0.545	0.561	0.559	**0.743**	**0.713**	**0.682**
AnomTrans	0.500	0.500	0.500	0.500	0.500	0.500	0.500
FCM	0.451	0.441	0.448	0.461	0.438	0.478	0.454
A2P	0.602	0.592	0.563	0.567	0.566	0.591	0.567

Selection alone raises MPA well above 0.5. The test-positive pseudo-events reach 0.551–0.653 for iForest, DAGMM, and LSTM-VAE in every window-maximum variant ($p \leq 0.002$). LSTM-VAE exceeds this selection level in all three windows, whereas the DAGMM separation is matched by the selected normal windows.

B.3.3 Per-Machine Training on SMD

The seven detectors of the main comparison were trained on pooled entities. A separate experiment fits an iForest to each of the 28 SMD machines on that machine's training data alone, and adds a normal-energy score and an input-independent random score as references. Each machine's training segment is split chronologically into fit, calibration, and verification parts in the ratio 60/20/20, thresholds are the 95th or 99th percentile of calibration scores, and the precursor detectability test uses the fit part as its normal reference. Under this train-only reference the test flags 92 of the 320 eligible SMD events, compared with 51 under the reference of the main text, and 30 of 362 normal pseudo-events.

Table 18: Window-maximum MPA on SMD for P events and for test-positive normal pseudo-events, with the difference and its 95% bootstrap interval. Test-negative pseudo-events lie between 0.49 and 0.51 for both detectors.

Model	Window	P events	Test-positive pseudo-events	Difference [95% CI]
LSTM-VAE	max, final 50	0.743	0.630	0.113 [0.031, 0.192]
LSTM-VAE	max, final 25	0.713	0.608	0.105 [0.018, 0.188]
LSTM-VAE	max, final 10	0.682	0.576	0.106 [0.011, 0.196]
DAGMM	max, final 50	0.662	0.653	0.009 [−0.082, 0.098]
DAGMM	max, final 25	0.635	0.619	0.016 [−0.078, 0.104]
DAGMM	max, final 10	0.612	0.573	0.039 [−0.064, 0.138]

Table 19: Mean-score MPA with per-machine training on SMD, with 95% bootstrap intervals.

Scorer	P events (92)	NoP events (228)	Positive pseudo (30)	Negative pseudo (332)
iForest	0.861 [0.760, 0.935]	0.577 [0.478, 0.696]	0.655 [0.550, 0.763]	0.489 [0.451, 0.526]
Energy	0.874 [0.762, 0.960]	0.568 [0.460, 0.698]	0.647 [0.600, 0.825]	0.508 [0.462, 0.558]
Random	0.495 [0.409, 0.619]	0.506 [0.468, 0.541]	0.552 [0.435, 0.650]	0.513 [0.471, 0.554]

With per-machine training, iForest and energy scores separate P events well above the level reached by test-positive normal pseudo-events, and the random score stays near 0.5 in every group.

B.4 Protocol Comparison

B.4.1 Point Adjustment on the Merged Target

This analysis applies PA to the merged target $P_e \cup A_e$ of Equation (4) for the seven pooled-trained detectors at the 95th-percentile threshold. For every event hit under merged-target PA, it records whether the pre-onset segment P_e contains an alarm. Pooled over the seven detectors, the pre-onset segment contains no alarm for 314 of 699 hits on SMD (0.449), 22 of 48 on PSM (0.458), 28 of 65 on SWaT (0.431), 87 of 281 on SMAP (0.310), and 37 of 181 on MSL (0.204). These events receive warning credit solely from post-onset alarms. Mean F1 on the merged target is 0.175 without PA and 0.570 with PA.

B.4.2 Input-Blind Alarms and Operating Points

Two input-blind alarm sequences are evaluated for every detector and entity. Bernoulli alarms fire independently at the detector's own alarm rate, and circular shifts move the detector's own alarm sequence by a random offset, which preserves the number and length of alarm runs while removing their timing relative to events. Each is drawn 20 times.

Table 20: Mean F1 over the five benchmarks and seven detectors at the 95th-percentile threshold.

Protocol	Detectors	Bernoulli	Circular shift
Detection+PA	0.551	0.740	0.476
Shifted+PA	0.177	0.431	0.167
Merged+PA	0.570	0.788	0.497

Across thresholds at the 90th, 95th, 99th, and 99.9th normal-score percentiles and pre-onset lengths $W \in \{25, 50, 100, 200\}$, Shifted-warning F1 without PA differs from Bernoulli alarms by at most 0.002, and Bernoulli alarms exceed the detectors under Shifted+PA and Merged+PA in all sixteen settings.

B.4.3 Per-Machine Traces under the Three Targets

Table 21 applies the three targets to the per-machine SMD traces of Appendix B.3.3. Late-only counts give the credited events whose pre-onset segment contains no alarm. At the 95th percentile, the per-machine random score hits 316 of the 320 pre-onset windows, in line with the hit probability of about 0.99 for independent alarms at this rate.

Table 21: F1 under Detection, Shifted, and Merged targets with and without PA for per-machine SMD traces.

Scorer	q	Detection		Shifted		Merged		Late-only
		Raw	PA	Raw	PA	Raw	PA	/ credited
iForest	0.95	0.154	0.285	0.083	0.229	0.203	0.464	72/302
iForest	0.99	0.203	0.462	0.068	0.294	0.207	0.600	65/236
Energy	0.95	0.164	0.279	0.083	0.225	0.213	0.456	68/310
Energy	0.99	0.192	0.435	0.074	0.266	0.208	0.618	125/293
Random	0.95	0.046	0.620	0.048	0.652	0.064	0.792	2/318
Random	0.99	0.016	0.822	0.016	0.692	0.018	0.876	41/246

B.4.4 Decision-Oriented Outcomes

Table 22 reports the operation of the per-machine SMD traces over the test period. Early recall counts an event only when a new alarm episode starts within the 100 minutes before onset. Episode precision is the share of alarm episodes that start in an eligible pre-onset window. The calibration budget chooses the threshold that allows at most one episode per 1,000 calibration samples and 1% time under alarm.

Table 22: Test-period operation on SMD with per-machine training, 320 eligible events and 631,031 normal samples. Episodes are normal-data alarm episodes per 1,000 samples, alarm time is the share of normal samples under alarm, and lead is the median first-warning lead in minutes.

Scorer	Threshold	Early recall	Precision	Episodes	Alarm time	Lead
iForest	$q = .95$	62.8%	4.51%	44.0	20.3%	88
iForest	$q = .99$	50.3%	5.38%	21.9	8.72%	73
iForest	budget	26.3%	4.82%	10.0	3.48%	65
Random	$q = .95$	98.8%	4.58%	46.7	4.92%	88
Random	$q = .99$	64.1%	4.48%	9.8	0.99%	61

Early recall alone rewards frequent alarms, as the random score reaches 98.8% at the 95th percentile. The calibration budget holds on every machine's calibration split, yet the test-period burden of iForest is 10.0 episodes per 1,000 samples, which reflects drift between calibration and test periods. Run online at one test per five samples, the precursor detectability test itself would raise 10 to 14 false alarms per day on the minute-level benchmarks and 647 on SWaT, which is sampled every second.