AGGREGATION SCALE SHAPES ANOMALY-DURATION BIAS: BOUNDS, READOUT EFFECTS, AND A NONPARAMETRIC DETECTOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Anomaly duration can change which time-series anomaly detectors perform best, yet aggregate benchmark scores obscure this variation. Across 40 published detectors on TSB-AD, relative performance varies with duration even after controlling for file fixed effects, and all three pretrained forecaster families favour short anomalies. Under stationary normal increments, a shift-invariant statistic reading w consecutive increments has expected coverage at most $\min\{1,\ 2w/L+\alpha\}$ on a flat level shift of length L at false-alarm rate α: detecting an anomaly's presence is distinct from covering its interior. Short-anomaly sweeps reveal empirical dilution under the tested aggregations. Controlled interventions on two frozen pretrained forecasters separate conditioning context, forecast horizon, and score aggregation. Over the tested grid, widening aggregation from 1 to 64 samples at a fixed one-step horizon extends the post-peak half-maximum duration by roughly an order of magnitude in matched configurations with nonzero peak coverage; context has smaller effects. Increasing horizon yields a similar extension. The combined effect is sub-additive on the log-duration scale, and wider readouts reduce peak coverage. On TSB-AD-U, the same aggregation substantially reduces the relative short-anomaly bias of both frozen forecasters. These findings motivate a nonparametric multiscale detector that combines calibrated short- and wide-scale evidence by a maximum. Without gradient training, it achieves competitive offline performance against tuned deep models on TSB-AD-U and TSB-AD-M under the reported development protocol. The results identify aggregation and readout as practical controls on duration bias and support duration-aware evaluation.

1 INTRODUCTION

Time-series anomaly detection is evaluated by aggregate scores on shared benchmarks (Schmidl et al., 2022; Paparrizos et al., 2022b; Liu & Paparrizos, 2024), and those scores are read as measures of detector quality. Aggregate scores also depend on the benchmark's anomaly-duration distribution, the axis beneath the point/pattern split of standard taxonomies (Lai et al., 2021; Blázquez-García et al., 2021). In particular, every pre-trained forecaster on the TSB-AD leaderboard (Ansari et al., 2024; Das et al., 2024; Rasul et al., 2023) — a family built and described as general-purpose — is biased towards short anomalies. Interventions on two frozen forecasters identify residual *readout* as a practical control, with comparatively smaller effects from changing conditioning context.

Published rankings change across duration buckets, and file-fixed-effect analysis shows heterogeneous duration slopes (Section 2). For a statistic reading increments over a fixed window, a flat level shift under stationary normal increments changes the input only at its edges. Expected coverage is at most $2w/L$ plus the false-alarm rate, regardless of capacity or shift amplitude, for that class (Section 3). The tested aggregations also dilute short anomalies within wide windows. These conditional bounds and empirical relations motivate duration-dependent aggregation (Figure 1).

For learned detectors, the receptive field (Luo et al., 2016; Bai et al., 2018) determines available observations; the readout scores the evidence. Zhu et al. (2026) show that longer horizons help long anomalies and hurt short ones, with scores aggregated over the forecast block. We separate conditioning context, forecasting horizon and residual aggregation to distinguish prediction changes

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.

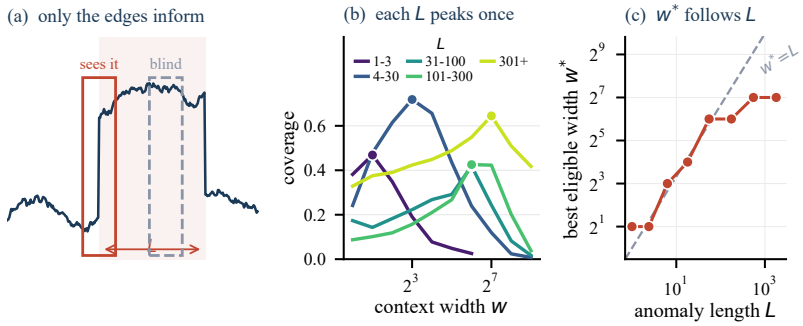

Figure 1: **Aggregation scale and duration.** **(a)** A flat shift changes increments only at its two edges. Under Proposition 2's assumptions, interior windows have the normal law and coverage is at most $2w/L$ plus the false-alarm rate; short-anomaly dilution is measured by (1). **(b)** Raw-window kNN coverage across eligible widths on 2,922 real anomalies. **(c)** The best eligible width tracks duration up to 300 and plateaus at 128 beyond. Eligibility and differing event subsets are detailed in Appendix C.7; x-coordinates are representative bucket positions, not sample means. Panel (a) is schematic. The centred, shift-invariant statistic also tracks duration on injections (Table 13); neither sweep guarantees constant coverage.

Table 1: **Duration effects across published detectors.** Slopes are VUS-PR per e-fold of mean anomaly length, relative to the mean slope over methods, with file fixed effects. CIs: file bootstrap; q: within-source permutation with BH-FDR over all methods. Negative slopes favour short anomalies. Full results: Appendix C.4.

Pretrained forecaster	Slope	95% CI	q
Chronos	-0.070	$[-0.080, -0.059]$	0.004
TimesFM	-0.066	$[-0.076, -0.055]$	0.004
Lag-Llama	-0.047	$[-0.057, -0.037]$	0.020

All 40 detectors, 350 files: 13 significant ($q < 0.05$), 6 short / 7 long.
Within-file R^2: 0.150 $\rightarrow$ 0.216; within-source permutation $p = 0.0002$.

from scoring changes. With predictions held fixed, wider aggregation extends the relative duration frontier (§3.4.3, Table 2) and reduces relative short-anomaly bias on TSB-AD-U (Table 3). Coupled horizon–aggregation widening trades peak coverage for frontier extension. **Context changes have smaller effects than readout changes over the tested grid**.

Our contributions connect three levels of evidence. *Finding:* duration-dependent performance is heterogeneous across 40 published detectors, survives file fixed effects and within-source checks, and places every pre-trained forecaster on the short side (§2). *Mechanism:* a conditional coverage bound, tested with assumption-specific controls, and interventions that separate context, horizon and aggregation and locate the effect in the readout (§3–4). *Construction:* a theory-motivated non-parametric detector that merges calibrated short- and wide-scale evidence by a maximum, without gradient training or per-file tuning; it ranks third of 41 on TSB-AD-U and second of 31 on TSB-AD-M overall, and among the top three by weakest duration bucket on either track under the official bucket boundaries and their perturbations — offline results under the conditions stated in §5.2.

2 Published Rankings Move With Anomaly Duration

The TSB-AD leaderboard (Liu & Paparrizos, 2024) publishes per-file VUS-PR (volume under the precision–recall surface) for 40 detectors on 350 univariate series. Across 23 sources, 80% of log mean anomaly duration variance is between-source. File effects remove method-common file differences and identify each method's duration slope relative to the mean across methods (Table 1) (bucket-level view: Appendix C.3). **Which method leads depends on the durations a benchmark contains, by an amount comparable to the gap between the leading methods.**

Table 1's two-sided interaction survives within-source permutation and the reported leave-one-source-out checks ($p \leq 10^{-3}$). Adding method-by-source intercepts with within-source centred log-duration accounts for source-specific method offsets: signs agree with file-effect contrasts for $32/40$ methods (Spearman $\rho = 0.86$), and all three pretrained forecasters retain negative slopes. In the file-effect model, slope SD is 0.032 VUS-PR per e-fold; over 7.5 e-folds, the steepest implies a 0.59 relative swing (linear extrapolation for scale), versus a 0.19 first-to-tenth gap (Appendix C.4).

Related work. The benchmark, leaderboard and forecasters' strength on point anomalies are Liu & Paparrizos (2024)'s; the measure is Paparrizos et al. (2022a)'s. Yeh (2026) splits results short versus long for one method family; we test graded, two-sided effects across 40 methods with file fixed effects. Multi-scale detectors combine features or scores (Wu et al., 2023; Li et al., 2025); §3.4.3 tests score rules, not those architectures. Sparse-detection and scan boundaries (Donoho & Jin, 2004; Chan & Walther, 2013) provide reference behaviours for the dilution relation. Observational data cannot separate duration from source where a source has one duration; Section 4's injections set duration by design. Appendix D details the contribution relative to prior work.

3 Coverage bounds and aggregation scale

The duration profiles motivate a question about information: which parts of an anomaly can a score distinguish from normal data? Proposition 1 gives the standard bound for any local test; Proposition 2 turns it into a duration-dependent coverage limit under shift invariance and a flat level shift. Short-anomaly dilution is then measured for the tested aggregations. The table distinguishes formal results, empirical findings and architectural hypotheses, together with their assumptions:

Statement	Status; hypotheses	Tested
Prop. 1, total-variation bound	theorem (standard); any w-local test	—
Prop. 2, $2w/L$	theorem; flat level shift, increment-only (shift-invariant) statistic	§4
Eq. (1), short-side dilution	empirical fit; raw-window kNN, pipeline-specific exponents	§4
Prop. 3, residual readout	theorem under exact or ε-approximate adaptation (App. A.8), else hypothesis	§3.4.3
AH1–AH3, architectures	hypotheses; aggregating receptive field	§3.4, §5

Widths. w is the context width of a generic local statistic (Definition 1); C the conditioning width of a forecaster and R its readout width, the samples scored jointly, which §3.4.3 factorises into a horizon H and an aggregation width A; W the rung widths of our detector; L the anomaly duration. The frontier $L_{1/2}$ is the first half-peak crossing after the coverage maximum, interpolated in log-duration; zero peaks are undefined and curves without a crossing are right-censored at the grid limit (Appendix C.5).

Definition 1 (w-local detector)**.** A detector has *context width* w if the score it assigns at time t has the form

$$S_t \;=\; \phi\big(Y_{t-a}, \ldots, Y_{t+b}\big), \qquad a + b + 1 = w,$$

where ϕ is an arbitrary measurable function.

The arbitrariness of ϕ is the point: a hand-designed statistic, a convolutional network and a pretrained transformer are all admissible, differing in how well they use the window, not in which window they may use.

Definition 2 (Coverage at a fixed false alarm rate)**.** Fix the per-sample false alarm rate on normal data at α and let τ_α be the threshold it induces. For an anomalous segment $A = [s, e]$ of length L,

$$\mathrm{cov}_\alpha(A) \;=\; \frac{1}{L}\sum_{t=s}^{e} \mathbb{P}\left[S_t \geq \tau_\alpha\right].$$

Fixing α rather than a detection budget judges each sample against the normal law rather than against its neighbours, so an information deficit is not mistaken for a thresholding artefact.

Proposition 1 (Neyman–Pearson bound)**.** *Let P_w^{N} be the law of a width-w window drawn from the normal regime and $P_w^{\mathrm{A},t}$ the law of the window centred at t inside the anomaly. Then for any detector of context width w operating at false alarm rate α,*

$$\mathbb{P}\left[S_t \geq \tau_\alpha\right] \;\leq\; \alpha + \mathrm{TV}\big(P_w^{\mathrm{N}},\, P_w^{\mathrm{A},t}\big), \qquad \textit{hence} \qquad \mathrm{cov}_\alpha(A) \;\leq\; \alpha + \frac{1}{L}\sum_{t=s}^{e} \mathrm{TV}\big(P_w^{\mathrm{N}},\, P_w^{\mathrm{A},t}\big).$$

The alarm is a measurable function of the window, so the variational characterisation of total variation bounds its power (Appendix A). The right-hand side depends on the two window laws, not the algorithm, parameter count or training set: increasing model capacity cannot overcome limited distinguishability within that view.

3.1 LONG FLAT ANOMALIES

A fixed-width, shift-invariant view can register a level shift at its edges while providing no evidence about the intervening samples.

Proposition 2 ($2w/L$ bound)**.** *Suppose the normal process has stationary increments, so that a window of w increments (a window of $w+1$ samples in Definition 1) has the same law wherever it is taken; that the anomaly is a pure level shift, $\delta_t = h\mathbb{1}[s \le t \le e]$; and that the detector's statistic reads only increments within its window (equivalently, is invariant to a constant shift). Then*

$$\mathrm{cov}_\alpha(A) \;\le\; \alpha + (1-\alpha)\,\min\!\left(1,\; \frac{2w}{L}\right) \;\le\; \min\!\left(1,\; \frac{2w}{L} + \alpha\right).$$

In increments the perturbation occurs at exactly two indices, each affecting at most w windows; every other window has the normal law (Figure 1a). Thus $2w$ bounds the number of informative windows rather than prescribing a decay, and the same count gives Kw for K increment transitions (Remark 5): the class is blind to the interior between transitions, not to long anomalies as such, at every amplitude and for every statistic in the class. A target coverage $c \in (\alpha, 1]$ requires $w \ge L(c-\alpha)/[2(1-\alpha)]$, a necessary scale condition, not a sufficiency guarantee. Detecting *presence* can thus remain easy while *coverage* is limited: stronger boundary evidence does not make the interior distinguishable. Per-sample measures such as VUS-PR reward coverage, so the ceiling is what a benchmark score sees; an event-level measure that credits presence would not.

Remark 1 (The hypotheses identify the mechanism). The bound needs a flat shift and a shift-invariant statistic; changing either removes the guarantee, and Section 4 crosses both factors to distinguish boundary blindness from a generic loss of power. The restriction is not academic: among the short anomalies where our own detector is weakest, between 38% and 75% of anomalous samples carry both a value and an increment within the prefix's observed ranges. On the long side, in the 99 TSB-AD-U files with mean anomaly length above 300, the median file has 99.8% of its anomalous samples carrying an increment inside the prefix's increment range (interquartile 96–100%). This range overlap neither estimates K nor establishes equality of the normal and anomalous increment laws (Appendix C.6).

3.2 SHORT ANOMALIES

The short-anomaly regime imposes the opposite pressure on the tested aggregation: widening it raises the amplitude needed for a fixed coverage. We measure this with raw-window nearest-neighbour scores, broadcast to their samples and merged by a maximum over overlapping windows. Let h^* be the amplitude at which mean coverage reaches 50%, at an empirical 1% false alarm rate on normal-labelled samples. We fit

$$\frac{h^*}{\sigma} \;\propto\; w^{\,\beta_w}\, L^{\,\beta_L}. \tag{1}$$

Section 4 estimates both exponents freely: $(0.69, -0.33)$ on complete cells and $(1.13, -0.38)$ under a censored model, two estimates of this pipeline's scaling rather than bounds on a universal exponent; the positive width effect is shared by both fits. Appendix A.6 compares them with idealised energy and mean tests, which the signal norm Lh^2 motivates, and Appendix D.5 with the scan alternative (Chan & Walther, 2013).

Nothing here binds a general w-local detector: ϕ may ignore all but the last few samples of its window, so no width can be forced to lose a short anomaly. Dilution is a property of an aggregation — a mean, a norm, an occupancy histogram — not of a receptive field, and every claim below is stated for the aggregation. The same argument runs along the channel axis (Proposition 5, Appendix C.1, where its weaker empirical status is reported); it motivates the multivariate construction of Section 5, but no claim here rests on it.

3.3 NO WIDTH IS NEUTRAL

The long-side theorem alone rules out a duration-neutral scale; the short-side measurements say where the best scale should lie.

Corollary 1 (No fixed scale is duration-neutral in this class)**.** *Under the hypotheses of Proposition 2, for any fixed width w and any target coverage $c \in (\alpha, 1]$, every flat shift longer than $L_0 = 2w(1-\alpha)/(c-\alpha)$ has expected coverage below c, whatever its amplitude.*

Remark 2 (Where the optimum lies). The corollary uses the long side only; on the short side (1) measures a cost of widening the raw-window aggregation, and the same exponent need not hold after centring. An optimum near $w \sim L$ is therefore a prediction to test, not a theorem.

Remark 3 (Multiscale design). Complementary duration profiles motivate combining several scales so that different components cover different durations; this also permits adaptive scale selection and does not make a ladder the only construction.

Corollary 2 (A conjunction of alarms is bounded by its narrowest scale)**.** *Let each scale w_i raise its own alarm $A_i = \{S_t^{(i)} \geq \tau_i\}$ at false alarm rate α_i, and let the combined detector alarm only when all do. Then $\mathbb{P}[\bigcap_i A_i] \leq \min_i \mathbb{P}[A_i]$, so if the narrowest statistic satisfies Proposition 2's hypotheses on a flat anomaly, the conjunction's coverage is at most $2w_1/L + \alpha_1$, whatever the wider scales see.*

The corollary is about alarms; for real-valued scores no general statement holds (a constant, uninformative factor leaves a product's ranking intact), so §3.4.3 measures product, sum and maximum at a common empirical false-alarm budget.

To test Remark 2's scale prediction, we measure it twice with the algorithm held fixed and only w swept: on real anomalies with a raw-window nearest-neighbour statistic (w from 1 to 512; amplitude-reading, outside Proposition 2's class) and on controlled injections with its centred, shift-invariant version (1 to 256; inside it, Appendix C.7). On real anomalies the best eligible widths rise across the six buckets up to length 300: $2, 2, 8, 16, 64, 64$ (Figure 1). Both buckets beyond 300 peak at 128 despite wider available rungs, so tracking weakens on the long side (Appendix C.7). On the centred injections $w^* = 2, 2, 4, 8, 32, 16, 64$ for $L = 1, 2, 4, 8, 16, 32, 64$ at 2.83σ, every optimum within a factor of two of the injected duration (Spearman $+0.95$, Table 13): the scale prediction holds for a shift-invariant statistic too.

At its *matched* width, coverage in every real-anomaly bucket lies between 0.33 and 0.92: **each tested duration bucket has a scale that recovers substantial coverage**, although on the centred injections coverage at w^* still falls with length (0.81 to 0.18), so matching locates the best scale without removing all duration dependence. These complementary profiles motivate the multiscale detector of §5.

3.4 FROM CONTEXT WIDTH TO ARCHITECTURES

3.4.1 THE CONSTRAINED QUANTITY IS THE RECEPTIVE FIELD

By Definition 1 the score is an arbitrary measurable ϕ applied to a window, so a transformer with billions of parameters is an admissible ϕ; to apply the propositions to a neural detector we need only identify the input samples that can influence S_t, its *receptive field*. Proposition 1 bounds *every* test measurable in the window, so over-estimating a receptive field is the safe direction (Remark 8); Table 8 (Appendix C.5) gives the quantity playing the role of w in each family, from subsequence length to transformer context.

3.4.2 RESIDUAL DETECTORS HAVE TWO WIDTHS, AND THE THEORY POINTS TO THE READOUT

For forecasting-based anomaly detection, the score is a *residual*,

$$r_t \;=\; d\big(x_t,\, \hat{x}_t\big), \qquad \hat{x}_t \;=\; g\big(x_{t-C}, \ldots, x_{t-1}\big). \tag{2}$$

Such a detector has *two* widths: the **conditioning width** C, how much history the model may look at, and the **readout width** R, how many samples are scored jointly ($R = 1$ for a step-wise residual, the patch size for a score over one reconstructed patch). Scaling is organised around C; the boundary argument says to test R and adaptation speed separately.

Once a conditioning window lies within a long flat anomaly, a forecaster that adapts to the shifted regime can predict its continuation, returning the residual to its normal scale. Evidence then remains near the boundaries. Zhu et al. (2026) describe this mechanism empirically for single-step forecasting (their §5.2); the exactly-adapting case below quantifies its coverage cost, while the interventions test its relevance to real models.

Proposition 3 (Residual readout, informal)**.** *For a residual detector* (2) *whose forecaster adapts to its conditioning window, coverage on a long flat anomaly is governed by the readout width R and the adaptation transient; the conditioning width C enters through that transient. A larger nominal context need not enlarge the boundary region retaining residual evidence.*

Appendix A.8 proves the exact case, with bound $2(C + R)/L + \alpha$ and C upper-bounding the transient, and an approximate form under a block-law total-variation condition (Remark 7). The general reading remains a hypothesis with two testable consequences.

Two predictions. First, *context length beyond the transient does not buy duration coverage if the transient is unchanged*, testable by varying C and R independently. Second, *forecasting foundation models should be short-anomaly specialists*: if their residual aggregates as the statistics of §3.3 do, a small R places them on the short side whatever the context length or pre-training scale — which Section 2 finds without exception, and which Liu & Paparrizos (2024) and Zhu et al. (2026) report as strength on point and weakness on sequence anomalies. The benchmark's implementation scores all three by the squared error of a single forecast step, so each has $R = 1$, TimesFM's output patch notwithstanding (Appendix D).

3.4.3 Architectural predictions and interventions

Read through Table 8, the scale implications of §3.3 suggest three hypotheses about architectures — hypotheses, because a learned score may select within its receptive field rather than aggregate it (Appendix D). **AH1: no aggregating receptive field is duration-neutral** — scaling a fixed field moves its peak rather than flattening the curve; the width sweep and the forty profiles of §2 are its evidence. **AH2: combining complementary scales can broaden duration coverage**; §5 tests this by construction. **AH3: a conjunctive merge of per-scale scores is held down by its blindest scale** — a theorem for alarms (Corollary 2), measured below for real-valued scores, and a prediction for trained networks (Li et al., 2025).

Separating context, horizon and aggregation. The horizon sweep of Zhu et al. (2026) changes both the forecast and the block over which it is scored. We separate them with controlled injections at an empirical false-alarm budget of 1% and predictions registered before each run (Appendix C.5). On two frozen pretrained forecasters, **the relative duration frontier moves much more with readout than with context over the tested grid**: the coupled $R = 1 \to 64$ sweep gives median per-context frontier ratios of 27 and 30 on Chronos-Bolt and TimesFM, against registered context ratios of 1.31 and 1.44 (Table 9).

The factorised follow-up holds forecasts fixed while changing aggregation width A, separately from horizon H. Table 2 reports the per-cell effects: aggregation alone and horizon alone both extend the relative half-maximum duration $L_{1/2}$, with sub-additive combined effects on the log-duration scale. These summaries require nonzero peaks at all four corners and cap positive-peak frontiers at 256. The paired comparison favours aggregation on Chronos but not TimesFM; it does not establish a shared advantage over horizon.

Relative reach trades off against peak coverage (Table 2). Some wide configurations never reach absolute coverage 0.25, so the registered absolute-frontier claim is not confirmed. Aggregation redistributes residual evidence rather than adding predictive information; it does not uniformly improve absolute coverage. The adaptation traces characterise typical residual decay, not the block-law condition of the approximate-adaptation theorem (Appendix C.5).

Readout interventions on the benchmark. On 350 TSB-AD-U files, A64 improves long-anomaly VUS-PR and reverses relative duration slopes at a short-anomaly cost; neither maxA nor maxH dominates (Table 3). Paired slope increases are $+0.078$ $[0.063, 0.091]$ for Chronos and $+0.083$ $[0.068, 0.099]$ for TimesFM (95% file-bootstrap intervals). Context invariance fails

Table 2: Readout factorised (Appendix C.5). Growth in $L_{1/2}$ when aggregation alone ($H{=}1$), horizon alone ($A{=}1$), or both increase from 1 to 64: per-(host, seed, C) factors, with medians computed on the log scale and exponentiated. Paired ratio: exponentiated mean per-cell log-difference (aggregation over horizon), with a host-bootstrap 95% interval. Peak coverage is from a separate coupled ($H{=}A$) sweep, $1 \to 64$ (11 hosts, 2 seeds per model). Paired effects retain 11 Chronos and 5 TimesFM hosts; n/N is retained/total cells for paired effects. Retained cells have nonzero peaks at all four corners; positive-peak frontiers beyond the grid are capped at 256.

Model	Aggregation alone	Horizon alone	Both	Paired ratio [95%]	Peak coverage $1 \to 64$	n/N
Chronos-Bolt	$\times 11.3$	$\times 9.5$	$\times 19.8$	1.41 [1.01, 2.00]	$0.73 \to 0.24$	$81/110$
TimesFM 2.0	$\times 13.0$	$\times 8.4$	$\times 17.0$	0.97 [0.49, 1.45]	$0.81 \to 0.32$	$10/14$

Table 3: **Readout interventions on all 350 TSB-AD-U files.** Official VUS-PR on the shortest and longest duration buckets and overall. A1: one-step absolute residual; A64: the same forecasts, residuals averaged over 64 samples. maxA/maxH combine widths/horizons $1, 16, 64$ by a maximum after label-free normalisation; maxH is our absolute-residual variant of the max-over-horizons rule, not a reimplementation. Slope is relative VUS-PR per e-fold of duration (Section 2); ranks compare each configuration with 40 published methods. Full buckets, A16 and published-model references: Table 11.

Model	Readout	$L = 1$–3	$L > 1000$	Overall (rank)	Slope
Chronos-Bolt	A1	0.817	0.360	0.290 (26)	-0.065
	A64	0.440	0.614	0.408 (9)	$+0.013$
	maxA	0.845	0.439	0.399 (10)	-0.064
	maxH	0.815	0.440	0.378 (15)	-0.057
TimesFM 2.0	A1	0.860	0.360	0.298 (24)	-0.071
	A64	0.442	0.610	0.407 (9)	$+0.012$
	maxA	0.878	0.431	0.403 (9)	-0.069
	maxH	0.843	0.429	0.382 (14)	-0.061

TimesFM's rank clause; the matched one-liner hypothesis fails. Appendix C.5 separates these from the unconfirmed absolute-frontier and host-count checks.

The fusion prescription, measured. The alarm bound does not establish AH3 for real-valued scores. Changing only the merge rule and re-thresholding at the same empirical false-alarm budget, a soft-score product loses 0.08 official VUS-PR (paired interval excludes zero). It also reverses the width-100 statistic's positive duration slope over 1,569 internal-protocol events. Maximum and two-arm average are indistinguishable in official VUS-PR (-0.004, $[-0.014, +0.006]$; Appendices C.8 and C.9). The bounds motivate non-conjunctive fusion without preferring maximum to average; AH3 remains untested for learned merges (Appendix D).

4 VERIFICATION

Figure 2 summarises the injection and amplitude checks; Appendix C.6 gives their designs and the boundary controls.

Injection separates duration from the series. Injections share hosts, seeds, signs and amplitude under fixed segment count or anomalous-sample budget; onsets vary, so neither varies duration alone. The oracle envelope is the per-cell best score among four POINT implementations minus that among five SUBSEQ implementations. Its correlation with duration is negative in all 20 conditions (2 shapes $\times$ 2 conventions $\times$ 5 metrics): median $\rho = -1.00$ to -0.35, every Holm-adjusted $p \leq 4 \times 10^{-7}$ over 39 hosts, with one correlation per host. POINT leads at short durations, SUBSEQ at long. These sets are not pure width classes; the within-algorithm sweep isolates width.

Controls for Proposition 2. Crossing flat/structured anomalies with shift-invariant/amplitude-reading statistics tests the two hypotheses (Remark 1). With tie-safe calibration at empirical normal-sample FPR $\leq 1\%$, bound exceedance is 0.8% where both hold, versus $64.5, 81.2, 94.1\%$ otherwise.

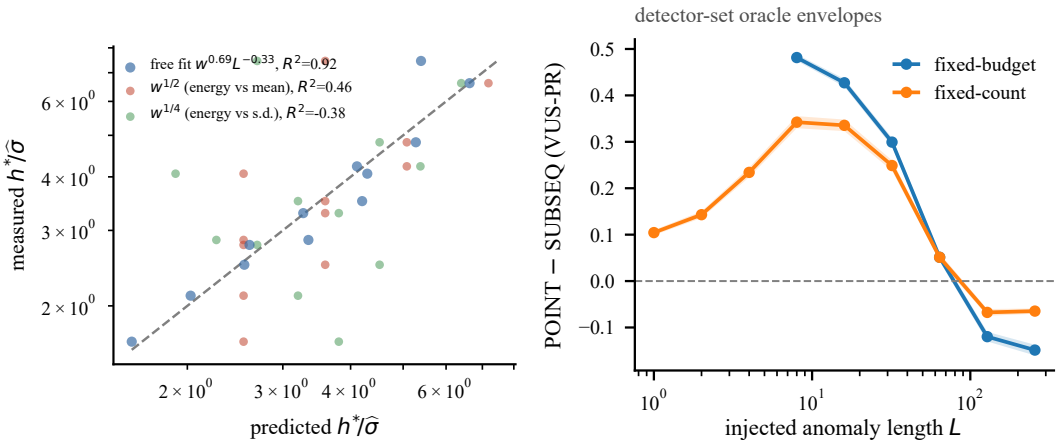

Figure 2: **Controlled verification. Left:** measured versus predicted raw-window threshold $h^*/\widehat{\sigma}$ ($\widehat{\sigma}$: scaled prefix MAD) on 12 complete (w, L) cells; the free fit is $w^{0.69}L^{-0.33}$ ($R^2 = 0.92$). The two energy heuristics are $bw^{1/2}L^{-1/2}$ and $bw^{1/4}L^{-1/2}$ with fitted b. **Right:** POINT-minus-SUBSEQ oracle-envelope VUS-PR, under two prevalence conventions, across 39 hosts and 5 seeds. Means and ± 1 SEM pool host–seed–shape cells; tests use one correlation per host. Onsets vary with duration; detector sets are not pure width classes. Design and censoring-aware fit: Appendix C.6.

Host-bootstrap intervals are $[0.002, 0.017]$ for the first and $[0.528, 0.758]$ for the nearest control; the no-injection rate is 0.1%. On synthetic processes with stationary increments, exceedances are 0.3–0.5%, by at most 0.004 in coverage. For $K \in \{2, 4, 8, 16, 32\}$ transitions the centred arm exceeds the post-hoc broadcast bound in 0.0–0.8% of cases; the original single-window prediction fails (Appendix C.6).

The dilution relation is empirical. The amplitude sweep has 180,999 raw/centred cells. Of 35 raw-window (w, L) combinations with $w > L$, 12 at widths 2–32 reach mean coverage $1/2$; 23 at widths 32 and above do not reach it at 16σ. The complete-case fit is $w^{0.69}L^{-0.33}$ ($R^2 = 0.92$, Figure 2); treating the rest as right-censored gives $w^{1.13}L^{-0.38}$. Both width intervals exclude $1/4$, only the censored interval excludes $1/2$, and both duration intervals exclude -1 and $-1/2$. A ratio-only fit has $R^2 = 0.47$: equal-and-opposite exponents are not supported. The complete-case and censored fits use different data and assumptions (Appendix C.6).

The bounds constrain coverage; the benchmark reports an area. Corollary 3 gives AUC-ROC $\leq \frac{1}{2} + \mathrm{TV} - \frac{1}{2}\mathrm{TV}^2$. For prevalence-dependent VUS-PR the connection is empirical: on 91 real series, coverage at empirical FPR $\leq 1\%$ correlates at $\rho = +0.75$ ($[0.61, 0.85]$; Appendix B).

5 A THEORY-MOTIVATED NONPARAMETRIC DETECTOR

Construction. Our detector, WIDTHLADDER, combines multiple scales with a non-conjunctive merge; the representation and scale set are empirical choices. Per channel, a width-2 3NN term on (increment, value) and five width-W 3NN terms on block occupancy grids ($W \in \{25, \ldots, 400\}$) are calibrated using prefix-score tails; the width-2 term is set against the mean of the five by a maximum, and a maximum is taken over channels (TSB-AD-M adds a joint rung at widths 1, 4, 16); nothing is tuned per file (Appendix B.2).

5.1 OFFICIAL-PROTOCOL RESULTS

Table 4 reports the baseline gaps and diagnostics. The wide ladder retains most performance; the product loses 0.081, while the maximum's advantage over the two-arm average is not significant. By weakest duration bucket, ours ranks third of 41 on U and **first of** 31 on M (second under scaled boundaries; Appendix C.11).

Table 4: **Official results and detector diagnostics.** Published baselines (Liu & Paparrizos, 2024); bold marks our reported detector. Each rank adds one configuration to the 40 U or 30 M baselines. Δ is row minus reported detector on the same track, with a file-paired 95% bootstrap interval. U informed the design; M is run 3 of 4 sequential evaluations, not held out (§5.2). Bottom: existing U ablations and the strict prefix-local control (Tables 15–16); full baseline intervals: Table 17.

Track	Method / configuration	VUS-PR	Rank	Δ [95% CI]
U (n=350)	TimeRCD-MAFT	0.5856	1	$+0.047$ $[+0.011, +0.083]$
	TSPulse (FT)	0.5516	2	$+0.013$ $[-0.024, +0.051]$
	WIDTHLADDER	**0.5390**	**3/41**	—
M (n=180)	StreamVAE	0.4353	1	$+0.021$ $[-0.013, +0.056]$
	WIDTHLADDER	**0.4145**	**2/31**	—
	xLSTMAD	0.3923	3	-0.022 $[-0.054, +0.010]$
U diagnostics (development data; reported VUS-PR 0.5390*)*				
	Wide ladder alone	0.5316	3	-0.007 $[-0.015, -0.001]$
	Two-arm average	0.5353	3	-0.004 $[-0.014, +0.006]$
	Soft-score product	0.4577	6	-0.081 $[-0.109, -0.053]$
	Strict prefix-local	0.4763	6	-0.063 $[-0.094, -0.032]$

5.2 DEVELOPMENT AND EVALUATION CONDITIONS

TSB-AD-U selected the fusion rule and width ladder; TSB-AD-M underwent four registered sequential runs, so neither score is held out (Appendix B.4). Before calibration and the joint rung, the first untouched M run scored 0.3770 (fifth of 31). The detector is offline: whole-series standardisation and bin ranges, plus two prefix-boundary crossings (Appendix B). Table 4's strict prefix-local control measures the cost for this construction, not locality in general (Appendix C.10).

6 LIMITATIONS

From mechanism to design. Of seven design prescriptions registered before running, one succeeded, one improved performance but failed its predictions, and five failed (Appendix C.12). These outcomes concern implementation, not the bound or readout tests. Maximum fusion is not sufficient: the shortest bucket scores 0.46, below the width-2 arm's 0.65. Nor is the duration profile flat (27th of 41 in spread).

Scope. Evidence covers TSB-AD and two pretrained forecasters; (1)'s exponents are pipeline-specific; Proposition 3 requires exact adaptation or a block-law total-variation condition. The channel-width rung gains 0.0174 on TSB-AD-M but fails its predictions. The benchmark applies no common normalisation. Half its detectors, including the three pretrained forecasters, use whole series. These forecasters are scored by one-step squared error (§3.4.2). Our prefix-local control is stricter (Appendix B).

7 CONCLUSION

Anomaly duration is an axis of relative detector performance. Our conditional bound separates detecting a shift's presence from covering its interior; interventions on two pretrained forecasters connect readout scale to a peak–reach trade-off and to reduced relative short-anomaly bias on TSB-AD-U; and a nonparametric multiscale construction achieves competitive offline performance without gradient training. Duration profiles should accompany aggregate scores.

AI USE STATEMENT

We used generative AI tools in this work and describe the division of labour rather than give a blanket disclosure. The research questions, the theoretical claims and their scope, the experimental designs, and every decision about what to claim are the authors'. AI tools were used (i) to search and summarise related work, which the authors then read and verified against the original sources; (ii)

to write and refactor the analysis code and to run and tabulate experiments specified by the authors; (iii) to draft and polish prose from author-specified content and structure; and (iv) to assist with the proofs in Appendix A, which the authors checked line by line.

Two safeguards are worth naming because they bear on the claims. First, every registered prescription was motivated by the theory and its prediction written into the implementing script before that script was run, and no prediction was edited afterwards (Appendix C.12), while the choices selected on TSB-AD-U rather than registered are listed as such (Appendix B.2); this procedure was followed for AI-assisted and hand-written experiments alike. Second, scripts regenerate the main benchmark and experimental tables from archived result files, reducing transcription errors.

We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, its claims and its artefacts.

Ethics statement

This work uses only public benchmark time series and introduces no human subjects, personal data, or new data collection. We see one issue worth naming. The paper's central claim is that published leaderboard rankings partly reflect the anomaly-duration composition of the benchmark rather than detector quality; if that claim is right, it bears on how existing published comparisons should be read. We have therefore stated it as a property of the evaluation rather than a criticism of any method or group, and we report the duration profile of our own detector on the same terms as everyone else's, including the buckets where it is worst.

Reproducibility statement

Code and reproducibility materials will be released on acceptance, including the detector and official-protocol scoring code, experiment and analysis scripts with processed results needed to reproduce the reported tables and figures, the metric self-test, and preserved pre-run prediction records supporting the claims in the paper. Appendix B documents the evaluation protocol, preprocessing, fixed seeds, and development choices; Appendix B.4 describes the sequential evaluation history, and Appendix C.12 summarises the registered prescriptions and their outcomes.

References

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. arXiv:2403.07815.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018. arXiv:1803.01271.

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):56:1–56:33, 2021.

Hock Peng Chan and Guenther Walther. Detection with the scan and the average likelihood ratio. *Statistica Sinica*, 23(1):409–428, 2013.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167, 2024. arXiv:2310.10688.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

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 (ICLR)*, 2026. arXiv:2505.13033.

Arik Ermshaus, Patrick Schäfer, and Ulf Leser. Window size selection in unsupervised time series analytics: A review and benchmark. In *Advanced Analytics and Learning on Temporal Data (AALTD 2022)*, volume 13812 of *Lecture Notes in Computer Science*, pp. 83–101. Springer, 2023.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Towards a rigorous evaluation of time-series anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36(7), pp. 7194–7201, 2022. doi: 10.1609/aaai.v36i7.20680.

B. Ravi Kiran. Multi-scale streaming anomalies detection for time series. *arXiv preprint arXiv:1706.06910*, 2017.

Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Hu. Revisiting time series outlier detection: Definitions and benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 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 *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, New York, 3rd edition, 2005.

Beibu Li, Qichao Shentu, Yang Shu, Hui Zhang, Ming Li, Ning Jin, Bin Yang, and Chenjuan Guo. Crossad: Time series anomaly detection with cross-scale associations and cross-window modeling. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025. arXiv:2510.12489.

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 37 (NeurIPS 2024) Datasets and Benchmarks Track*, 2024.

Xuechun Liu, Heli Sun, Xuecheng Wu, Ruichen Cao, Yunyun Shi, Dingkang Yang, and Haoran Li. Darts: A dual-path robust framework for anomaly detection in high-dimensional multivariate time series, 2025. arXiv:2512.13735.

Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems 29 (NeurIPS 2016)*, 2016.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954.

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, 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.

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, 2023. arXiv:2310.08278.

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.

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 (ICLR)*, 2025.

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 31 (NeurIPS 2018)*, 2018.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

Guenther Walther and Andrew Perry. Calibrating the scan statistic: Finite sample performance vs. asymptotics. *Journal of the Royal Statistical Society Series B*, 84(5):1608–1639, 2022. doi: 10.1111/rssb.12549.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations (ICLR)*, 2023.

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. doi: 10.1109/TKDE.2021.3112126.

Chin-Chia Michael Yeh. Matrix profile for time-series anomaly detection: A reproducible open-source benchmark on tsb-ad, 2026. arXiv:2604.02445.

Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. Matrix profile i: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. In *IEEE International Conference on Data Mining (ICDM)*, 2016.

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 (ICLR)*, 2026. OpenReview forum H27kvyG4qf.

A Proofs

A.1 What is proved and what is argued

Propositions 1, 2 and 5 follow from the variational characterisation of total variation. Equation (1) quantifies the measured amplitude–width–duration relation of the raw-window pipeline; Appendix A.6 explains the signal geometry and the reference tests against which it is compared. Its fitted exponents are empirical. Proposition 4 proves the residual-readout bound under exact adaptation; the general transient statement in §3.4.2 is a hypothesis tested by the model interventions.

A.2 A variational fact

All three theorems reduce to one elementary fact, which we record once.

Lemma 1. *For probability measures P, Q on a measurable space and any measurable φ with $0 \le \varphi \le 1$,*

$$\mathbb{E}_Q[\varphi] - \mathbb{E}_P[\varphi] \;\le\; \mathrm{TV}(P, Q).$$

Proof. $\mathrm{TV}(P,Q) = \sup_B |P(B) - Q(B)|$ over measurable B, and this supremum equals $\sup_{0\le\varphi\le1} |\mathbb{E}_P[\varphi] - \mathbb{E}_Q[\varphi]|$: the right-hand side dominates the left by taking $\varphi = \mathbb{1}[B]$, and is dominated by it because $\mathbb{E}_Q[\varphi] - \mathbb{E}_P[\varphi] = \int \varphi \,\mathrm{d}(Q - P)$ is maximised over $[0,1]$-valued φ by the indicator of the set where the signed measure $Q - P$ is positive. □

The reason to phrase the argument this way, rather than through the Neyman–Pearson lemma directly, is that Lemma 1 covers randomised decision rules with no extra work. A detector that breaks ties at random, or that thresholds a stochastic score, is still an admissible φ.

A.3 Proposition 1: the Neyman–Pearson bound

Proof. Fix t. The detector assigns $S_t = \phi(Y_{t-a}, \dots, Y_{t+b})$ and raises an alarm when $S_t \geq \tau_\alpha$. Write $\varphi = \mathbb{1}[\phi(\cdot) \geq \tau_\alpha]$, a measurable $[0,1]$-valued function of the width-w window and nothing else. Operating at false alarm rate α means $\mathbb{E}_{P_w^{\mathrm{N}}}[\varphi] = \alpha$. Under the anomalous law the alarm probability is $\mathbb{E}_{P_w^{\mathrm{A},t}}[\varphi]$, so Lemma 1 gives

$$\mathbb{P}\left[S_t \geq \tau_\alpha\right] \;=\; \mathbb{E}_{P_w^{\mathrm{A},t}}[\varphi] \;\leq\; \alpha + \mathrm{TV}\big(P_w^{\mathrm{N}}, P_w^{\mathrm{A},t}\big).$$

Averaging over $t \in [s, e]$ and applying Definition 2 gives the statement for $\mathrm{cov}_\alpha(A)$. □

Two remarks on what the proof does not use. It does not use any property of ϕ beyond measurability, so it holds for a learned ϕ of arbitrary capacity trained on arbitrary amounts of data. And it does not use any property of the threshold beyond the false alarm rate it induces, so it cannot be evaded by a cleverer calibration.

A.4 Proposition 2: the $2w/L$ bound

Proof. Let $Y_t = X_t + \delta_t$ with $\delta_t = h\,\mathbb{1}[s \leq t \leq e]$, and let $\Delta Y_t = Y_t - Y_{t-1}$. Then

$$\Delta Y_t \;=\; \Delta X_t + h\big(\mathbb{1}[t = s] - \mathbb{1}[t = e+1]\big), \tag{3}$$

so the increment sequence differs from the unperturbed one at exactly two time indices, s and $e+1$, whatever the value of $L = e - s + 1$.

By hypothesis the statistic at time t is a function of the increments in a window W_t of w consecutive indices. If $s \notin W_t$ and $e+1 \notin W_t$, then by (3) the restriction of $(\Delta Y_\tau)_{\tau \in W_t}$ to that window is identically equal to $(\Delta X_\tau)_{\tau \in W_t}$, hence — by the stationarity of increments assumed in the statement, which is what lets a window at one location stand for a window drawn anywhere in the normal regime — equal in law to a normal window: $P_w^{\mathrm{A},t} = P_w^{\mathrm{N}}$ and $\mathrm{TV} = 0$. Proposition 1 then gives $\mathbb{P}[S_t \geq \tau_\alpha] = \alpha$ exactly, not merely as a bound.

Each of the two indices s and $e+1$ belongs to at most w of the windows $\{W_t\}$, so at most $m = \min(L, 2w)$ of the L anomalous time points escape the argument. Bounding the alarm probability by 1 on those and by α on the remaining $L - m \geq 0$,

$$\mathrm{cov}_\alpha(A) \;\leq\; \frac{m + (L-m)\,\alpha}{L} \;=\; \alpha + (1-\alpha)\frac{m}{L} \;\leq\; \alpha + (1-\alpha)\min\left(1,\; \frac{2w}{L}\right),$$

which is at most $\min(1,\, 2w/L + \alpha)$, the form quoted in the abstract. □

Remark 4 (Where the hypotheses bite). Both hypotheses are load-bearing and neither is cosmetic. If the anomaly is not flat, its interior carries structure and $P_w^{\mathrm{A},t} \neq P_w^{\mathrm{N}}$ there. If the statistic is not shift-invariant — an amplitude statistic, say — then a level shift outside the normal amplitude range is visible at every interior point and the bound does not apply. The 2×2 design in Section 4 examines the role of each hypothesis. Outside this class the bound is not guaranteed, although it may still hold. The experiments broadcast window scores to samples over overlapping windows (Appendix C.6); the count above survives that, since the windows touching a boundary reach at most w samples into the anomaly.

Remark 5 (More than two transitions, and the general count). Nothing in the proof uses that there are two boundaries. Let the anomaly be piecewise constant, $\delta_t = \sum_{j=1}^{K} a_j\,\mathbb{1}[t \geq s_j]$ with change points $s = s_1 < \cdots < s_K = e+1$ (a flat shift is $K = 2$, $a_2 = -a_1$). Then $\Delta Y_t = \Delta X_t + \sum_j a_j \mathbb{1}[t = s_j]$ departs from the normal increments at exactly K indices, each in at most w windows, so $m \leq \min(L, Kw)$ and

$$\mathrm{cov}_\alpha(A) \;\leq\; \alpha + (1-\alpha)\min\left(1,\; \frac{Kw}{L}\right).$$

What a shift-invariant statistic of width w cannot cover is the interior between transitions, whatever the anomaly's length; Kw is again an upper bound on the informative windows, not a statement that coverage grows with K. The noise-type structured anomaly of Section 4 changes the increments at almost every index, so K is of the order of L and the bound is vacuous there — the contrast the 2×2 design measures. For an anomaly of arbitrary shape the same averaging gives a total-variation form: with $d_t = \mathrm{TV}(P_w^{\mathrm{N}}, P_w^{\mathrm{A},t})$ and $m_\varepsilon = |\{t \in A : d_t > \varepsilon\}|$, Proposition 1 applied pointwise and averaged over the L anomalous samples yields $\mathrm{cov}_\alpha(A) \le \alpha + \varepsilon + (1 - \alpha - \varepsilon)\, m_\varepsilon / L$ whenever $\alpha + \varepsilon \le 1$; the proposition is the case $\varepsilon = 0$, $m_0 \le 2w$. We state these as remarks because they are the same argument. The K-transition test of Appendix C.6 uses $K \in \{2, 4, 8, 16, 32\}$. For its broadcast score, an interior transition can reach at most $2w - 1$ samples rather than w, giving the informative-count upper bound $2w + (K-2)(2w-1)$. The appendix reports empirical exceedances separately from the original single-window prediction.

A.5 From a coverage bound to a reported score

The bounds constrain coverage at a fixed false alarm rate; the leaderboard reports an area. For segment-level coverage, let P_w^{A} denote the uniform mixture $L^{-1} \sum_{t \in A} P_w^{\mathrm{A},t}$. Then $\mathrm{TV}(P_w^{\mathrm{N}}, P_w^{\mathrm{A}}) \le L^{-1} \sum_{t \in A} \mathrm{TV}(P_w^{\mathrm{N}}, P_w^{\mathrm{A},t})$, so either the mixture distance or this average supplies a common bound across operating points.

Corollary 3 (The whole ROC curve is pinned)**.** *If* $\mathrm{TV} = \mathrm{TV}(P_w^{\mathrm{N}}, P_w^{\mathrm{A}})$ *bounds the detector at every false alarm rate, then its receiver operating characteristic satisfies* $\mathrm{TPR}(\alpha) \le \min(1, \alpha + \mathrm{TV})$ *for all* $\alpha \in [0, 1]$*, and therefore*

$$\text{AUC-ROC} \;\le\; \tfrac{1}{2} + \mathrm{TV} - \tfrac{1}{2}\mathrm{TV}^2.$$

Proof. Proposition 1 is stated for an arbitrary operating point, so it holds at every α simultaneously; integrating the pointwise bound, $\int_0^1 \min(1, \alpha + \mathrm{TV})\, \mathrm{d}\alpha = \int_0^{1-\mathrm{TV}} (\alpha + \mathrm{TV})\, \mathrm{d}\alpha + \mathrm{TV} = \frac{1}{2} + \mathrm{TV} - \frac{1}{2}\mathrm{TV}^2$. □

The same step does not reach a precision-based area, because precision depends on prevalence as well as on the two laws, and the benchmark's files differ widely in prevalence. For VUS-PR we therefore rely on measurement rather than implication, and Section 4 reports it.

A.6 The short-anomaly dilution relation

Signal geometry is exact; detection thresholds require a score law. Let $b \in \{0, 1\}^w$ indicate the $L < w$ shifted coordinates, all contained in the window. The signal has $\|hb\|^2 = Lh^2$. For an unperturbed window x and a fixed reference r,

$$\|x + hb - r\|^2 - \|x - r\|^2 = 2h\langle x - r, b\rangle + Lh^2.$$

Thus Lh^2 is the signal's squared norm, not the realised change in every distance. For window centring, $P = I - \mathbf{1}\mathbf{1}^\top / w$,

$$\|P(hb)\|^2 = h^2 L(1 - L/w).$$

When $L \ll w$ the leading signal energy is similar before and after centring; this alone does not equate their detection thresholds or exponents.

Reference tests under independent Gaussian noise. For known zero mean and $x \sim \mathcal{N}(0, \sigma^2 I)$, an energy test uses $Q = \|x\|^2 / \sigma^2$. Its null law is χ^2_w and its shifted law is noncentral $\chi^2_w(\lambda)$ with $\lambda = L(h/\sigma)^2$. At fixed false alarm rate α and fixed power $q \in (\alpha, 1)$, the large-w reference scale is $\lambda \asymp \sqrt{w}$, since the null fluctuation is of order $\sqrt{w}$ and the alternative variance has the same order in this regime. Consequently $h^*/\sigma \asymp w^{1/4} L^{-1/2}$. Comparing the signal instead with the null *mean*, $w\sigma^2$, yields $\sqrt{w/L}$, but that is a signal-to-background heuristic, not a fixed-level detection threshold. A signed mean has signal-to-noise $Lh/(\sigma\sqrt{w})$ and instead gives $h^*/\sigma \propto w^{1/2} L^{-1}$.

What the fitted exponents measure. Our statistic is the mean distance to three nearest prefix windows, followed by broadcasting and a maximum over overlapping windows; it is not the energy test above. Its threshold is an empirical normal-labelled quantile, and h^* is defined by mean sample coverage. Neighbour selection, host dependence, width-dependent reference sets, and the score-to-sample mapping enter this pipeline. Their individual contributions to the exponent have not been isolated. Section 4's complete-case and censored estimates $(0.69, -0.33)$ and $(1.13, -0.38)$ both show a positive width cost. The complete-case width interval $[0.46, 1.00]$ admits $\frac{1}{2}$ and excludes $\frac{1}{4}$; the censored interval $[0.87, 1.60]$ excludes both under its additional assumptions. These comparisons locate the empirical result relative to reference tests; they do not derive its numerical exponents or reject those tests in their own settings.

Unknown position: the scan boundary. The energy and signed-mean tests above aggregate the whole window. The sparse-detection literature's reference statistic instead *scans* for the displaced block, and its boundary is worth deriving for our parameterisation because it predicts a width dependence of a different kind. Let the window be white noise of scale σ and let the L displaced coordinates form a contiguous block at an unknown offset s among $m = w - L + 1$ candidates. For each candidate the standardised block mean T_s is $\mathcal{N}(0, 1)$ under the null and $\mathcal{N}(h\sqrt{L}/\sigma,\, 1)$ at the true offset; the generalised likelihood-ratio statistic is $T = \max_s T_s$. By the union bound and the Gaussian tail, the level-α threshold obeys $\tau_\alpha \le \sqrt{2\log(m/\alpha)}$, and since the candidates contain $\lfloor w/L \rfloor$ disjoint, hence independent, blocks, $\tau_\alpha \ge \sqrt{2\log(w/L)}\,(1 - o(1))$ as $w/L \to \infty$ at fixed α. If the true block's mean equals τ_α then that block alone alarms with probability one half, and $T \ge T_{s^*}$, so the scan's power is at least one half:

$$\frac{h^*}{\sigma}\sqrt{L} \;\le\; \sqrt{2\log\big((w - L + 1)/\alpha\big)}. \tag{4}$$

This is a finite-sample upper bound on the amplitude sufficient for half-power. We state no matching lower bound: other candidates can alarm, so the scan's power at that amplitude exceeds the single block's (for $w = 2$, $L = 1$, $\alpha = 0.05$ it is 0.513), and the regime in which the true block's term dominates the maximum — how w and L grow, whether L is known or scanned over, how the scales are calibrated — is the subject of the scan literature (Chan & Walther, 2013; Walther & Perry, 2022), whose results concern presence detection with one alarm per series and do not transfer to per-sample coverage without a rule for which samples an alarmed block covers. What we take from that literature is only the reference shape of its model, $h^*\sqrt{L}/\sigma \propto \sqrt{\log(w/L)}$ up to constants: a duration exponent of $-\frac{1}{2}$ and a width that enters through a logarithm — a multiplicity price, not a dilution one. Appendix D.5 fits that shape to the measured thresholds and finds that, over the swept range, the data prefer the power law.

None of these whole-window comparisons constrains an arbitrary w-local detector. Definition 1 permits ϕ to read only the last few samples of its window, so widening a receptive field cannot force a loss; what dilutes is an aggregation that is obliged to use the whole window.

A.7 Proposition 5: channel width

Proof. Let $\mathcal{C}$ be the channel set carrying the anomaly and suppose that for every $\mathcal{K} \subsetneq \mathcal{C}$ with $|\mathcal{K}| = k$, the law of the anomalous window restricted to the channels in $\mathcal{K}$ coincides with the normal law restricted to $\mathcal{K}$. A statistic reading only the channels in $\mathcal{K}$ is a measurable function of that restriction, so $P_w^{\mathrm{A},t} = P_w^{\mathrm{N}}$ as laws on the restricted space and $\mathrm{TV} = 0$. Proposition 1 gives detection probability α. □

The structure is identical to Proposition 2 with the channel axis in place of the time axis: in both, the anomaly is invisible to every view narrower than the one in which it is written. We note that the hypothesis — an anomaly whose every proper marginal is normal — is strong, and that we verify Proposition 5 only partially (Section 6).

A.8 Proposition 3: residual readout

Section 3.4 separates conditioning from residual readout and tests their effects independently. Here we prove interior indistinguishability under exact adaptation, yielding a coverage bound containing both widths. Their relative empirical effects are assessed in Appendix C.5.

Definition 3 (Exactly adapting forecaster)**.** A forecaster g is *exactly adapting* for the level shift $\delta_t = h\mathbb{1}[s \le t \le e]$ if, whenever its conditioning window lies entirely inside $[s, e]$, its prediction satisfies $\hat{x}_t(Y) = \hat{x}_t(X) + h$, where $\hat{x}_t(X)$ is the prediction it would have made on the unperturbed series.

The definition says only that a forecaster which sees nothing but the shifted regime forecasts the shifted regime. It is the formal content of "the anomaly has become the model's new normal".

Proposition 4 (Residual readout, exact case)**.** *Let g be exactly adapting in the sense of Definition 3, with conditioning window $\{u - C, \dots, u - 1\}$ at time u, and let $r_u = d(x_u, \hat{x}_u)$ for a translation-invariant discrepancy d. Let J be a fixed set of R consecutive integer offsets containing 0, and define the readout $J_t = t + J$ and score $S_t = \phi((r_u)_{u \in J_t})$ for a fixed measurable ϕ. Suppose the normal law of S_t is the same at every t and a single threshold gives false-alarm rate α under that law. If $\{u - C, \dots, u\} \subseteq [s, e]$ for every $u \in J_t$, then the score has its normal law and alarm probability α. Consequently,*

$$\mathrm{cov}_\alpha(A) \;\le\; \frac{2(C+R)}{L} + \alpha.$$

Proof. For every $u \in J_t$, the condition places both x_u and its entire conditioning window inside the shift. Exact adaptation and translation invariance give $r_u(Y) = r_u(X)$ for all these u jointly, hence $S_t(Y) = S_t(X)$ pathwise and alarm probability α. The complete input interval is $D_t = \{\min J_t - C, \dots, \max J_t\}$, of width $C + R$, and contains t. Thus $D_t \not\subseteq [s, e]$ can occur only within $C+R$ samples of an anomaly boundary, affecting at most $\min(L, 2(C+R))$ anomalous time points. Bounding their alarm probability by 1 and averaging as in Proposition 2 gives the stated conservative bound. □

Remark 6 (Why this does not make C helpful). The bound contains C, so read literally it grows with the conditioning width — the opposite of the usual expectation that more context helps. The mechanism is the one described in Section 3.4: conditioning is what allows the model to absorb the anomalous regime, and the escape region is the part of the segment where it has not yet done so. A longer transient can retain residual evidence over more samples, but need not yield higher coverage. Exact adaptation does not say the transient equals C: the persistence forecaster $\hat{x}_t = x_{t-1}$ satisfies Definition 3 for every C and adapts in one step, so C is a conservative upper bound on the transient, not its value. The bound is an upper bound: that it loosens with C says nothing about coverage rising with C, and the adaptation reading is an empirical claim, tested in Appendix C.5, not a consequence of the inequality.

Remark 7 (Approximate adaptation). Exact adaptation can be relaxed without changing the argument, provided the relaxation is stated on the block the score reads. For a readout of R samples let P_R^{N} be the normal law of a width-R residual block and $P_{R,t}^{\mathrm{A}}$ the law of the block read out at t inside the anomaly. Call the forecaster *ε-adapting with transient τ_ε* if $\mathrm{TV}(P_R^{\mathrm{N}}, P_{R,t}^{\mathrm{A}}) \le \varepsilon$ for every t whose block lies more than τ_ε samples from both s and $e + 1$. Proposition 1 then gives alarm probability at most $\alpha + \varepsilon$ at every such t, at most $2(\tau_\varepsilon + R)$ anomalous samples escape, and the averaging of Remark 5 gives

$$\mathrm{cov}_\alpha(A) \;\le\; \alpha + \varepsilon + (1 - \alpha - \varepsilon)\,\min\!\left(1,\; \frac{2(\tau_\varepsilon + R)}{L}\right), \qquad \alpha + \varepsilon \le 1.$$

Proposition 4 is the case $\varepsilon = 0$ with $\tau_0 \le C$. The transient may depend on the readout, horizon and model; it is not identified by the nominal conditioning width. The condition must control the joint block law. Uniform per-step conditional TV bounds at every shared past give at most $\min(1, R\varepsilon)$ by sequential coupling, which can become vacuous as R grows; marginal residual laws alone do not suffice. The threshold-based time measured in H6 (Appendix C.5) is a descriptive proxy: it neither estimates τ_ε nor verifies the block-law TV condition. That proxy is about 34–50 samples on the pretrained models; the TCN does not meet the threshold within 256 samples, a censored observation. The bound does not predict the half-maximum frontier: at $R = 1$, the measured frontier (5–8 samples) is below the proxy transient, and depends on the early evidence decay and the peak-relative definition.

Remark 8 (Information, not estimation; and which way to err). Proposition 1 bounds the power of *every* test measurable with respect to the window, so it cannot be escaped by a better estimator, a larger model, more pre-training data or a better loss: if $\mathrm{TV}(P_w^{\mathrm{N}}, P_w^{\mathrm{A}})$ is small the distinguishing

samples are not in the window, and no function of the window can recover them. It follows that overestimating a receptive field is the safe direction. A bound for width w also holds for any detector whose receptive field fits inside a width-w window, so where the effective field is smaller than the nominal one — attention, whose support is data-dependent, or a recurrent state, whose influence decays — the nominal width gives a weaker but still valid bound. Claiming a width *smaller* than the true receptive field would be the unsafe direction, and we never do.

Remark 9 (What is assumed and what should be measured). Exact adaptation guarantees equality to the normal residual in the interior; it does not order forecasters by adaptation speed or specify their boundary transients. Real models may adapt partially or retain residual evidence. The general transient interpretation in §3.4.2 is therefore empirical. The registered tracker test of §3.4.3 has four predictions; under tie-safe calibration three are confirmed and one is not (Appendix C.5).

B Evaluation protocol

B.1 What we ran

Results are computed under the official TSB-AD protocol, so that they can be placed beside the published leaderboard without adjustment.

- **File lists.** The official evaluation lists: 350 univariate series for TSB-AD-U and 180 multivariate series for TSB-AD-M.
- **Scoring region.** The whole series is scored, with no truncation. Each file names its own learning prefix; the nearest-neighbour references and the calibration tails are fitted on it, with two exceptions at block boundaries stated in Appendix B.2. Label-free quantities are taken from the whole series: the histogram bin ranges (minimum and maximum of value and increment) on both benchmarks, and on TSB-AD-U also the per-series standardisation (mean and standard deviation; on TSB-AD-M it is the prefix's). These are offline (transductive) preprocessing: the score is not a function of its window and the prefix alone, and Appendix C.10 measures what a strictly prefix-local version costs.
- **Sliding window.** The window that the metric uses is set per series by the benchmark's own autocorrelation rule rather than by us. We ported that rule from the reference implementation; across the 350 univariate files it returns a median of 90 samples, ranging from 4 to 299.
- **Metric.** VUS-PR, with the reference threshold count.
- **Baselines.** Every competing number is the published per-file value from the leaderboard, not a re-implementation. This is deliberate: it avoids errors from reimplementing baseline detectors and lets a reader check the duration analysis of Section 2 without running any of our code. The numbers are the benchmark repository's per-file result tables merged with the community submissions listed there, as downloaded on 2026-09-04: 40 entries on TSB-AD-U and 30 on TSB-AD-M (the web table lists 39 on TSB-AD-U, omitting the plain LOF entry of the result files, which ranks far below ours). The merged tables are released with the code, so the ranks refer to that snapshot rather than to the live page. The benchmark's reference scripts (`benchmark_exp/Run_Detector_U.py` and `TSB_AD/model_wrapper.py`, read on 2026-09-18) apply no normalisation to the data — the one exception is StreamVAE, which standardises on the training slice inside its wrapper — and divide the detectors into an unsupervised pool, which receives the whole series and no prefix (by that classification 20 of the 40 published TSB-AD-U entries, including Chronos, TimesFM, Lag-Llama, MOMENT (ZS), TSPulse (ZS) and Time-RCD), and a semi-supervised pool, which is fitted on the prefix and scores the whole series. Our whole-series preprocessing therefore lies within the leaderboard's own convention, and the strict prefix-local control of Appendix C.10 is stricter than the protocol requires of any entry.

Metric implementation. We use a vectorised implementation of VUS-PR rather than the reference one, because the duration analyses require scoring the same series many times over. Two checks are included with the code: a self-test on 18 synthetic settings with ties and mixed anomaly lengths, and a comparison on every series of both benchmarks — the real label vectors and sliding

windows with a deterministic synthetic score — which agrees with the reference to 10^{-9} on all 530 files (in fact to 2×10^{-15}; script and log are in the archive).

Hardware and cost. Scoring is CPU-only, with no training phase and no GPU at any point. All 350 univariate series take 151 seconds of wall-clock on a 64-core, 128-thread workstation and all 180 multivariate series 690 seconds. Those are parallel figures, so we also give the serial cost, obtained by timing a random sample of files on one core and extrapolating by total series size: about 20 minutes for the univariate benchmark (20 files, 4% of bytes) and about three hours for the multivariate one (16 files, 15%). The latter is dominated by a few series with up to 248 channels and 650,000 samples and the per-file spread is wide, so the serial figures are orders of magnitude rather than measurements. An earlier five-file sample extrapolated to 126 minutes where the sixteen-file re-measurement gives 183; we report the larger sample and the caveat rather than the smaller number. All timings are on one workstation: an AMD Ryzen Threadripper 3990X (64 cores, 128 threads, 2.9 GHz base) with 256 GB of RAM; the pretrained-forecaster experiments of Appendix C.5 used a single NVIDIA GeForce RTX 4090 (24 GB), and everything else ran on the CPU.

B.2 The detector's construction

Let $x_{1..n}$ be one channel, p the length of its labelled-normal learning prefix, and $z_t = (x_t - \bar{x})/\mathrm{sd}(x)$ the standardised series. On TSB-AD-U the mean and standard deviation are those of the whole series; on TSB-AD-M they are the prefix's, and only the bin ranges below read the whole series. Both are label-free, neither is prefix-local (Appendix C.10). The width-2 reference set is drawn from $t \le p$ and contains those prefix samples that survive its subsample, so such a sample's own distance of zero is one of its three neighbours; the calibration tails are fitted on those self-inclusive prefix scores.

1. **Width-2 term.** $u_t = (z_t - z_{t-1},\, z_t)$; $s_t^{(2)}$ is the mean Euclidean distance from u_t to its 3 nearest neighbours among at most 4,000 prefix points (uniform subsample, fixed seed). The subsample matters per file: a different draw of it, on the 112 files whose prefix exceeds 4,001 samples, moves a file's VUS-PR by up to 0.062 and the benchmark mean by 10^{-5}, so per-file numbers of the detector are reproducible only with the seed and draw order fixed, as the archive fixes them.
2. **Width-W terms**, $W \in \{25, 50, 100, 200, 400\}$. The series is cut into consecutive blocks of W (the first increment of a block uses the last sample of the previous one). Each block is summarised by a 5×5 binary occupancy grid over (increment, value): each axis's whole-series range is divided into four equal intervals, every sample spreads bilinear weights over the four grid nodes around it, and a node is occupied if its total weight is positive. $s^{(W)}$ of a block is the mean Euclidean distance between its 25-vector and those of its 3 nearest reference blocks, assigned to every sample of the block. The reference set is the first $\max(3, \lfloor p/W \rfloor)$ blocks: when the prefix holds fewer than three complete blocks it reaches past the prefix (111 of 350 TSB-AD-U files, at $W = 200$ or 400), and the block containing sample p straddles the prefix end, so its score — computed partly from samples after p — enters the calibration sample of the next item (334 files). Neither behaviour is prefix-local; both are removed in the strict control of Appendix C.10, and the reported detector is an offline method in this respect too.
3. **Calibration.** For a statistic a with sorted prefix values $a_{(1)} \le \cdots \le a_{(k)}$, $\mathrm{cal}(a) = -\log\big((1 + \#\{i : a_{(i)} \ge a\})/(k+1)\big)$, and for $a > a_{(k)}$ it is continued as $-\log(1/(k+1)) + (a - a_{(k)})/\mathrm{MAD}$, where MAD is the scaled median absolute deviation of the prefix values plus 10^{-9} (a zero MAD makes the continuation steep, not undefined). Ties share a value; a prefix with fewer than 20 values falls back to a robust z-score; no constant is tuned.
4. **Per channel.** $S_t^{\mathrm{ch}} = \max\big(\mathrm{cal}(s_t^{(2)}),\ \frac{1}{5}\sum_W \mathrm{cal}(s_t^{(W)})\big)$.
5. **Across channels.** $S_t = \max_{\mathrm{ch}} S_t^{\mathrm{ch}}$. On TSB-AD-M a joint rung is added inside the maximum: the width-w nearest-neighbour distance of the full channel vector to prefix windows, per-channel MAD-scaled, for $w \in \{1, 4, 16\}$, each calibrated as above. With d channels, each window is flattened to a dw-vector. Reference windows lie wholly in the prefix, at stride $\max(1, \lfloor w/2 \rfloor)$; at most 300 are sampled uniformly without replacement,

using the file's seed-0 random generator after the per-channel terms. The score is the mean Euclidean distance to the three nearest references, including self-matches. Query windows start at stride $s = \max(1, \lceil n/Q \rceil)$, where $Q = \max(200, \lfloor 3 \times 10^9/(300dw) \rfloor)$ is a fixed computational-budget rule. Each score is broadcast from its start over $\max(w, s)$ samples, clipped at the series end; overlaps take the maximum, and uncovered samples receive the minimum covered score. A rung is omitted if $w \geq \max(\lfloor p/4 \rfloor, 4)$, $w \geq \lfloor n/4 \rfloor$, or fewer than ten reference windows are available before subsampling. If all joint rungs are omitted, the per-channel maximum is retained.

The table separates theoretically motivated choices from choices selected on TSB-AD-U:

Choice	Status	Where it comes from
several scales, merged disjunctively	motivated	Remark 3; Corollary 2
maximum rather than product or sum	selected on TSB-AD-U	fusion comparison, Table 14
mean over the five wide terms	selected on TSB-AD-U	Table 14, "averaged ladder" rows
width-2 amplitude-reading term	outside Prop. 2's class; selected on TSB-AD-U	§3.3
scale set $\{25, \ldots, 400\}$, occupancy histogram	empirical, fixed early on TSB-AD-U	—
calibration by prefix tail	design (commensurability)	§B.2
joint-channel rung, $w \in \{1, 4, 16\}$	motivated by Prop. 5; registered before its own TSB-AD-M run (run 3 of App. B.4), after runs 1–2 had been seen	Appendix C.1

Calibration is what makes the disjunction fair across statistics with different tail shapes — without it a heavy-tailed arm dominates the maximum regardless of evidence. The width-2 term reads amplitude and so lies outside the class Proposition 2 covers. It preserves absolute-level information; this is a design choice, not a necessary condition for detecting short anomalies. The detector's own long-anomaly coverage is therefore not bounded by $2 \cdot 2/L$.

B.3 Which numbers come from which protocol

Three kinds of measurement appear in this work and they must not be mixed. The **official protocol** above is the leaderboard's: whole series, VUS-PR, published baselines. The **internal, event-level protocol** analyses individual anomaly events on the real series, compares within a series so that dataset identity is controlled, and measures coverage at a fixed empirical false-alarm budget; it is the right instrument for a claim about which samples inside a segment are recoverable and the wrong one for comparing against a leaderboard. **Controlled injection** plants anomalies of chosen length and amplitude into anomaly-free prefixes and measures the same coverage; it is the only interventional one of the three. Every measurement is identified below; the metric bridge combines the two observational protocols:

Measurement	Unit	Metric	Protocol	Calibration
Leaderboard analysis, §2, App. C.4	file	published VUS-PR	official (their numbers)	none (rank metric)
Width sweep and w^*, §3.3, App. C.7	event	coverage at 1% FPR	internal	tie-safe (App. C.6)
Fusion comparison, §3.4.3, App. C.8	event	coverage slope; VUS-PR by source	internal	tie-safe
Injection, Prop. 2 cells, K-transition, synthetic, eq. (1), §4	injected series	coverage; empirical FPR $\leq 1\%$	controlled injection	tie-safe
Readout experiments, App. C.5	injected series	coverage; empirical FPR $\leq 1\%$	controlled injection	tie-safe
Coverage–VUS bridge, §4	real series	coverage; VUS	observational bridge	tie-safe
Table 4, leakage, worst bucket, channel rung, §5, App. C.1	file	VUS-PR	official	none (rank metric)

The bridge uses one score vector per real series, with coverage and its normal-labelled FPR threshold computed after the learning prefix, and VUS computed on the whole series under the official metric convention. It is an observational association, not an injection result. Coverage uses a strict comparison above the higher empirical 99th percentile of normal scores. This controls the calibration-sample FPR at at most 1%; it neither fixes realised rates to be identical across tied scores nor guarantees an out-of-sample rate. Re-scoring the original 100 files leaves VUS unchanged; after the original exclusion of nine degenerate-source files, the coverage–VUS-PR correlation is 0.748 (file-bootstrap 95% interval $[0.607, 0.847]$; 0.662 with all 100 files).

B.4 BENCHMARK DEVELOPMENT AND EVALUATION HISTORY

The initial fusion rule and width ladder were selected on TSB-AD-U. TSB-AD-M was held out from that selection: the channel-maximum extension was fixed before any multivariate file was scored, giving the first, untouched result 0.3770 (fifth of 31). Table 4 reports 0.4145 from run 3, which includes calibration and a joint-channel rung. There were four sequential runs; each addition was fixed before its own run, but later hypotheses could use knowledge of earlier M results. The chronology is:

Run	What was scored on TSB-AD-M	Fixed before the run	What the result changed
1	The disjunctive detector as selected on TSB-AD-U, with the channel-maximum extension rule (0.3770, 5th) and its pre-declared ablation, channel mean (0.3745)	Extension rule motivated by Corollary 2; no multivariate file had been read	Nothing. First held-out number
2	The same detector with tail calibration, and the gapless and plain-ladder variants (0.3970 calibrated, 2nd; its pre-declared channel-mean ablation 0.3864)	Calibration and its three predictions registered and confirmed on TSB-AD-U first; the M prediction (larger gain) registered before this run	Adopted calibration — a decision the U result had already fixed
3	Run 2 plus the joint-channel rung (0.4145, 2nd)	Rung motivated by Proposition 5; three predictions registered	The rung was kept as registered; two of its three predictions failed and are reported as failures
4	Run 3 with a per-arm Šidák correction	Three predictions registered	Nothing: VUS-PR unchanged to four decimals
—	Worst-bucket audit and runtime measurement on the run-3 scores	—	Text only; no re-scoring

The calibration decision followed its univariate predictions; the joint rung was retained as registered despite failing two predictions. Its registration also named run 2's weakest multivariate duration bucket, so later hypotheses were not independent of M. The registrations are the authors' pre-run script docstrings and run logs, not an external time-stamped registry; the original records are preserved. These distinguish pre-run commitments from post-hoc selection, but do not make run 3 a fresh confirmatory evaluation of the final detector.

B.5 PRE-REGISTRATION

Every registered prescription was motivated by the theory, had its prediction written into the docstring of the script that implements it before that script was run, and the prediction was not edited afterwards; the constituent choices that were selected on TSB-AD-U rather than registered are listed as such in Appendix B.2. Appendix C.12 reports five failures, one partial success and one clear success. The original pre-run predictions, including failures, are preserved in the scripts and will be released on acceptance.

The preserved readout endpoint description and its implementation disagree on which half-maximum crossing to use. Appendix C.5 discloses the discrepancy and a comparison on the saved curves; no historical description has been rewritten to match the implementation.

The same discipline governs the test of Proposition 3 (§3.4.3). Its first implementation used a forecaster that diverged; it was replaced rather than repaired, and the broken version is retained. Its predictions were registered before the run; a first configuration passed all four and the larger one fails one of them under tie-safe calibration, and we report the larger one.

B.6 REPRODUCING THE TABLES AND FIGURES

The main benchmark and experimental tables and every figure are generated by scripts from the result files, including the corrected fusion table, which is generated from its event and series result files; the development-history table of Appendix B.4 and the prescription summary of Appendix C.12 carry values transcribed from the run records rather than read from result files. This is so that a figure cannot drift from its source through repeated copying. Re-running those scripts after any change regenerates the tables and figures together. The experiment and analysis scripts, required result files, metric self-test, and preserved pre-run prediction records supporting the paper's claims will be released on acceptance. These materials will support regenerating the reported tables and figures from archived results and rerunning the experiments with the public TSB-AD release.

Table 5: The ordering of two of our variants reverses across the benchmarks. TSB-AD-U informed the fusion rule and width ladder; TSB-AD-M was not used for that initial selection and was evaluated sequentially afterwards (Appendix B.4). The gapless variant wins on the former and collapses on the latter. Ranks count 40 published U methods or 30 published M methods, plus both variants (42 and 32), one more than Table 4's.

	TSB-AD-U (design decisions leak)		TSB-AD-M (sequential runs; App. B.4)	
Variant	VUS-PR	Rank	VUS-PR	Rank
gapless ladder	0.5449	3	0.3680	6
WIDTHLADDER (reported)	0.5390	4	0.4145	2

Table 6: Six anomaly-duration buckets on TSB-AD-U, six different winners among the 40 published detectors — in the point estimates. $P(\text{win})$ is the winner's probability under a file bootstrap within the bucket (2,000 draws); all six winners are distinct in 12% of draws. The spread of a single method across buckets exceeds the gap between first and tenth place overall.

Mean anomaly length	Files	Best published method	Its VUS-PR	$P(\text{win})$	Runner-up	Gap
1-3	48	TimesFM	0.8377	0.35	Chronos	0.004
4-30	56	TimeRCD-MAFT	0.4310	0.79	CHARM	0.037
31-100	61	TSPulse (FT)	0.5700	0.55	TimeRCD-MAFT	0.013
101-300	86	CHARM	0.6009	0.61	TSPulse (FT)	0.012
301-1000	63	Sub-PCA	0.6147	0.58	TimeRCD-MAFT	0.014
1001+	36	StreamVAE	0.8482	0.60	Sub-PCA	0.008

C SUPPORTING DETAIL

C.1 CHANNEL WIDTH

Proposition 5 (Channel-width bound)**.** *If the trace of an anomaly lies only in the joint distribution of a channel set $\mathcal{C}$, so that every proper marginal of the anomalous law agrees with the normal one, then any statistic reading fewer than $|\mathcal{C}|$ channels jointly has $P^{\mathrm{A}} = P^{\mathrm{N}}$ and detection rate α.*

Proposition 5 extends the argument of Proposition 2 from the time axis to the channel axis: an anomaly is invisible to every view narrower than the one in which it is written. It motivates the multivariate construction of Section 5, in which channels are combined by the same disjunctive rule as widths, and it predicts that a detector whose final statistic depends on fewer than $|\mathcal{C}|$ channels is blind to anomalies carried purely by their joint distribution. A detector that scores channels separately and then reads those scores jointly is not covered: the joint law of the per-channel scores can change even when every marginal does not.

Its empirical status is the weakest of the four propositions and we do not rest any claim of the paper on it. Adding a rung that reads all channels jointly improves TSB-AD-M by 0.0174 ($p = 0.035$), but the predictions we registered before running failed: the gain concentrates in series with 21 to 60 channels and is approximately zero elsewhere, and it does not grow with channel count as predicted. A plausible mechanical explanation — that a nearest-neighbour statistic with a few hundred reference points carries little information in 248 dimensions — is untested. Proposition 5 says a channel-wide view is necessary; it does not say which statistic should read it, and our attempt to choose one was not successful.

C.2 SELECTING ON THE LEAKED BENCHMARK

The ordering of these variants reverses across benchmarks (Table 5). Because TSB-AD-M was evaluated sequentially (Appendix B.4), this is a descriptive comparison rather than a fresh held-out validation.

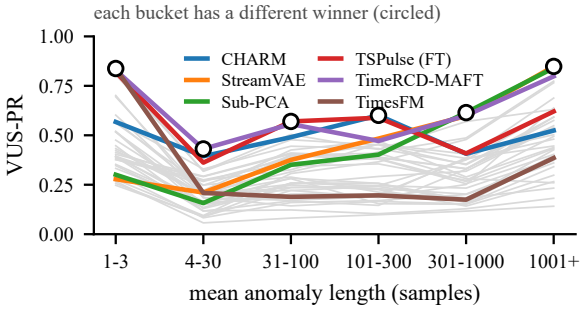

Figure 3: VUS-PR of all 40 published TSB-AD detectors against the mean anomaly length of the files they are scored on. Grey lines are individual methods; coloured lines are the six that win a bucket, and the circle marks each bucket's winner. No method wins twice.

Table 7: Duration slope of every published TSB-AD-U detector with a fixed effect for each file. $\hat{\beta}_m$ is the change in VUS-PR per e-fold of mean anomaly length, as a contrast to the mean slope over methods (the main effect of duration is absorbed by the file); the interval is a 95% percentile bootstrap over files and q is BH-FDR over the 40 methods from a permutation null that shuffles durations only within a source (*: $q < 0.05$). Negative favours short anomalies. Adding the interaction to a model with file and method effects raises within-file R^2 from 0.150 to 0.216 (within-source permutation $p = 0.0002$).

Method	$\hat{\beta}_m$	95% CI	q	Method	$\hat{\beta}_m$	95% CI	q
Chronos*	-0.070	[-0.080, -0.059]	0.004	MatrixProfile	-0.004	[-0.015, +0.006]	0.65
TimesFM*	-0.066	[-0.076, -0.055]	0.004	MMPAD	-0.002	[-0.018, +0.014]	0.73
Lag-Llama*	-0.047	[-0.057, -0.037]	0.02	IForest	-0.001	[-0.012, +0.011]	0.86
LOF	-0.033	[-0.044, -0.022]	0.22	Sub-KNN	+0.003	[-0.009, +0.014]	0.71
TSPulse (ZS)*	-0.033	[-0.045, -0.020]	0.0098	CHARM	+0.003	[-0.011, +0.016]	0.87
SR*	-0.026	[-0.038, -0.013]	0.0067	FITS	+0.005	[-0.002, +0.013]	0.71
Sub-MCD	-0.025	[-0.036, -0.015]	0.076	Donut	+0.006	[-0.001, +0.014]	0.71
TSPulse (FT)	-0.021	[-0.033, -0.010]	0.15	LSTMAD	+0.012	[-0.000, +0.025]	0.6
Sub-LOF	-0.020	[-0.032, -0.009]	0.33	MOMENT (FT)	+0.015	[+0.004, +0.025]	0.71
KMeansAD	-0.020	[-0.034, -0.006]	0.22	Sub-HBOS*	+0.016	[+0.006, +0.026]	0.018
Time-RCD	-0.014	[-0.026, -0.000]	0.66	Sub-IForest	+0.017	[+0.007, +0.026]	0.1
SAND	-0.013	[-0.028, +0.002]	0.6	MOMENT (ZS)	+0.017	[+0.006, +0.028]	0.55
Sub-OCSVM	-0.013	[-0.022, -0.003]	0.55	Series2Graph*	+0.018	[+0.007, +0.030]	0.006
KShapeAD	-0.012	[-0.030, +0.004]	0.6	TranAD	+0.020	[+0.010, +0.030]	0.34
CNN	-0.011	[-0.023, +0.000]	0.22	POLY*	+0.025	[+0.015, +0.035]	0.009
AnomalyTransformer	-0.011	[-0.019, -0.003]	0.79	OmniAnomaly	+0.036	[+0.024, +0.049]	0.15
TimesNet	-0.011	[-0.019, -0.003]	0.097	xLSTMAD*	+0.060	[+0.050, +0.070]	0.009
TimeRCD-MAFT	-0.008	[-0.019, +0.004]	0.64	USAD*	+0.066	[+0.056, +0.077]	0.021
AutoEncoder	-0.007	[-0.019, +0.004]	0.99	Sub-PCA*	+0.075	[+0.064, +0.086]	0.0064
OFA*	-0.005	[-0.013, +0.002]	0.028	StreamVAE*	+0.078	[+0.068, +0.088]	0.0053

C.3 Bucket winners in full

Figure 3 plots every method; this table gives the exact value of each bucket's best published detector.

C.4 The method–duration interaction with the file held fixed

Why the file, and what can be identified. The bucket winner of Appendix C.3 is an argmax over 40 methods, and its obvious confound is source: TSB-AD's 350 files come from 23 source datasets, 80% of the variance in log mean-anomaly-length lies between sources, 39 files belong to sources that occupy a single duration bucket, and only ten sources span three or more buckets (UCR alone spans all six). We therefore fit

$$\text{VUS-PR}_{m,f} \;=\; \delta_f + \alpha_m + \beta_m \log L_f + \varepsilon_{m,f}, \tag{5}$$

with a fixed effect δ_f for every file. Because every method is scored on every file, demeaning within file removes δ_f exactly and with it the main effect of $\log L$, so what is identified is the contrast $\beta_m - \bar{\beta}$: how a method's score moves with duration relative to the other methods on the same file. The contrasts are the same whether the file-level control is a file effect or a source effect —

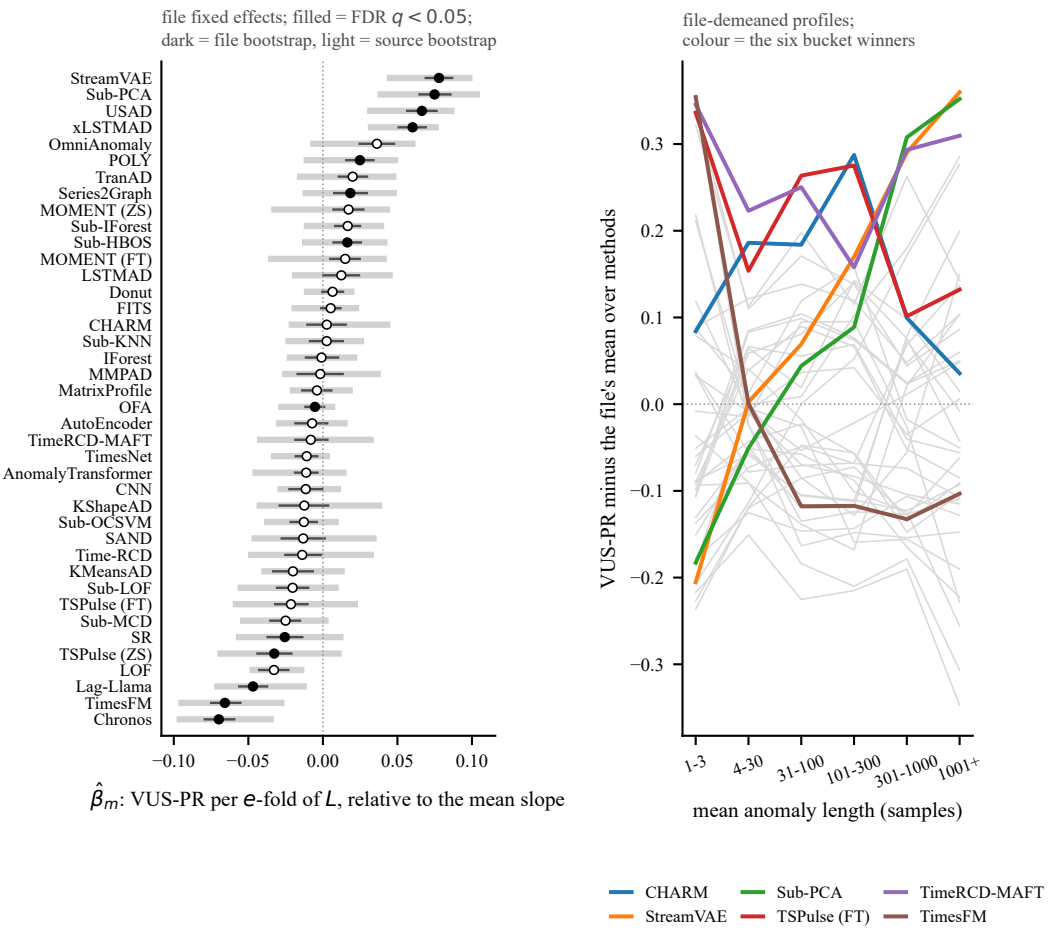

Figure 4: **Left:** the duration slope of every published detector under a fixed effect for each file, sorted; dark intervals are a percentile bootstrap over files, light ones a cluster bootstrap over the 23 sources, and filled markers are significant at FDR 0.05. **Right:** VUS-PR minus the file's mean over methods, per duration bucket, for the six bucket winners (colour) and every other method (grey). The right panel is the residualised profile and is descriptive; no test is run on it.

they come from between-method variation within a file, which neither touches — so nothing below depends on that choice; only the main effect does, and it is discussed last. The file effects alone explain 36% of the variance in VUS-PR.

The interaction. Adding $\beta_m \log L_f$ to a model with file and method effects raises within-file R^2 from 0.150 to 0.216 (partial R^2 0.078). Two permutation nulls are reported. Shuffling durations across all files gives a null 95th percentile of 0.005 for the gain; shuffling them only within each source — which keeps every between-source difference in duration and tests only what within-source variation contributes — gives 0.056. The observed gain of 0.066 exceeds both ($p = 2 \times 10^{-4}$ under each, 5,000 permutations). The second null is the one that answers the source objection, and its high baseline is itself informative: most of the interaction's explanatory power is carried by between-source differences in duration, which observational data cannot attribute to duration rather than to source, and it is the within-source part that is shown to be significant. Dropping any one of the 15 sources with at least eight files leaves the within-source permutation $p \leq 10^{-3}$; without UCR, the source with the widest duration range, the gain is 0.081.

Effect sizes. Table 7 gives every slope and Figure 4 plots them with both bootstraps. Their standard deviation across methods is 0.032 VUS-PR per e-fold of duration and the axis spans 7.5 e-folds, so the steepest slope (+0.078, StreamVAE) implies a relative swing of 0.59 across the axis and the two

Table 8: The context width w of Definition 1, realised in each detector family. "Exact" means the receptive field is fixed by the architecture; "upper bound" means the nominal width contains the effective one, which by Remark 8 still yields a valid bound; "scorer-dependent" means the span is fixed by the scoring rule rather than by the architecture. The convolutional formula is the standard one (Bai et al., 2018). Patching tokenises the input (Nie et al., 2023); readout width is set by the scoring rule.

Family	Quantity playing the role of w	Status
Sliding-window statistic, fixed references	subsequence length	exact
Convolutional / temporal convolutional network (TCN)	receptive field $1 + \sum_\ell (k_\ell - 1)\, d_\ell$ (stride 1)	exact
Transformer, fixed input window	input span used by the scoring rule	upper bound
Recurrent / state-space	effective memory horizon	approximate
Forecasting foundation models	full conditioning + scored forecast/readout span (§3.4.2)	scorer-dependent

extremes (-0.070 for Chronos) a gap of 1.1 between them — linear extrapolations of a bounded score, quoted for scale. Thirteen methods are significant at BH-FDR $q < 0.05$ under the within-source null, six negative and seven positive; the three pre-trained forecasters are the three most negative slopes and remain below zero under the source-cluster bootstrap. The signs agree with the rank correlation an earlier draft reported for 37 of 40 methods. The statistic to read is the spread of slopes, not their mean: with source fixed effects in place of file effects the main effect of duration on the average method is $+0.002$ per e-fold ($[-0.012,\ 0.015]$), so within a source performance does not rise or fall with duration on average — what changes is which method is ahead, which is what the theory predicts.

What Section 2 is for. Section 2 is the phenomenon and the motivation; Section 3 is the mechanism; Section 4's controlled injection is the interventional evidence, setting duration by design jointly with onset and, under a fixed budget, event count. Duration and source are partly entangled in any observational benchmark and no fixed effect separates them where a source has one duration; the interaction shows that source differences do not fully explain the association, but does not identify a causal effect of duration. Expectations for this analysis were written into the script before the models were fitted; all four held.

Robustness, and how fragile the hook is. Three further checks, with expectations registered before running. First, a model with a method-by-source intercept and a slope on the *within-source centred* log-duration — so that β_m is identified from within-source variation only — agrees in sign with the file-effect contrasts for 32 of 40 methods (Spearman $+0.86$); the slopes shrink (standard deviation 0.024 against 0.032), as they should when between-source duration differences are removed. Chronos, TimesFM and Lag-Llama retain negative slopes (-0.058, -0.057, -0.022); Chronos and TimesFM remain the two most negative. Second, fitting the file-effect model inside each of the 14 sources with duration variation, the pooled sign is shared by a majority of sources for 28 of 40 methods. Third, the bucket winners are fragile, which is why they are reported only in Appendix C.3: resampling files within buckets ($2{,}000$ draws) gives winner probabilities of 0.35 (1–3, TimesFM over Chronos by 0.004), 0.79, 0.55, 0.61, 0.58 and 0.60, and all six winners are distinct in 12% of draws (Table 6).

C.5 Scope of the width–architecture correspondence

Section 3.4 identifies the theory's context width with a model's receptive field. We record here what that identification does and does not establish.

It establishes that the propositions apply to learned detectors, because their hypothesis is a property of the score function's support rather than of its form, and that the relevant support is computable exactly for convolutional and sliding-window detectors and bounded above for attention-based ones (Remark 8). One qualification concerns reference sets. A self-join statistic — the matrix profile, or a nearest-neighbour distance against every other window of the same series — reads a query window of the stated width but compares it against the whole series, so its score is not a function of that window alone, and the propositions do not apply to it as stated: reference and query are dependent, and we offer no error bound. For references fixed in advance the propositions apply exactly, provided every other input to the score is fixed in advance too. Our reported detector fixes its references on

the prefix but takes its standardisation and histogram bin ranges from the whole series and reaches past the prefix at two block boundaries (Appendix B.2), so it is not a local test; the strict variant of Appendix C.10 is, in the sense of Proposition 1 (its width-2 term reads amplitude and so lies outside Proposition 2's class), and the table there says for every experiment which statistic, which preprocessing and which reference set were used. Table 8's first row is the fixed-reference, fixed-preprocessing case.

The registered test of Proposition 3, and its outcome. The test is summarised in §3.4.3 and run as follows: inject as in Section 4 with the segment count fixed, score with an exponentially weighted level tracker of half-life τ read out in blocks of R over a context of C, and record L^*, the anomaly length at which coverage at an empirical FPR budget of 1% is highest. Separating τ from C is what makes the design informative: a tracker that simply averages its whole context has an adaptation transient equal to C by construction, so it would measure only itself. Four predictions were registered before running, with thresholds:

- P1 *Readout governs.* Where $\tau \le C/4$, $\mathrm{Spearman}(R, L^*) \ge +0.70$ and the median L^* at the largest R is at least $4\times$ that at $R = 1$. **Not confirmed**: $+0.341$ ($p = 0.056$), factor 1.
- P2 *Conditioning does not.* $|\mathrm{Spearman}(C, L^*)| < 0.30$. **Confirmed**: $+0.079$.
- P3 *The right summary is* $R + \min(\tau, C)$, beating C by at least 0.30 in rank correlation with L^*. **Confirmed**: $+0.490$ against $+0.012$.
- P4 *The transient governs too*: $\mathrm{Spearman}(\tau, L^*) \ge +0.50$ at fixed R. **Confirmed**: $+0.642$.

Two procedural notes. A first version of the experiment used a locally fitted autoregressive forecaster run free for R steps; it diverged, reaching a normal-region residual of 6×10^6 standard deviations at the widest readout, so that the false alarm threshold tracked the divergence rather than the anomaly. Nothing was learned from it and it was replaced rather than repaired. And a first run on a third of the hosts passed P1 and P4; the larger run originally failed both. Re-running its $33{,}048$ cells with tie-safe calibration leaves P1 failed but changes P4 to passed. The original numerical predictions are preserved; we report the corrected larger run.

The failure of P1 has a diagnosable cause, which we give as a post-hoc observation rather than a rescue: L^* sat at the smallest injected length in 32 of 36 cells, censored from below, because the level tracker is strong on short anomalies and the optimum lies at or beneath the grid floor. A statistic less affected by this floor — the long-side half-maximum length — tests the same claim better; we did not substitute it after seeing these results, but registered it as the primary endpoint of the follow-up experiment below.

The same question, registered on pretrained forecasters. The level tracker is a forecaster we control, not one anyone uses. Zhu et al. (2026) swept model size and context length for five families of pretrained time-series models and found their anomaly-detection performance no better than a moving-window variance or a squared difference. They also lengthened the forecasting horizon, scoring a block of h steps by its mean squared error, and found on real anomalies that a longer horizon helps long anomalies and hurts short ones, that the best horizon tracks anomaly length, and proposed a maximum over horizons (their §6.1 and Appendix L). Their h couples what we call the readout's two parts, the forecast horizon and the aggregation width, and their sweep did not vary the context alongside it. We therefore registered a second experiment that varies both widths on the models themselves, with a coverage endpoint at fixed false-alarm rate on controlled injections rather than VUS-PR on real anomalies: Section 4's fixed-count level-shift generator (four segments of amplitude $3\widehat{\sigma}$, lengths $1, 2, 4, 8, 16, 32, 64, 128, 256$), a conditioning window $C \in \{64, \ldots, 1024\}$, a readout $R \in \{1, 4, 16, 64\}$ defined as the number of consecutive samples read into one score — the block is forecast jointly from the context ending at its start and its evidence is averaged — and, in the same runs at the same R, the two one-liner controls of Zhu et al. (2026) and the tracker above. The primary endpoint is the long-side half-maximum duration $L_{1/2}$: the first crossing below half the peak after the first maximum, interpolated linearly in log-duration. A zero-peak curve has no defined frontier; a positive-peak curve without a crossing is right-censored at 256. All reported frontiers use this implementation. The preserved pre-run description instead specifies the last crossing. These definitions differ for nonmonotone curves; the timing and reason for the discrepancy could not be established from the retained records. Recomputing both definitions on

Table 9: Experiment 3: the long-side half-maximum duration $L_{1/2}$ at each readout width R, median over conditioning widths $C \in \{64, \ldots, 1024\}$, computed from coverage averaged over hosts and seeds, with the four registered statistics: H1 requires, over R, a median $\max_C L_{1/2} / \min_C L_{1/2} \leq 2$ *and* $|\mathrm{median}_R \mathrm{Spearman}(C, L_{1/2})| < 0.30$; H2 requires, over C, a median $\mathrm{Spearman}(R, L_{1/2}) \geq 0.70$ *and* a median $L_{1/2}(64)/L_{1/2}(1) \geq 2$. †: right-censored at the largest injected length in at least half the widths. Runs are hosts $\times$ seeds and differ by arm because of cost.

Arm	runs	$R = 1$	4	16	64	H1 ratio	H1 ρ_C	H2 ρ_R	H2 ratio	censored
Chronos-Bolt (small)	11×2	5.2	12.9	35	145	1.31	+0.00	+1.00	27	0%
TimesFM 2.0 (500M)	11×2	5.5	15.9	48	153	1.44	+0.60	+1.00	30	0%
small TCN, trained per host	11×2	$256^\dagger$	$256^\dagger$	$256^\dagger$	$256^\dagger$	1.00	+0.00	+0.00	1	95%
squared difference	11×2	2.0	7.5	31	89	1.00	+0.00	+1.00	44	0%
moving-window variance	11×2	$256^\dagger$	$256^\dagger$	$256^\dagger$	$256^\dagger$	1.98	+0.71	+0.00	1	80%
level tracker ($\tau = 8$)	11×2	23	28	45	108	1.01	+0.71	+1.00	5	0%

the saved coverage gives identical endpoints for all 120 coupled and 192 factorial pooled curves, but different endpoints for $186/2{,}414$ and $274/3{,}442$ defined host–seed curves, respectively (across all arms). Thus the pooled tests are unchanged; paired summaries below are conditional on the implemented first-crossing definition. Three hypotheses were registered before any arm ran: (H1) at fixed R, C moves $L_{1/2}$ by less than a factor of two and with $|\mathrm{Spearman}(C, L_{1/2})| < 0.30$; (H2, primary) at fixed C, $L_{1/2}$ rises with R, Spearman ≥ 0.70 and at least twofold from $R = 1$ to 64; (H3) the one-liners show the same readout dependence and match the learned models' $L_{1/2}$ within a factor of two at matched R.

Implementation settings. Each eligible host has at least 12,000 normal prefix samples: the first 4,000 form the learning prefix and the next 8,000 the injection canvas. The injected amplitude uses the operational scale $\widehat{\sigma} = 1.4826\ \mathrm{median}_{t \in \mathrm{prefix}} |x_t - \mathrm{median}_{s \in \mathrm{prefix}} x_s|$; this robust estimate is not assumed equal to the theoretical noise parameter. The coupled grid uses $C \in \{64, 128, 256, 512, 1024\}$ and $R \in \{1, 4, 16, 64\}$; the host/seed counts and reduced factorial design are given below. Chronos uses `amazon/chronos-bolt-small` and its median quantile forecast. TimesFM uses `google/timesfm-2.0-500m-pytorch` and the point forecast, with configured maximum context 2048 and horizon 64; the actual supplied context is C, not always 2048. Both receive raw-valued inputs, leaving any model-internal normalisation to their implementations. The TCN instead uses learning-prefix median/MAD normalisation, kernel 2, dilations $1, 2, 4, \ldots, C/2$, 16 channels and a 64-step head, trained with Adam for 400 steps (learning rate 0.003, batch 256, seed 0). Forecast blocks start at the canvas origin, advance by the horizon, and use the immediately preceding C observations. Absolute forecast residuals are averaged in non-overlapping readout blocks aligned to that origin and broadcast to their samples; a final partial block averages its available values. In the factorial design, forecasting uses H and residual averaging uses A. Host selection scans sorted filenames, taking up to two eligible hosts per source (one for the reduced TimesFM factorial design), then rejects nonfinite or labelled initial 12,000-sample segments and nonpositive prefix MAD. The coupled runs use seeds $0, 1$; the reduced factorial uses seed 0.

Calibration and replay audit. The corrected runs use the same injected cells and conservative empirical calibration (Appendix C.6); raw evidence and labels are cached. Chronos is replayed with its original inference batches. TimesFM removes masked padding where possible, retaining the conditioning window, weights and batch size; any change in legacy-rule coverage triggers an original-padding replay. Every reported pretrained coverage under the old rule is reproduced before applying the corrected rule; floating-point differences in score medians are logged. The original TCN weights were not saved and its CUDA training was not exactly reproducible. We therefore retrain that control on CPU with the same architecture, prefix, seed and 400 Adam steps, save every checkpoint, and recompute both coverage and adaptation from those weights. This control is a fresh training run, not a calibration-only replay. Numerical predictions written before the original experiments are retained, including failures.

Table 9 gives the outcome. Chronos-Bolt (small) and TimesFM 2.0 (500M) use 11 hosts and 2 seeds with frozen weights. Coverage is averaged over hosts and seeds before taking the frontier; table entries then take the median over C. From $R = 1$ to 64, that frontier rises from 5.2 to 144.5 on Chronos and 5.5 to 153.1 on TimesFM. The registered median of per-context ratios is 27.38 and

Table 10: Experiment 3b: the readout of Experiment 3 factorised into the forecast horizon H (rows) and the aggregation width A (columns). Entries are the long-side half-maximum duration $L_{1/2}$, median over C after averaging coverage over hosts and seeds; the diagonal $H{=}A$ is Experiment 3's R by the same procedure (coverages identical on every common cell; TimesFM here is the reduced subset of 7 hosts, one seed, $C \in \{64, 1024\}$). Bottom block: adaptation time τ_{ad}, the first offset after the onset of a level shift of length 256 at which the step-wise evidence stays below $1.5\times$ its normal median for eight consecutive samples, median over hosts, per C; 256 means it never does.

$H\backslash A$	Chronos-Bolt				TimesFM 2.0			
	1	4	16	64	1	4	16	64
1	5.2	12.5	34.9	104.1	7.7	15.3	34.4	129.8
4	7.3	12.9	32.0	76.8	12.9	19.3	40.4	101.6
16	19.8	27.7	35.3	118.0	28.3	33.8	54.2	106.1
64	58.1	74.4	89.5	144.5	82.3	91.9	124.2	181.6

τ_{ad} at $C = 64, 128, 256, 512, 1024$: Chronos-Bolt 45, 38, 34, 50, 50
TimesFM 2.0 41, –, –, –, 49
per-host TCN 256, 256, 256, 256, 256

29.54, with Spearman $+1.00$ on both: H2 passes. H1's median max/min ratio over C is 1.31 and 1.44, but its rank statistic is 0.00 and $+0.60$, so H1 passes on Chronos and fails on TimesFM. None of these 20 averaged curves per pretrained arm is right-censored; individual host curves can be, as the paired analysis below records.

The small causal TCN, trained per host on its learning prefix alone, has median frontier 256 at every readout. H1 passes and H2 fails, with 95% of its averaged curves right-censored: the grid generally does not reach its half-maximum crossing. This control therefore does not establish readout independence beyond the grid. Its persistent residual evidence is characterised by the adaptation trace below. All three forecasters are frozen at test time; their response to a shifted input, rather than whether weights are updated at test time, is what differs.

Absolute coverage and host-level robustness. The endpoint is relative to each configuration's own peak. In the coupled sweep the median peak falls from 0.73 to 0.24 on Chronos and 0.81 to 0.32 on TimesFM. Coverage integrated over log-duration also falls overall (Spearman with R: -1.00 and -0.80). Define an absolute frontier as the first post-peak crossing below 0.25, undefined if the peak never reaches 0.25. At $R = 64$ it is defined in only $2/5$ and $3/5$ conditioning widths, versus $5/5$ at the other readouts. Although medians over the defined endpoints increase with readout, the original absolute-frontier robustness claim E1 is not confirmed: dropping the undefined settings would compare different subsets.

For the host-level check, seeds are pooled and the frontier is summarised over C separately for each host. A host with an undefined endpoint at any required C is excluded. The $R = 64$ frontier exceeds the $R = 1$ frontier in $8/8$ evaluable Chronos hosts and $9/9$ TimesFM hosts; median ratios are 21.7 and 27.3. Host-bootstrap 95% intervals for the mean log-ratio, exponentiated, are $[10.3, 39.2]$ and $[17.7, 43.4]$. These are conditional on evaluability. E2 originally required an increase in at least 10 of the 11 hosts; neither arm meets that count, so E2 is not confirmed. The denominator is not relaxed after excluding zero-peak curves.

Horizon against aggregation. The coupled sweep jointly forecasts and averages a block of R samples. The registered follow-up factorises the horizon $H \in \{1, 4, 16, 64\}$ and aggregation width $A \in \{1, 4, 16, 64\}$, computing evidence once per H and averaging it at every A. Chronos uses the same design; TimesFM uses the registered reduced design of 7 hosts, one seed and $C \in \{64, 1024\}$. H4 requires median Spearman with A at least 0.70 and at least a twofold increase, on both relative and absolute frontiers. H5 requires horizon alone to extend the frontier by less than aggregation alone. H6 requires median adaptation time at most 64 for the pretrained models and censored at 256 for the TCN, at each tested C.

For H4/H5 the frontier is taken from coverage averaged over hosts and seeds; statistics are then summarised over the registered C, H cells. H4's relative clause passes: median Spearman $+1.00$ on

both models, with aggregation ratios 8.7 and 5.4. Its absolute clause can be evaluated across all A in only $8/20$ Chronos and $4/8$ TimesFM (C, H) cells. The defined subset gives median ratios 4.2 and 3.7, but missing endpoints cannot confirm the full registered clause; H4 is not confirmed. H5 passes at its registered aggregation level: aggregation alone gives 19.4 and 16.8, versus horizon alone 10.9 and 10.6. This comparison of aggregate curves is distinct from the paired reading.

Paired effects and censoring sensitivity. For each (host, seed, C), effects compare the four corners $(H, A) = (1, 1), (1, 64), (64, 1), (64, 64)$. A curve with zero peak has no relative frontier, leaving $81/110$ Chronos cells across 11 hosts and $10/14$ TimesFM cells across 5 hosts. Of the retained cells, 24 and 2 have at least one positive-peak frontier capped at 256. These are conditional, grid-limited summaries, not estimates for every host or of uncensored reach beyond the grid.

Aggregation alone, horizon alone and both together give median factors 11.3, 9.5, 19.8 on Chronos and 13.0, 8.4, 17.0 on TimesFM (Table 2). The mean paired log-difference between aggregation and horizon is $+0.343$ and -0.025; the exponentiated host-bootstrap intervals are $1.41\,[1.01, 2.00]$ and $0.97\,[0.49, 1.45]$. Thus the Chronos interval barely excludes equality, whereas the TimesFM interval does not; the latter has only five retained hosts. Aggregation is ahead in 69% and 50% of cells. The interaction is negative in 86% and 90%, with median multiplicative interaction 0.16 and 0.17: the factors are sub-additive on the log-duration scale. Excluding every cell with a censored corner leaves 57 and 8 cells; aggregation/horizon medians are $11.3/9.5$ and $9.3/6.4$, and interaction medians 0.14 and 0.18. The qualitative extension and sub-additivity remain, while the TimesFM effect size depends on which cells are retained. These sensitivity summaries do not repair the missing absolute endpoints.

Aggregation spreads existing step-wise evidence over a block; it does not create new evidence or establish uniformly improved localisation. The registered absolute-frontier condition is incomplete, so the supported claim is a redistribution of transient evidence with a relative-frontier extension, not improved absolute coverage at every width or a uniquely aggregation-driven mechanism. Horizon changes also contribute.

Adaptation traces. For $H = 1$ and a shift of length 256, the step-wise evidence at each offset is averaged over four segments and divided by its normal-sample median, without smoothing. The proxy τ_{ad} is the first offset with eight consecutive values below 1.5, or 256 if none occur. H6 passes on host–seed medians: 34–50 samples for Chronos and 41 and 49 for TimesFM, versus 256 at every C for the retrained TCN. Hosts remain heterogeneous: Chronos is at most 64 in 59–73% of host–seed cells, and the TCN in 0–14%. This is evidence about the typical residual trace, not equality of normal and anomalous block laws or an estimate of the theorem's total-variation transient.

The diagonal $H = A = R$ has exactly the same corrected coverage on all $3{,}960$ Chronos and 504 TimesFM cells common to the two experiments. The TimesFM diagonal differs between the tables only because the factorial experiment uses its reduced design.

The one-liner hypothesis H3 fails as registered. Squared difference moves from $L_{1/2} = 2.0$ to 88.7, and the tracker from 23.1 to 108.2, as R goes from 1 to 64. Moving-window variance instead has median frontier 256 at every readout, with 80% of averaged curves censored: its window already aggregates over C. The pretrained frontiers are within a factor of two of the pooled one-liner median at only $1/4$ and $2/4$ readouts. A common aggregation width therefore does not establish architecture-independent duration coverage.

The readout knob on the benchmark itself. The interventions above set duration by injection. To see whether the same knob moves a forecaster's profile on the benchmark's own files, we score frozen Chronos-Bolt (context 50) and TimesFM 2.0 (context 96), the contexts of the benchmark's own wrappers (Appendix D; TimesFM is run with its input padded to 128 rather than $2{,}048$ samples, same weights and point forecast), on all 350 TSB-AD-U series, whole series as the benchmark's unsupervised pool does, with the official VUS-PR, and vary only the readout: a one-step absolute residual (A1), that residual averaged over blocks of 16 and 64 samples with the forecasts unchanged (A16, A64), coupled readouts with the horizon equal to the block (H16, H64, reported in the result file), the maximum over A1, A16 and A64 after label-free median/MAD normalisation (maxA), and a max-over-horizons baseline inspired by Zhu et al. (2026), over horizons 1, 16 and 64 (maxH). Our baseline uses block mean absolute residuals and whole-series me-

Table 11: The readout knob on the benchmark itself: official VUS-PR of a frozen forecaster on all 350 TSB-AD-U files, by duration bucket, with the forecasts held fixed for A-only configurations. A1 uses a one-step absolute residual; A16 and A64 average that residual over blocks of 16 and 64 samples; maxA takes the maximum of the three after label-free normalisation; maxH varies the forecasts over horizons 1, 16 and 64, inspired by Zhu et al. (2026), using our absolute-residual scores and normalisation rather than their MSE. "Slope" is the file-fixed-effect duration slope of Section 2 (VUS-PR per e-fold, relative to the 40-method file mean); the published entry is the leaderboard's own score with its own implementation. Rank counts the 40 published methods plus the configuration.

Model	Readout	1–3	4–30	31–100	101–300	301–1000	1001+	Overall (rank)	Slope
Chronos-Bolt	A1	0.817	0.198	0.189	0.192	0.164	0.360	0.290 (26)	-0.065
	A16	0.640	0.405	0.411	0.293	0.268	0.564	0.403 (9)	-0.030
	A64	0.440	0.282	0.518	0.371	0.319	0.614	0.408 (9)	$+0.013$
	maxA	0.845	0.343	0.414	0.274	0.239	0.439	0.399 (10)	-0.064
	maxH	0.815	0.307	0.337	0.288	0.236	0.440	0.378 (15)	-0.057
	published Chronos	0.834	0.185	0.143	0.165	0.153	0.341	0.272 (27)	-0.070
TimesFM 2.0	A1	0.860	0.198	0.196	0.198	0.159	0.360	0.298 (24)	-0.071
	A16	0.614	0.436	0.435	0.298	0.268	0.555	0.408 (9)	-0.029
	A64	0.442	0.282	0.507	0.383	0.313	0.610	0.407 (9)	$+0.012$
	maxA	0.878	0.350	0.411	0.287	0.225	0.431	0.403 (9)	-0.069
	maxH	0.843	0.283	0.367	0.295	0.224	0.429	0.382 (14)	-0.061
	published TimesFM	0.838	0.209	0.189	0.196	0.174	0.385	0.300 (23)	-0.066

dian/MAD normalisation; their original score uses horizon-wise MSE. Forecasts are held fixed only for the A-only configurations. Three predictions were registered before any file was scored (`bench_readout.py`): the A1 arm is short-biased and A64's file-fixed-effect slope exceeds it by at least 0.02 with a file-bootstrap interval excluding zero; A64 raises the 1001+ bucket and lowers the 1–3 bucket; maxA's overall score is at least the better of A1 and A64 minus 0.01. All three hold on both models (Table 11). The A1 arm recovers a similar short-biased profile to the leaderboard entries: its slope is -0.065 (Chronos) and -0.071 (TimesFM) against the published -0.070 and -0.066, and its bucket means lie within 0.05 of theirs, although the Chronos weights differ (Bolt-small here, the tokenised base model there). TimesFM also requires a version distinction: the current public TSB-AD wrapper uses `timesfm-1.0-200m-pytorch`, whereas our intervention uses `timesfm-2.0-500m-pytorch`. The exact historical checkpoint behind the downloaded TimesFM leaderboard scores could not be verified. Published rows are reference profiles, not matched-weight controls; only A1–A64 comparisons within each frozen intervention model isolate aggregation. Widening the aggregation alone to 64 samples moves the slope to $+0.013$ and $+0.012$ (paired differences $+0.078$ $[0.063, 0.091]$ and $+0.083$ $[0.068, 0.099]$), raises the 1001+ bucket from 0.36 to 0.61 on both models, lowers the 1–3 bucket from 0.82 and 0.86 to 0.44, and lifts the overall score from 0.29 and 0.30 (ranks 26 and 24 of 41) to 0.41 (rank 9) on both. The maximum over aggregations keeps the short bucket (0.85, 0.88) but recovers less than half of the long-side gain (0.44, 0.43 at 1001+), as Section 6 says a maximum need not, and the max over horizons behaves the same (0.378 and 0.382 overall); the coupled readouts move the slope as far ($+0.040$, $+0.043$ at H64) but score lower overall (0.317 on both), so on the benchmark aggregation of a one-step residual is the better of the two readouts. What the injections measured as a frontier is, on the benchmark, a duration profile that the readout alone rotates from short-biased to slightly long-biased, at a cost on the shortest bucket.

Finally, the correspondence is weakest for recurrent and state-space models, where the effective memory horizon must be measured rather than read off the architecture. No claim in this paper depends on a particular value of w for such a model.

C.6 Design of the verification experiments

Controlled injection. For each host and seed we reuse random position fractions, signs and the per-sample amplitude $3\widehat{\sigma}$, with $\widehat{\sigma} = 1.4826\times$ learning-prefix MAD. The generator divides the 8,000-sample canvas into n slots and places event i at $i\,\mathrm{slot} + 8 + \lfloor u_i \max(\mathrm{slot} - L - 16, 1) \rfloor$. Thus a shared u_i does not fix the onset as L varies. Fixed count uses $n = 4$ and lets prevalence rise; fixed budget uses $nL = 256$ and changes the number of events. The conventions coincide at $L = 64$. The 5,850

injected cells span 39 hosts, five seeds, two anomaly shapes and the admissible length/convention combinations; each is scored by nine detectors under five metrics.

For each cell and metric m, the comparison is the oracle-envelope difference

$$\Delta_m = \max_{d \in \mathcal{P}} m(d) - \max_{d \in \mathcal{S}} m(d).$$

Here POINT ($\mathcal{P}$) contains `point_delta`, `ewma`, `ours_s2`, `ours_cas`; SUBSEQ ($\mathcal{S}$) contains `mp_discord`, `knn_subseq`, `pca_recon`, `iforest`, `lof`. Selection uses the measured metric in each cell, not a label-free selection rule. These are implementation group names, not pure narrow/wide classes: `ours_cas` multiplies the width-2 score by a width-100 occupancy score. The two `ours` methods use whole-series mean/sd scaling, and the cascade uses whole-series bin ranges; the five SUBSEQ methods use prefix median/MAD input scaling and prefix reference windows or fitted models (`mp_discord` additionally centres and scales each window). The point-delta and EWMA controls scale their evidence by its prefix median/MAD. This comparison does not isolate width or establish the performance of a deployable selector. The real-anomaly width sweep (Appendix C.7) holds the algorithm fixed.

Within each host and seed we correlate Δ_m with L, take the median over the host's five seeds, and apply a two-sided Wilcoxon signed-rank test to the 39 host summaries, with Holm correction over 20 conditions (2 shapes $\times$ 2 conventions $\times$ 5 metrics). Seeds from one host share the underlying series and are not independent hosts. The earlier test over 195 host–seed pairs is superseded by this host-level analysis. Figure 2 pools the two shapes for display only.

How flat the benchmark's long anomalies are. Proposition 2 describes an anomaly whose increments depart from the normal law at K indices (Remark 5). For every TSB-AD-U file we count the anomalous samples whose increment $|x_t - x_{t-1}|$ lies inside the learning prefix's increment range (`analysis_scripts/flatness_long.py`, result file `flatness_long.csv`; the same script also computes the fraction with both value and increment in range, whose 1–3 and 4–30 bucket means are the 38% and 75% of Remark 1). By bucket, the mean over files of the increment-only fraction is 0.41, 0.86, 0.92, 0.92, 0.98 and 0.95 (medians 0.27, 0.92, 0.98, 0.99, 1.00, 1.00). On the 99 files with mean anomaly length above 300 the median is 0.998 with interquartile range $[0.960, 1.000]$, 92% of files exceed 0.9, and against the prefix's 99th percentile instead of its maximum the median is 0.976. Large increment-range excursions are thus rare in these long anomalies. This is a descriptive range-overlap diagnostic: it neither estimates the transition count K of Remark 5 nor verifies its normal-law or window-law TV conditions. Distributions and temporal dependence can differ even when all increments lie within the same range; the statistic also does not establish that the level is flat. The numerical expectation registered in the script, a median above 0.90 on the long side, held; its original interpretation as indistinguishability from the prefix is not supported by this diagnostic.

The 2×2 design for Proposition 2. Proposition 2 carries two hypotheses — a flat anomaly and a shift-invariant statistic — so testing it requires varying both. We cross flat against structured anomalies with shift-invariant against amplitude-reading statistics, giving four cells of which exactly one satisfies both hypotheses. The pre-registered prediction was that the bound holds in that cell; the other three are negative controls. The proposition makes no claim outside its hypotheses, so their failure is not required by it; the three controls show that both hypotheses are load-bearing rather than cosmetic.

Across 59,670 injected cells, restricting the exceedance test to cells with bound below one, the flat/centred arm exceeds the bound in 0.8% of cases, against 64.5, 81.2, 94.1% in the other three arms. These are full reruns with the tie-safe calibration specified below; the additional 29,835 no-injection cells give an exceedance rate of 0.1% on the centred arm. The original non-strict threshold comparison inflated coverage in tied-score cells, so the earlier attribution of all residual violations to substrate non-stationarity is withdrawn. The no-injection control describes residual variation; it does not identify its cause.

The bound on substrates that satisfy its hypotheses exactly. The real prefixes are only approximately stationary, so we also ran the same test on synthetic normal processes: white noise, an AR(1) with coefficient 0.8, and a Gaussian random walk, 20 seeds each, with Section 4's injection (3σ

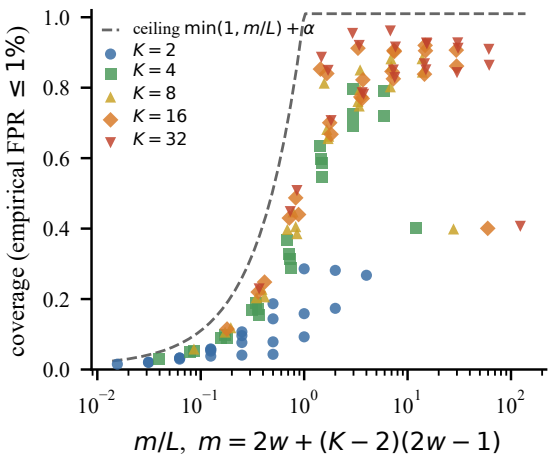

Figure 5: Coverage of piecewise-constant anomalies with K transitions (centred kNN, the class Proposition 2 covers) against the broadcast informative-count upper bound over the duration, m/L with $m = 2w+(K-2)(2w-1)$; one marker per (K, L, w) cell, mean over 39 hosts and 5 seeds; the dashed line is the ceiling $\min(1, m/L) + \alpha$. Empirical normal-sample FPR is at most 1%. $K = 2$ is the flat shift of Section 4.

level shifts in increment units, lengths 8 to 1024) and a shift-invariant statistic (nearest-neighbour distance on increment windows of width 2 to 32) against an amplitude-reading control (the same distance on raw windows). Each window score is assigned to its last sample, without broadcasting. Expectations were registered before the run. On the shift-invariant arm the bound $2w/L + \alpha$ is exceeded in 0.3–0.5% of the 600 informative cells per substrate, by at most 0.004 in coverage. These are empirical exceedances of an expected-coverage bound, not deterministic counterexamples. The amplitude-reading control exceeds the bound in 100.0% of white-noise cells, 81.0% of AR(1) cells and 6.3% of random-walk cells. The registered expectation that this control exceeds the bound in more than half the informative cells on every substrate therefore still fails: a shift of three increment-scales is small against wandering random-walk levels. The shift-invariant control passes the registered small-exceedance and between-substrate similarity checks.

Between two transitions and many: the K-transition test. Remark 5 counts the K indices where anomalous increments depart from the normal law. We inject piecewise-constant anomalies with $K \in \{2, 4, 8, 16, 32\}$ transitions, $K - 1$ nearly equal plateaus of alternating sign, and lengths $L \in \{32, 64, 128, 256\}$. Host, sites, signs and per-sample amplitude are fixed across K; widths 2–64, 39 hosts and 5 seeds give 46,800 cells across centred and raw statistics. The original predictions remain in `prop2_ktrans.py`; all results here use the corrected calibration. The registered single-window ceiling $\min(1, Kw/L) + \alpha$ is exceeded on the centred arm in $0.8, 8.8, 35.4, 34.0, 24.4\%$ of informative cells for the five increasing K values. Thus P1 (at most 5% at every K) fails.

The post-hoc diagnosis concerns broadcasting (next paragraph): an interior transition can influence at most $2w - 1$ anomalous samples, giving the informative-count upper bound $m = 2w + (K - 2)(2w - 1)$. Overlap and stride can make the actual count smaller. Against $\min(1, m/L) + \alpha$, the centred arm's exceedance rates are $0.8, 0.1, 0.0, 0.0, 0.0\%$. None of the 120 cell means exceeds this broadcast bound. The raw arm exceeds the *original* ceiling in $66.0, 65.6, 64.2, 63.9, 63.6\%$ of cells, so registered P2 (more than 50% at every K) passes; its $66.0, 57.6, 51.2, 46.9, 43.8\%$ rates against the broadcast bound are a separate post-hoc result, not that registered prediction. Registered P3 also passes: the correlation with Kw/L is Spearman $+0.843$ and coverage is non-decreasing in K for 23 of the 24 (L, w) pairs. The post-hoc correlation with m/L is $+0.865$ (Figure 5). The count bounds how much of a structured anomaly can retain evidence; it does not determine the attained coverage.

From a window statistic to a sample score. In the real-prefix nearest-neighbour experiments the statistic is computed on subsequences of width w taken at stride $\max(1, \lfloor w/2 \rfloor)$ and broadcast to the samples of its window, overlapping windows resolved by a maximum; a sample score therefore reads up to $2w - 1$ samples, wider than the single window the proposition speaks of. The boundary

count in the proof of Proposition 2 is unchanged by this: a window that contains a boundary extends at most w samples into the anomaly, so a sample deeper than w from both boundaries is scored only by windows lying wholly inside the flat shift, whose centred contents — and hence the maximum over them — have the normal law; at most $2w$ samples escape, as in the proof, and the bound holds for the broadcast score with the same constant. For a piecewise-constant anomaly the constant does change: an interior transition is anomalous on both sides, so the windows containing it reach at most $2w-1$ samples, giving the upper bound $2w+(K-2)(2w-1)$ used in the K-transition test. Overlap and stride can reduce the actual count. The synthetic-substrate check instead assigns each increment-window score to its last sample, as specified above. For the five coverage experiments on nearest-neighbour scores — flat/structured, synthetic, K-transition, the amplitude sweep behind (1) and the real-anomaly width sweep of Appendix C.7 — alarms use a strict inequality above the higher empirical $(1-\alpha)$ quantile of the normal-labelled scores. This guarantees an empirical normal-sample FPR of at most $\alpha = 0.01$, including ties; the CSVs record the realised FPR and threshold-tie fraction. The original non-strict comparison could alarm on every normal sample when the threshold was zero, and, because the broadcast score is piecewise constant, it alarmed on tied normal samples at wide widths (legacy realised FPR median 1.3% at $w = 128$ and 2.3% at $w = 512$ on the width sweep, 3.2% at $w = 256$ on the flat-shift cells), which inflated coverage there by up to 0.1; every result above uses corrected full reruns, and the legacy values are kept beside them in the CSVs. Calibration uses the same injected series, including normal-labelled samples whose windows straddle a boundary. Thus finite-sample calibration and boundary contamination both distinguish this empirical FPR from the normal-law FPR in Proposition 2; the empirical constraint is not a guarantee on unseen normal data. The readout experiments, fusion comparison and coverage–VUS bridge also use this tie-safe rule after rerunning their original designs. Block-mean readouts repeat one value across samples and therefore create ties even from continuous residuals. The reruns retain raw evidence or scores, realised FPRs and legacy-rule coverages for audit; the threshold-free VUS evaluations are unchanged (Appendix B).

Fitting (1)**.** The fit uses the *raw-window* arm of the amplitude sweep, with prefix median/MAD scaling, mean Euclidean distance to three nearest prefix windows, and the broadcast/maximum rule above. The centred arm is used separately for the scale-optimum check in Appendix C.7. A ratio-only fit, $h^*/\sigma \propto (w/L)^{0.439}$, has $R^2 = 0.47$; freeing the width and duration exponents gives $(0.69, -0.33)$ and $R^2 = 0.92$. The correlation of $\log w$ with $\log L$ over the 12 cells is 0.71 (variance inflation 2.0), moderate rather than severe, so the improvement is not a collinearity artefact.

Uncertainty is assessed by resampling the 39 base series, preserving their shared cells and seeds. The reported percentile width interval $[0.46, 1.00]$ uses 10^4 refits and admits $\frac{1}{2}$; cells sharing a host are not treated as independent replicates. The duration interval is $[-0.44, -0.19]$, which excludes $-\frac{1}{2}$. The right tail reflects between-series variation: skewness is $+0.73$, 2.6% of refits exceed 1.0, and those refits retain the same 12 cells and design condition number. The upper endpoint moves from 0.92 at 200 refits to 1.00 at 2,000 and 1.00 at 10^4; at each count, $\frac{1}{2}$ is admitted and $\frac{1}{4}$ excluded.

Which cells were fitted, and a censoring-aware refit. The threshold amplitude is the smallest swept amplitude at which mean coverage over hosts and seeds reaches one half, log-interpolated between grid points. Of the 35 (w, L) cells with $w > L$ it is defined in 12; the other 23 never reach half coverage at the largest amplitude swept, 16σ. The excluded cells are not the large-w/L ones alone: they are every cell at $w \in \{64, 128, 256\}$ and three of the five at $w = 32$ (those with $L \leq 4$), so the fitted relation spans widths 2 to 32 with two cells at 32; a complete-case point prediction at wider windows is an extrapolation. Under the original calibration 16 cells were defined and the fitted exponents were $(0.72, -0.45)$; the tie-safe calibration removes four wide-window cells whose half coverage was reached only with tied alarms. Figure 6 shows every coverage-against-amplitude curve, so the reader can see what was and was not reached. Two checks follow. Leaving one width out at a time and predicting its cells, the power law beats the best logarithmic-width form at all five widths (RMSE in $\log h^*$ from 0.04 to 0.36 against 0.11 to 0.66). And treating the 23 cells as right-censored at 16σ and refitting by maximum likelihood with Gaussian residuals in $\log h^*$ moves the exponents to $w^{1.13}L^{-0.38}$, with series-level intervals $[0.87, 1.60]$ on the width and $[-0.43, -0.34]$ on the duration (1,000 refits), against the complete-case $[0.47, 0.96]$ and $[-0.44, -0.20]$ at the same refit count. Under that censored model the width dependence is estimated stronger than the complete case gives; the direction is what dropping the widest cells would produce, but the model adds assumptions the data do not test — that every censored threshold is finite rather than a structural ceiling of the

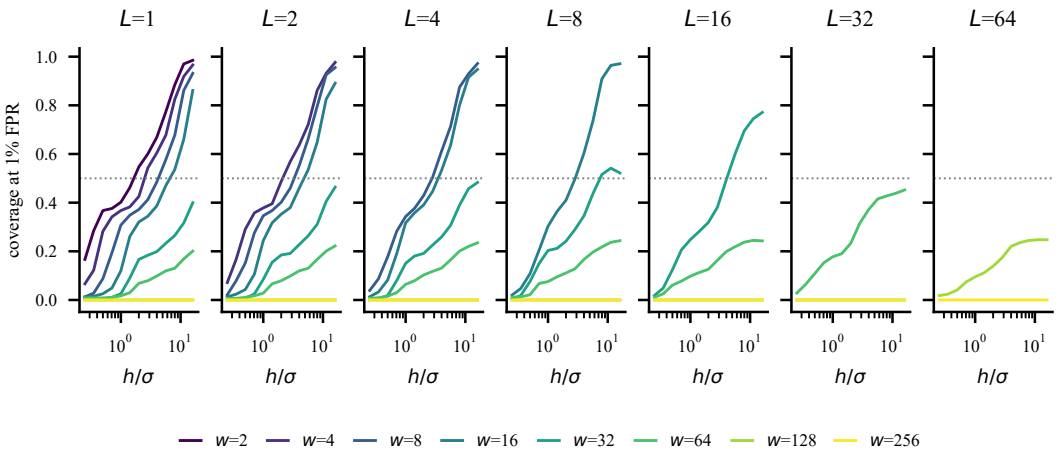

Figure 6: Mean coverage at a 1% false alarm rate against injected amplitude, one panel per anomaly length and one curve per width $w > L$, coloured from dark ($w = 2$) to light ($w = 256$) with the same colour for the same width in every panel. The dotted line is the half-coverage level that defines the threshold amplitude; curves that never cross it are the 23 cells excluded from the complete-case fit of (1).

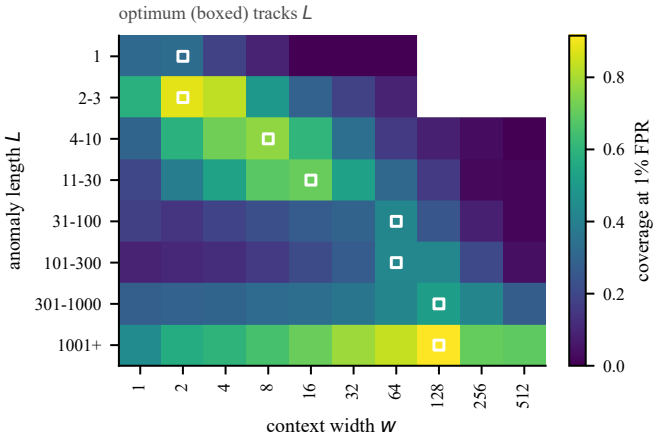

Figure 7: Coverage at a 1% false alarm rate over anomaly length and context width, with the algorithm held fixed and only the width swept. The boxed cell in each row is the best eligible width in that length bucket; cells with fewer than half the row's maximum valid-event count are masked.

statistic, Gaussian residuals of common variance in $\log h^*$, and exchangeable cells although several share a host — so Section 4 reports both fits with their own intervals. The two point estimates are not endpoints of an uncertainty interval or bounds on a universal exponent. The two fits also see different data: the complete-case fit uses $w \in [2, 32]$ with two cells at 32; the censored fit adds the bounds at $w = 32$, 64, 128 and 256.

C.7 Sweeping width with the algorithm held fixed

To test Remark 2's scale prediction, we vary width and nothing else: a single algorithm (nearest neighbours on subsequences, scored as in Appendix C.6) with w swept from 1 to 512.

For each duration bucket, a width is eligible if its nonmissing coverage count is at least 50% of the largest count over widths in that bucket. Means exclude missing scores (width/prefix and width/test-length guards or too few reference/query windows can make a width unavailable), and ineligible cells are masked before selecting the maximum. Thus widths can use different event subsets; w^*

Table 12: Valid event counts in the real-anomaly width sweep. A dagger marks a width excluded from the bucket's optimum because its count is below 50% of the row maximum. Missing coverage values are excluded from each mean; the remaining event sets can differ by width. Figure 1(b) pools these into five coarser buckets and applies the same rule.

L bucket / w	1	2	4	8	16	32	64	128	256	512
1	867	867	867	867	867	867	867	$4^\dagger$	$4^\dagger$	$2^\dagger$
2–3	303	303	303	303	303	303	303	$122^\dagger$	$119^\dagger$	$105^\dagger$
4–10	286	286	286	286	286	286	286	260	192	161
11–30	330	330	330	330	330	330	330	289	266	224
31–100	306	306	306	306	306	306	306	214	209	178
101–300	604	604	604	604	604	604	604	373	354	327
301–1000	155	155	155	155	155	155	155	129	116	103
1001+	71	71	71	71	71	71	71	65	56	56

Table 13: The duration-optimal width on the controlled injections (2.83σ level shifts, the amplitude-grid point nearest the 3σ of §4; 39 hosts; widths 1 to 256), for the raw-window nearest-neighbour statistic of the real-anomaly sweep and for its centred, shift-invariant version, which lies inside the class of Proposition 2. w^* is the width of highest mean coverage at a 1% false alarm rate; the bracket is the 95% range of w^* under a host bootstrap (2,000 draws).

	raw-window kNN		centred kNN (in class)	
L	w^* [95%]	coverage	w^* [95%]	coverage
1	2 [2, 2]	0.60	2 [2, 2]	0.81
2	2 [2, 2]	0.59	2 [2, 2]	0.73
4	4 [4, 4]	0.55	4 [4, 4]	0.64
8	8 [4, 16]	0.50	8 [8, 16]	0.53
16	16 [4, 16]	0.48	32 [16, 32]	0.43
32	32 [32, 32]	0.55	16 [16, 64]	0.27
64	64 [32, 64]	0.60	64 [64, 64]	0.18

means best among eligible widths, not a maximum on a common cohort. Table 12 reports the counts. Figure 1(c) uses display positions $1, 2.4, 6.3, 18, 55, 175, 550, 1800$, not sample means; 1800 places the open-ended $1001+$ bucket. Ratios below use the actual bucket means.

The best eligible width is non-decreasing across the six duration buckets up to length 300 — $2, 2, 8, 16, 64, 64$ — with w^* between 0.7 and 1.1 times the bucket's mean anomaly length for four of the five buckets above length 1 (bucket means 2.7, 7.6, 17.6 and 72) and 0.4 for the 101–300 bucket (mean 152), where widths 64 and 128 are within 0.003 of each other (0.426 against 0.423; at length 1 the optimum is 2, the narrowest width that reads an increment). These are the tie-safe values of Appendix C.6; under the original calibration the 101–300 optimum was 128, and the widest columns lose the most from the correction (at $w = 512$ that bucket's mean coverage is 0.04 against 0.24 before). The two longest buckets (301–1000 and $1001+$) peak at 128 although widths 256 and 512 are admissible for most of their events, so w^* stops tracking L there. Two things are consistent with the plateau and we do not separate them: this statistic reads amplitude, so on a long shift it is not boundary-blind and gains nothing from a window already inside the segment; and at 256 and 512 the prefix supplies fewer reference subsequences (evaluable events fall from 155 to 103 and from 71 to 56 in the two buckets), which lowers the mean coverage at those widths (0.51 at 128 against 0.42 and 0.27; 0.92 against 0.70 and 0.69). The duration slope of performance increases monotonically in w and changes sign near $w \approx 16$ to 32: narrow statistics degrade with duration and wide ones improve with it; on this amplitude-reading statistic the crossover is consistent with complementary duration profiles, but does not test Proposition 2, whose shift-invariance assumption it lacks.

At its matched width, coverage in every duration bucket lies between 0.33 and 0.92. In this sweep no duration is out of reach of some scale.

The same sweep with a shift-invariant statistic. The raw statistic above reads amplitude; its duration-tracking optimum does not establish that a shift-invariant statistic has the same behaviour. The injection experiment of Section 4 scores every cell with both that statistic and its centred version — each window minus its own mean, hence shift-invariant and inside the class — on the same

Table 14: Fusion rule swapped, everything else held fixed: same data, same constituent statistics, each fused score re-thresholded at the same empirical false-alarm budget of 1% using a tie-safe rule. Rules act on real-valued scores; the product is not the alarm conjunction of Corollary 2. Slope is the change in coverage per doubling of anomaly duration, with series fixed effects over $1{,}569$ events in 219 series; closer to zero is flatter. Accuracy is VUS-PR averaged over data sources. *These are internal-protocol numbers and are not comparable with the official leaderboard figures of Table 4.*

Combination	Rule	Duration slope	t	Accuracy
width-2 statistic alone	—	-0.0598	-15.5	0.3708
width-100 statistic alone	—	$+0.0299$	$+4.6$	0.4424
both, product	product	-0.0525	-10.4	0.4058
both, sum	sum	-0.0290	-4.2	0.4961
both, maximum	max	$+0.0099$	$+1.5$	0.4780
width-2 and averaged ladder, sum	sum	-0.0298	-3.7	0.5652
width-2 and averaged ladder, maximum	max	-0.0167	-2.1	0.5735

injected series, so the in-class sweep costs no new run. Table 13 reads off, at 2.83σ (the point of that experiment's amplitude grid nearest the 3σ of Section 4) and over widths 1 to 256, the width of highest coverage for each injected length. For the in-class statistic w^* rises with L ($2, 2, 4, 8, 32, 16, 64$ for $L = 1, \ldots, 64$; Spearman $+0.95$, median $w^*/L = 1$), stays within a factor of two of L at every length, and does so in the majority of host-bootstrap resamples at every length; the one non-monotone step, 32 at $L = 16$ against 16 at $L = 32$, has bootstrap ranges $[16, 32]$ and $[16, 64]$. At 4σ the ratio w^*/L is 1 or 2 at every length. Matching the width does not, however, remove the dependence on duration: the centred statistic's coverage at w^* falls from 0.81 at $L = 1$ to 0.18 at $L = 64$, so what the matched width recovers is the location of the optimum, not a duration-independent detectability. The amplitude-reading statistic tracks in the same way. The duration-tracking optimum thus also appears in a shift-invariant statistic. This is direct evidence for the scale prediction in Remark 2; it does not transfer the raw arm's fitted exponents to the centred arm.

C.8 Fusion rules compared

The table substitutes one fusion rule for another with the data, the constituent statistics and the false-alarm budget held fixed, over $1{,}569$ anomaly events in 219 real series. Comparison is within series, so dataset identity is controlled without synthetic injection.

Under the corrected calibration, the width-100 statistic has a positive within-series duration slope ($+0.0299$, $t = +4.6$); the product reverses it to -0.0525, while the sum gives -0.0290 and the maximum $+0.0099$. The product's source-averaged VUS-PR is 0.4058, below the wide statistic's 0.4424, whereas the sum and maximum reach 0.4961 and 0.4780. Thus the product loses both long-side coverage and aggregate accuracy here; these score comparisons are empirical, not the alarm-conjunction theorem. Each fused score uses a strict threshold above the higher empirical 99th percentile, ensuring normal-sample FPR at most 1%; realised rates can differ because of ties. The original non-strict rule could alarm on every normal sample for a tied score and gave the wide statistic an apparently flat slope of -0.0104. Corrected event coverages and their legacy counterparts are retained in the result file; VUS-PR is unchanged.

C.9 Component ablation under the official protocol

Table 15 takes the reported detector apart under the official TSB-AD-U protocol, one combination rule or arm at a time with the statistics, prefix references and calibration held fixed, and gives the file-paired difference to the reported score with a bootstrap interval over the 350 files. These ablations were run on TSB-AD-U, which informed the design; no corresponding ablation was run on TSB-AD-M. This restriction does not make the final M evaluation a fresh held-out test: its sequential run history is reported in Appendix B.4. Four expectations were registered before the run; three held and one failed. Every single arm is worse than the combination — the width-2 term alone by 0.17, one wide scale by 0.11, the averaged wide ladder by 0.007 with an interval of $[-0.015, -0.001]$ — so the width-2 term adds little on average and does not hurt. Calibration matters: the same combination without it loses 0.010, with an interval excluding zero. The soft-score product loses 0.081, and a bare maximum over all six scales, with the wide arms not averaged, loses 0.096 — the largest loss among fusion changes retaining all six terms, consistent with a multiplicity cost of a maximum

Table 15: Component ablation under the official TSB-AD-U protocol: same statistics, prefix references and calibration, only the combination changed. s_2 is the calibrated width-2 term, h_W the calibrated width-W term and $\overline{h} = \frac{1}{5}\sum_W h_W$; every rule acts on these real-valued scores. Difference is the file-paired mean change in VUS-PR from the reported combination with a 95% percentile bootstrap over the 350 files; rank is among the 40 published methods plus the variant. TSB-AD-U is the benchmark our design leaks into and TSB-AD-M is not scored, so this is a diagnostic, not a selection.

Combination	VUS-PR	Rank	Worst bucket	Δ	95% CI
$\max(s_2,\ \overline{h})$ — reported	0.5390	3	0.367 (4-30)	—	—
$s_2 + \overline{h}$ (ranks as the two-arm average)	0.5353	3	0.353 (4-30)	-0.0037	$[-0.0135,\ +0.0060]$
mean of all six terms	0.5371	3	0.351 (4-30)	-0.0019	$[-0.0101,\ +0.0068]$
$\overline{h}$ alone (averaged wide ladder)	0.5316	3	0.369 (4-30)	-0.0073	$[-0.0151,\ -0.0005]$
reported, no tail calibration	0.5295	3	0.382 (4-30)	-0.0095	$[-0.0183,\ -0.0012]$
$s_2 \cdot \overline{h}$ (soft-score product)	0.4577	6	0.294 (4-30)	-0.0813	$[-0.1094,\ -0.0527]$
max over all six terms (no averaging)	0.4431	7	0.251 (4-30)	-0.0959	$[-0.1164,\ -0.0759]$
h_{100} alone	0.4333	8	0.230 (4-30)	-0.1057	$[-0.1261,\ -0.0860]$
plain kNN ladder, max over widths	0.4349	8	0.288 (4-30)	-0.1040	$[-0.1407,\ -0.0684]$
s_2 alone (width-2 term)	0.3739	15	0.219 (4-30)	-0.1651	$[-0.2022,\ -0.1287]$

Table 16: Two controls against the reported detector, which standardises and sets its histogram bin ranges on the whole series and reaches past the prefix at two block boundaries (Appendix B.2). *Preprocessing-only*: standardisation and bin ranges fixed on the prefix, samples outside the range clipped to the edge bin, boundaries as reported. *Strict*: additionally, only blocks wholly inside the prefix serve as references and as the calibration sample, a width with fewer than three such blocks being dropped for that file; this score is a function of its window and the prefix alone (truncation-invariant, E5). Official TSB-AD-U protocol; difference is whole-series minus control, file-paired, 95% percentile bootstrap over 350 files; rank is the control's among the 40 published methods plus it.

Component	Reported	Control	Difference	95% CI	Rank
width-2 term s_2 alone (either control)	0.3739	0.3739	-0.0000	$[-0.0000,\ +0.0000]$	15
$\overline{h}$ alone, preprocessing-only control	0.5316	0.4153	$+0.1164$	$[+0.0873,\ +0.1472]$	9
$\max(s_2, \overline{h})$, preprocessing-only control	0.5390	0.4502	$+0.0888$	$[+0.0595,\ +0.1196]$	7
$\overline{h}$ alone, strict prefix-local control	0.5316	0.4249	$+0.1067$	$[+0.0768,\ +0.1381]$	8
$\max(s_2, \overline{h})$, strict prefix-local control	0.5390	0.4763	$+0.0627$	$[+0.0317,\ +0.0941]$	6

over heterogeneous arms, though we have not isolated that mechanism from the arms' dependence or their false-positive rates; it is the reason the wide arms were averaged before the disjunction, a choice made on TSB-AD-U (Appendix B.2). The registered expectation that the reported maximum would outperform the mean of all six terms was not supported. The comparison with the two-arm average was also inconclusive. The sum $s_2 + \overline{h}$ ranks every file exactly as the two-arm average $(s_2 + \overline{h})/2$ does, and it loses 0.004 with an interval $[-0.014, +0.006]$; the mean of all six terms, which weights s_2 at one sixth rather than one half, loses 0.002 ($[-0.010, +0.007]$). We did not detect a difference between the maximum and either average on whole-benchmark VUS-PR, which is not a demonstration that they are equal. On the official protocol, then, the product loses and an average does not measurably, and Section 3.4.3 says so; the product here is a rule on real-valued scores, not the alarm conjunction of Corollary 2.

C.10 Sensitivity to preprocessing and prefix-local construction

The reported detector is not a local test in the sense of Proposition 1, for three reasons, all label-free and all documented in Appendix B.2: its standardisation and histogram bin ranges are computed on the whole series; when the prefix holds fewer than three complete blocks of a width, that width's reference set reaches past the prefix (111 of 350 TSB-AD-U files, 161 file–width pairs, at $W = 200$ or 400); and the block containing the last prefix sample straddles the prefix end, so a score computed partly from samples after it enters the calibration sample (334 files, 1,257 pairs). The theorems say what a local test cannot do and nothing about what offline preprocessing may add, so we measured the difference with two controls, expectations registered before each ran (Table 16). The *preprocessing-only* control fixes the standardisation and bin ranges on the prefix, clipping samples

outside the range to the edge bin, and leaves the two boundary behaviours in place; the *strict* control additionally uses, for every width, only the blocks lying wholly inside the prefix as references and as the calibration sample, dropping a width with fewer than three such blocks for that file (all five widths survive on 239 files, four on 31, three on 80). Precisely: blocks are cut from the start of the series, so their phase is fixed; the score of block k reads samples $kW-1$ to $(k+1)W-1$ (its first increment uses the last sample of the previous block) and the prefix blocks $0,\dots,\lfloor p/W\rfloor-1$, and nothing else; and the per-channel score averages the calibrated wide terms of the widths kept for that file, $\frac{1}{|\mathcal{W}|}\sum_{W\in\mathcal{W}}\mathrm{cal}(s_t^{(W)})$, in place of the reported $\frac{1}{5}$; block 0 starts at sample 1, so no increment reaches before the series. The strict score is, by this definition, a function of its window and the prefix alone, and the following check confirms the implementation: on every fifth file with room for it (62 files) the strict score on the first m samples, m the largest multiple of 400 — a multiple of every width, so that every block is complete on both sides — with $p+400\le m\le n-400$, is identical to the last bit whether computed from the whole series or from the series truncated at m, while the reported preprocessing differs under the same truncation on 94% of them (the registered E5 and E6). This truncation check is an implementation check of the stated dependence, not a general proof of finite-window locality.

The results are the price of this control on this benchmark — it changes the clipping, the references, the calibration sample and, on 111 files, the set of widths at once, so it is the cost of one local construction and not of locality in general. The width-2 term moves by at most 2×10^{-3} on a file and 10^{-4} in the mean under either control; an affine change of the standardisation rescales its distances by a common factor and the rank-based calibration is invariant to that, but the tail continuation divides by a MAD plus a fixed constant, so we claim the smallness and not exact invariance (the registered threshold of 10^{-6} was set too tight and that expectation is recorded as failed). The histogram ladder moves a great deal: the preprocessing-only control costs it 0.116 VUS-PR ($[0.087, 0.147]$) and the strict control 0.107 ($[0.077, 0.138]$); the full detector loses 0.089 ($[0.060, 0.120]$, from 0.539 to 0.450, seventh of 41) under the preprocessing-only control and 0.063 ($[0.032, 0.094]$, 0.476, sixth) under the strict one — the strict control's cleaner boundaries recover part of the loss. On 188 of 350 files the preprocessing-only change exceeds 0.05 in either direction, and it is larger where more post-prefix samples fall outside the prefix range (Spearman $+0.21$, $p<10^{-4}$), the registered mechanism. These controls measure sensitivity to preprocessing and reference/calibration construction. They do not identify an independent benefit from crossing the prefix boundary: the strict construction recovers some of the loss relative to the preprocessing-only control, and also changes the available widths. The reported offline detector remains the benchmark construction; the local control scores 0.476 and sixth. The strict control's width-2 term reads amplitude and so lies outside Proposition 2's class; it is a local test in the sense of Proposition 1 and no more.

Because the paper's experiments use several statistics under several preprocessings, the following table says which is which, so that the theorems' hypotheses can be checked against each:

Table 17: How robust the official-protocol standing is. **Top:** file-paired difference in VUS-PR, ours minus each of the five best published methods, with a 95% percentile bootstrap over files, the share of files on which ours is at least as good, and the share of source datasets whose mean favours ours. **Bottom:** rank of ours by weakest duration bucket under the official bucket boundaries $\{3, 30, 100, 300, 1000\}$, the boundaries scaled by 0.7 and 1.4, the buckets with at least 20 files, and the median rank with its 95% range and the probability of first place under a source-cluster bootstrap and a file bootstrap (2,000 draws each).

Benchmark	Published method (rank)	Their mean	Ours − theirs	95% CI	Files ours ≥	Sources
TSB-AD-U	TimeRCD-MAFT (1)	0.5856	−0.0466	[−0.0833, −0.0105]	43%	43% of 23
	TSPulse (FT) (2)	0.5516	−0.0126	[−0.0510, +0.0239]	49%	52% of 23
	Time-RCD (3)	0.5224	+0.0166	[−0.0270, +0.0593]	50%	65% of 23
	CHARM (4)	0.5011	+0.0379	[−0.0063, +0.0815]	53%	52% of 23
	TSPulse (ZS) (5)	0.4801	+0.0588	[+0.0165, +0.0994]	59%	61% of 23
TSB-AD-M	StreamVAE (1)	0.4353	−0.0209	[−0.0563, +0.0129]	49%	53% of 17
	xLSTMAD (2)	0.3923	+0.0222	[−0.0102, +0.0540]	56%	41% of 17
	CHARM (3)	0.3914	+0.0231	[−0.0220, +0.0688]	53%	53% of 17
	TSPulse (FT) (4)	0.3866	+0.0279	[−0.0114, +0.0674]	69%	71% of 17
	TSPulse (ZS) (5)	0.3610	+0.0535	[+0.0144, +0.0927]	71%	76% of 17

Worst-bucket rank of ours	TSB-AD-U (of 41)	TSB-AD-M (of 31)
official boundaries	3 (0.367)	1 (0.362)
boundaries ×0.7	2 (0.385)	2 (0.270)
boundaries ×1.4	3 (0.375)	2 (0.317)
buckets with ≥ 20 files	3 (0.367)	1 (0.375)
source bootstrap: median [95%], P(first)	3 [1, 14], 0.12	1 [1, 13], 0.52
file bootstrap: median [95%], P(first)	3 [1, 7], 0.07	1 [1, 3], 0.53

Experiment	Statistic	Standardisation	References	Local
Official protocol (Tables 4, 5, 15); fusion (Table 14)	width-2 kNN + occupancy-histogram ladder	U: whole-series mean/sd; M: prefix mean/sd; bin ranges whole-series	prefix (self-inclusive); ladder reaches past the prefix at two boundaries	no
Preprocessing-only control (Table 16)	same	prefix mean/sd; bin ranges prefix, clipped	as above	no (boundaries)
Strict control (Table 16)	same, widths with < 3 prefix blocks dropped	prefix mean/sd; bin ranges prefix, clipped	complete prefix blocks only	yes (E5)
Width sweep, real anomalies (Fig. 7, App. C.7)	raw-window kNN at width w (amplitude-reading; outside Prop. 2's class)	prefix median/MAD	prefix subsequences	yes
Width sweep, injections (Table 13)	centred-window kNN (in class) and raw-window kNN	prefix median/MAD	prefix subsequences	yes
Prop. 2 cells (§4)	centred-window kNN (shift-invariant); raw-window kNN control	prefix median/MAD	prefix subsequences	yes
Equation (1) (§4)	raw-window kNN (primary fitted arm); centred arm for width-optimum check	prefix median/MAD	prefix subsequences	yes
Controlled injection (Fig. 2, right)	oracle maxima over POINT and SUBSEQ sets (App. C.6)	method-dependent; two methods use whole-series statistics	prefix references/models; cascade bins use whole series	no (whole-series transforms)
Synthetic substrates (App. C.6)	increment-window kNN; raw-window control	prefix median/MAD	prefix	yes
Readout experiments (Tables 9, 10)	forecaster residual, block-mean aggregation	pretrained: raw input; TCN: prefix median/MAD	none (frozen model; TCN fitted on prefix)	yes

For readout experiments, finite support includes the conditioning and forecast/readout intervals, with the fitted model held fixed; it is not limited to the conditioning width C.

C.11 How robust the official standing is

A rank says which side of a competitor a mean fell on, not by how much or how reliably. Table 17 supplies both, from the published per-file numbers and ours, with expectations registered before the computation. On TSB-AD-U the detector is behind the first-placed published method by 0.047 with

Table 18: Every theory-motivated design prescription, its prediction as registered in the implementing script before the run, and its outcome. Predictions were not edited after the fact.

Prescription	Registered prediction	Outcome	Note
Multi-scale statistics	targeted bucket improves	fail	targeted bucket responded least
Gapless width ladder	4–30 improves	fail	targeted bucket degraded
Calibrated disjunction	3 registered predictions	**pass**	3/3, on both benchmarks
Channel-wide rung	gain grows with channel count	partial	$+0.0174$ (p=0.035) but predictions failed
Relevance weighting	relevance ranks arms by usefulness	fail	confounded with window width ($\rho = -0.927$)
Šidák arm correction	suppressed arm recovers, score rises	fail	suppression fixed, score unchanged to 4 d.p.
Width-decoupled relevance	positive in every bucket	fail	confound sign inverted ($\rho = -0.948$ vs. width)

an interval excluding zero; its differences from the second (-0.013, $[-0.051, +0.024]$) and third are not statistically distinguishable from zero, which is not a demonstration of equivalence, since the intervals admit differences of either sign of the same size as the gap to the first; and it is ahead of the fifth with an interval excluding zero. Each interval is a file-percentile bootstrap for that one comparison, with no adjustment across comparisons. It is at least as good as the second-placed method on 49% of files and in 12 of 23 sources. On TSB-AD-M the gap to the first-placed method, -0.021, has an interval that includes zero ($[-0.056, +0.013]$), and the detector is at least as good as it on half the files and in 9 of 17 sources. The overall ranks are therefore a competitive standing and not a separation from the top: two registered expectations to that effect held.

The worst-bucket headline is robust in one respect and not in another, and we report both. Under the official boundaries the detector is third of 41 on TSB-AD-U and first of 31 on TSB-AD-M, with or without the buckets under 20 files. Scaling all five boundaries by 0.7 or by 1.4 keeps it second or third on TSB-AD-U but moves it to second of 31 on TSB-AD-M, so the registered expectation that first place survives a boundary shift failed. Under a source-cluster bootstrap — sources drawn with replacement, each drawn source contributing all of its files, buckets left empty by a draw ignored, ties given the minimum rank — the median rank is 3 on TSB-AD-U (ranked first in 12% of resamples, in the top three in 60%, with a 95% rank range of $[1, 14]$) and 1 on TSB-AD-M (first in 52%, top three in 90%, range $[1, 13]$; 53% and 99% under a file bootstrap). The wide ranges say that a few sources decide the ordering, so neither benchmark supports a claim of source-invariant rank. The defensible statement is the one Section 5 makes: first by weakest bucket on TSB-AD-M under the official boundaries, first in about half of the source resamples, and among the top three on either benchmark under most perturbations.

C.12 The prescriptions we registered

Every entry of Table 18 was a theory-motivated design prescription with its prediction written into the implementing script before that script was run; no prediction was edited afterwards. This table records design outcomes, not the success rate of the paper's coverage-bound or readout experiments. Two claims of earlier drafts are withdrawn here rather than repaired: that our detector is the flattest across durations (it is 27th in spread; the defensible statement is worst-bucket performance, Appendix C.11), and a count of bucket-level wins that was a pooling artefact. Two of the failures produced results worth stating on their own.

A per-arm correction that changed nothing measurable. A Šidák correction applied per arm changed the suppression rate of the affected arm from 46% to 0% and left VUS-PR identical to four decimal places on every file. We do not generalise this: a shift applied to one arm before a maximum can reorder the fused score, so it is a design change in principle. Our reading of why it was inert here — that the arm it promoted had been winning only where the fused score was already low, a region a rank-based metric weighs little — is a hypothesis about this run, not a property of the metric.

Both relevance measures were measuring window width. We twice attempted a label-free measure of whether a given arm carries signal in a given series. The first, an excess exceedance rate, correlates -0.927 with standalone performance across arms on the shortest bucket — it ranks arms by width, not by usefulness, because a wide-window statistic switches on across its whole window. The second was designed to remove that confound by counting excursions rather than samples and normalising by the arm's own prefix rate; it inverted the sign instead ($\rho = -0.948$ against width). The mechanical cause is that wide-window scores have many ties. In this diagnostic's non-strict thresholding, 37.4% of prefix samples on average meet or exceed the empirical 99th-percentile threshold for the width-400 arm in the 1–3 bucket; the excursion counts are therefore not comparable across widths.

What the disjunction costs on the shortest anomalies. On the 1–3 bucket, the reported combination scores 0.4648, below its width-2 component alone (0.6530), so combining the arms can reduce short-anomaly performance. On the 4–30 bucket it scores 0.3671 against 0.4764 for an external width-16 raw-window nearest-neighbour comparator; that statistic is not a component of the final univariate detector. These comparisons do not isolate whether the loss arises from the merge rule, score calibration or the constituent statistics' false positives. We did not tune a repair on these TSB-AD-U diagnostics. Scoring a repair on TSB-AD-M would extend the sequential history in Appendix B.4; a fresh held-out test would require data not yet used.

D Related work and contributions

D.1 The benchmark and its measure

The benchmark, the leaderboard we analyse and the evaluation measure we report are all from the same line, which itself answers an earlier round of comprehensive evaluations (Schmidl et al., 2022; Paparrizos et al., 2022b) and the finding that the benchmarks then in use were flawed enough to manufacture apparent progress (Wu & Keogh, 2023). Liu & Paparrizos (2024) assemble TSB-AD and argue that the field's comparisons have been unreliable for three reasons — flawed datasets, biased measures and inconsistent protocols — and identify VUS-PR as the most reliable measure among those they test. VUS (Paparrizos et al., 2022a) in turn generalises the range-aware precision and recall of Tatbul et al. (2018), which replaced point-wise counting with a scoring that gives partial credit when a predicted range partially overlaps a true one.

We adopt TSB-AD's files, protocol and measure without modification, and every baseline number in this paper is theirs rather than a re-implementation of ours. This avoids errors from reimplementing baseline detectors. We also never report point-adjusted F_1, following Kim et al. (2022), who show that under point adjustment even a random score can be made to look state of the art.

D.2 The closest prior work: duration is already a reported axis

Two published findings overlap with our Section 2 and we state both plainly.

First, Liu & Paparrizos (2024) report that pre-trained forecasting models perform promisingly *on point anomalies*. Our measurement of a short-anomaly bias in every such model is a sharpening of a published observation, not a discovery, and Section 2 says so.

Second, and closer still, Yeh (2026) benchmarks a matrix-profile detector on TSB-AD and regroups its results by dataset characteristic, including a short-versus-long split of *anomaly duration* at the track-wise median of mean duration. Against the strongest fixed baseline in each track, it finds that the method it studies loses ground on longer anomalies in the univariate track and gains ground on them in the multivariate one — a duration effect, measured independently of us.

What that work does not do is what this paper is for. Its split is binary and concerns one method family; ours is a graded axis across all 40 published detectors, on which the six duration buckets have six different winners. And it does not isolate the effect of aggregation width: its subsequence length is fixed by a period estimator rather than varied, and context width, receptive field and window length do not enter its analysis — the explanations it offers are broader or more repeated anomalous behaviour and cross-channel structure. Our contribution is the account — a bound that says which durations a given aggregation scale can cover, and the reading of that scale as a receptive field —

together with the measurement that the effect is general rather than confined to one family, and of a magnitude comparable to differences among leading methods.

D.3 Detectors, grouped by the context they read

The grouping this paper needs is not architectural family but how much of the series a score reads at once, which cuts across the usual taxonomy. Point and increment statistics read a handful of samples; subsequence methods such as the matrix profile and nearest-neighbour discord search read a subsequence whose length is an explicit parameter (Yeh et al., 2016); patched transformers tokenise fixed-length input segments (Nie et al., 2023); reconstruction and forecasting networks read whatever their receptive field admits (Luo et al., 2016); and the newest entries on the leaderboard are foundation models pre-trained for anomaly detection itself, such as Lan et al. (2026), which score a query segment against its surrounding context and are therefore residual-free but still windowed. How long that window should be is a studied question of its own: Ermshaus et al. (2023) review and benchmark window-size selection for unsupervised time-series analytics, where a single width is chosen per series from its dominant period or autocorrelation, as the benchmark's own metric rule does (Appendix B). That line asks which one width fits a series; Corollary 1 says that no one width fits every duration, and §3.3 measures which width fits which.

The pre-trained forecasters we single out differ from each other along exactly the axis of §3.4.2. Ansari et al. (2024) scales each series and quantises individual observations into a fixed vocabulary, so it emits one token per time step. Rasul et al. (2023) builds its input from lag features and its head emits a distribution over the next single step. Their native output units are individual time steps. Das et al. (2024) is patched: the published model uses an input patch length of 32 and an output patch length of 128, and each decoding step emits a whole output patch. These output units do not fix the anomaly-score readout: a scorer can aggregate residuals across steps or patches. The benchmark's reference scorer requests a horizon of one step (`horizon_len` $= 1$ over a context of 96 samples, stride one) and scores that step's squared error, exactly as it does for Chronos (context 50, `prediction_length` $= 1$) and Lag-Llama (context 96, one step), applying no normalisation of its own (`TSB_AD/models/{TimesFM,Chronos,Lag_Llama}.py` and `model_wrapper.py` of the public repository, read on 2026-09-18). The current TimesFM wrapper names the 1.0–200M checkpoint, unlike our 2.0–500M intervention; the downloaded leaderboard scores do not establish their historical checkpoint (Appendix C.5). These reference scorers all have readout width one, a property of the scoring rule rather than the token or patch size (Table 8). A score computed over one reconstructed patch has that patch's width; it is not an upper bound for other scoring rules.

The closest study of the forecasting mechanism is Zhu et al. (2026). They examine MOMENT (Goswami et al., 2024), Chronos, TimesFM, Time-MoE (Shi et al., 2025) and TSPulse (Ekambaram et al., 2026) across model sizes and context-window lengths and report anomaly-detection performance not significantly different from two one-liner baselines, a moving-window variance and a squared difference. They also explain the failure on sequence anomalies as §3.4.2 does — a context that already contains the anomaly makes it predictable (their §5.2) — and, lengthening the forecasting horizon h with a mean-squared readout over the h-step block, find that longer horizons help long anomalies and hurt short ones, that the best horizon tracks anomaly length, and propose a maximum over horizons as the remedy (their §6.1 and Appendix L). The horizon–length trade-off, and the maximum over horizons, are therefore theirs. What this paper adds is stated as such: the conditioning width varied alongside the readout on the same runs, which their sweep holds apart; a coverage endpoint at fixed false-alarm rate on controlled injections, where duration is set rather than observed; the factorisation of their h into horizon and aggregation (Appendix C.5), which shows that either alone reproduces the effect and that the two are sub-additive on the log-duration scale. In particular, changing aggregation with forecasts held fixed shows that a horizon change is not necessary to move the frontier: temporal redistribution of existing evidence is a separate control. We also run the one-liners at matched readout and provide the conditional long-duration coverage bounds together with the empirical measurement of short-anomaly dilution. The third hypothesis of that test — that both one-liners share the models' readout dependence and match their frontier at equal readout — failed as registered, because the moving-window variance has no readout dependence at all: its window is its aggregation scale. We therefore do not offer their similarity a scale-based explanation.

D.4 MULTI-SCALE MODELS AND HOW THEY MERGE SCALES

Section 3.4.3 measures merge rules, so the practice it bears on needs naming. The prevailing choices for combining scales in multi-scale time-series models are averaging, concatenation followed by a linear projection, pyramid aggregation across resolutions, and learned attention weighting. Wu et al. (2023), for example, reshapes a series by several detected periods and merges the per-period results using softmax weights computed from Fourier amplitudes of the current layer's learned features, without a separate gating network. That this is the prevailing state is not only our reading: Li et al. (2025), arguing for a cross-scale alternative, describe the field as one in which "existing methods often model multi-scale information independently or rely on simple feature fusion strategies". Multi-scale anomaly scores for time series predate the neural line: Kiran (2017) computes a streaming anomaly score at several scales from PCA over lag matrices, and the question of how such scores should be combined is as old as that construction. Liu et al. (2025) is a recent instance that combines two of these forms: each short-term window attends over the long-term context with softmax weights, and the window-level outputs are then concatenated and passed through a linear layer. Its cross-scale merge is therefore attention-weighted, and the caveat below applies to it.

One caveat applies. Corollary 2 bounds a conjunction of *alarms* by its narrowest scale; for real-valued scores what we measure is protocol-specific: on the internal event-level protocol a product and an unweighted average both steepened the duration slope relative to a maximum (Appendix C.8), while under the official protocol the product lost 0.08 VUS-PR and an average of the two arms did not measurably differ from the maximum (Appendix C.9). None of it is a theorem, and a constant uninformative factor is the counterexample to any general version. Attention-weighted fusion is outside even the measured claim: a learned gate can in principle drive a blind branch's weight to zero, which is selection rather than aggregation. Our claim is therefore narrower than "learned fusion is wrong" — it is that, on the statistics we measured, merging scales by a soft-score product was held down by the blind branch, an unweighted average was held down on one protocol and not measurably on the other, and a scheme which can suppress a branch need not be. The maximum itself is not new to this setting: Zhu et al. (2026) propose a maximum over forecasting horizons for the same reason, that no single horizon serves both short and long anomalies; Corollary 2 is the alarm-level statement of that rule, and Appendix C.9 the measurement of when it beats an average.

D.5 LIMITS OF DETECTION AND REFERENCE BEHAVIOURS

Proposition 1 is the standard variational bound on the power of any test at a fixed level, and is not new (Lehmann & Romano, 2005). What is new here is applying it along the duration axis and reading its width as a receptive field. A related distinction arises in change-point detection, from sequential monitoring (Page, 1954) to offline segmentation (Truong et al., 2020): their targets are a change alarm or change locations, rather than per-sample coverage of the shifted segment. Proposition 2 addresses that coverage for local increment statistics under its stated hypotheses; its interior-blindness premise is specific to this class.

Equation (1) measures a pipeline-specific short-anomaly cost. Detecting displaced coordinates among many normal ones is studied in sparse mixture testing (Donoho & Jin, 2004); contiguous blocks additionally supply structure that scan statistics exploit. The following reference tests clarify the role of aggregation and location search under independent Gaussian noise (Appendix A.6).

- A statistic that aggregates the whole window blindly. A signed mean over all w coordinates has signal-to-noise $Lh/(\sigma\sqrt{w})$, giving $h^*/\sigma \propto w^{1/2}L^{-1}$; an energy ($\ell_2$) statistic compares Lh^2 against a null fluctuation of order $\sqrt{w}\,\sigma^2$, giving $w^{1/4}L^{-1/2}$ (Appendix A).
- A statistic that scans for the unknown location of the displaced block pays a multiplicity price rather than a dilution one, giving a duration exponent of $-\frac{1}{2}$ and a width dependence that is *logarithmic*, not a power ((4) in Appendix A).

Our measurement (Section 4) is $h^*/\sigma \propto w^{0.69}L^{-0.33}$, with series-level intervals $[0.46, 1.00]$ on the width exponent and $[-0.44, -0.19]$ on the duration exponent. The duration exponent therefore excludes both the mean aggregator's -1 and the $-\frac{1}{2}$ shared by the energy statistic and the scan: the fitted raw-window pipeline pays a shallower duration cost than either reference, and differs from a

signed window average. The width comparison is less decisive: its interval admits the mean's $\frac{1}{2}$ but excludes the energy statistic's $\frac{1}{4}$, and a scan predicts no power of w at all.

The logarithmic alternative, tested. Because over the fitted widths, $w \in [2, 32]$, a $\sqrt{\log w}$ shape and a weak power law could be confused, we refitted the 12 threshold amplitudes of Section 4 against the scan form (4) and against a generic logarithmic-width form, with the expectation registered beforehand that the power law would win by at least 0.10 in R^2. It does. The scan form with its $L^{-1/2}$ and a free offset inside the logarithm fits worse than the mean of the data ($R^2 = -1.65$); freeing the duration exponent lifts it to 0.72; the best logarithmic form, $L^c(\log w)^d$, reaches 0.80 against the power law's 0.92, a gap of 0.12 (and 11 AIC units) that the power law holds in 99.6% of series-level resamples (median gap 0.12, $[0.03, 0.17]$), and leaving one width out at a time the power law predicts the held-out cells better at every width. Over the fitted range $w \in [2, 32]$ the tested power model therefore predicts these thresholds better than the tested logarithmic forms — a statement about this grid, not a determination of a universal functional form. This compares empirical models of the kNN thresholds, not the performance of a scan detector. The censored model gives a second estimate, 1.13 (Appendix C.6), with additional assumptions and wider-grid non-crossing constraints. An exploratory analysis, registered as such, also examines substrate variation: per-series fits give width exponents with median 0.60 and interquartile range $[0.39, 0.66]$, so the exponent varies across substrates; it does not track the substrate's aggregated-variance (Hurst-type) exponent, which is what we registered (Spearman $+0.19$, $p = 0.26$), but it does track the substrate's lag-one autocorrelation ($+0.76$, $p = 3 \times 10^{-8}$, $n = 38$), a post-hoc association rather than an explanation of the exponent. Together these analyses support a positive width cost for this pipeline and identify substrate dependence without claiming a universal numerical law.

D.6 Contribution relative to prior work

- **Established by prior work:** that foundation models do comparatively well on point anomalies (Liu & Paparrizos, 2024); that pretrained forecasters do not beat one-liner statistics at anomaly detection across sizes and context lengths, that a single-step forecaster misses sequence anomalies because its context already contains them, that a longer forecasting horizon helps long anomalies at the expense of short ones, and that a maximum over horizons remedies it (Zhu et al., 2026); that anomaly duration is a dimension along which TSB-AD results can be regrouped (Yeh, 2026); the variational bound itself (Lehmann & Romano, 2005); the benchmark, protocol and measure (Liu & Paparrizos, 2024; Paparrizos et al., 2022a; Tatbul et al., 2018); the observation that multi-scale models commonly fuse scales simply (Li et al., 2025).
- **This paper adds:** that the duration effect is general across the published field rather than a property of one family, survives a fixed effect for every file, and is of a magnitude comparable to differences among leading methods; a bound that ties detectable duration to the scale at which a statistic aggregates, respected where its hypotheses hold and violated in the contrasting cells; the reading of that scale as a receptive field, with the test that, on two forecasters, the frontier moved with the readout — horizon and aggregation each, once separated — and little with the context over the tested range; and a nonparametric detector whose multiscale structure is motivated by the bounds and whose implementation choices were made on TSB-AD-U, which places among tuned deep models with no gradient training.