PARC: LEARNING HISTORICAL REFERENCES FOR ONLINE CHANGE-POINT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Online change-point detection is challenging when no part of the observed sequence is known to follow the pre-change distribution. Waiting for a long reference can delay detection, while limited history may poorly characterize normal distributional variation. A large learned contrast need not exceed historical fluctuations in the same direction. We propose PARC, a rolling detector that estimates proposal-conditioned historical response distributions. Each proposal is defined by a recent-versus-earlier contrast at a particular time and scale, describing how the distribution may be changing. Conditioned on that proposal, PARC estimates the corresponding historical response distribution and learns how much that type of variation has occurred before. Different candidate changes are therefore evaluated against different historical variability. New proposals update the collection of references, while each comparison is fixed before later observations are tested. We analyze score sensitivity, the effect of startup on later thresholds, and a separate finite-horizon false-alarm guarantee under exchangeable normal sequences. Across two complete archives, PARC achieves higher detection coverage and lower capped detection loss than twelve evaluated baselines, with fewer early alarms than the strongest detection comparator on each archive under a matched protocol.

1 INTRODUCTION

Industrial sensors, wearables, and deployed prediction systems need early change detection to trigger inspection or model revision. Yet a new sequence provides no known pre-change segment. Even on annotated benchmarks, the true change time cannot be used to choose the detector's reference (Truong et al., 2020; Van den Burg & Williams, 2020).

Long initial windows can miss early changes, while short histories can make persistent fluctuations appear anomalous. Streaming methods update summaries or discrepancies (Keriven et al., 2020; Kalinke & Gavioli-Akilagun, 2025). Classifier and witness tests learn comparisons and separate construction from evaluation (Jang et al., 2022; Kübler et al., 2022). Adaptation and later-data testing are therefore not, by themselves, new.

The remaining question is how to characterize historical variation for the particular change being tested. A sequence may have large frequency fluctuations but a stable level. A reference that describes one pattern poorly may obscure a change in the other. Updating history alone does not specify how its variability should be measured for each candidate change.

We propose PARC (Proposal-Adaptive Reference Comparisons), a rolling detector that estimates proposal-conditioned historical response distributions (Figure 1). A recent-versus-earlier contrast proposes feature weights describing a possible change at a particular time and scale. Applying those weights to historical blocks reveals how strongly the same pattern varied before. Different proposals can therefore produce different reference distributions from the same history. The reference adapts to the proposed pattern, not only to the arrival of more data.

The empirical reference records both historical extremeness and variability. A later block must exceed its largest sampled response, and the excess is scaled by historical variability. This prevents a large contrast from the historical average alone from producing a positive score. The detector caches these functionals rather than fitting a density. Each comparison is fixed before later testing. New

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.

proposals update the collection as old ones expire. Learned short-window features enable startup and are replaced by native long-window features once available.

Our contributions are:

- We estimate a historical response distribution for each proposed change, so its pattern, time, and scale determine the historical variation used to judge later observations, without a supplied unchanged segment.
- We identify when historical coverage suppresses recurring variation while an aligned change produces excess, and characterize representation and startup costs. A separate finite-horizon guarantee requires whole-sequence exchangeability.
- We demonstrate higher detection coverage and lower capped loss than twelve baseline scores on complete TSSB and TCPD archives with held-out families, supported by component, theory-validation, and computational studies.

2 Related work

Online distribution comparisons. CUSUM accumulates evidence for a specified change model (Page, 1954). Energy and kernel distances compare distributions more broadly (Székely & Rizzo, 2013; Gretton et al., 2012). NEWMA maintains summaries at two time scales, online RFF tests use random features, and online kernel density-ratio methods adapt comparisons between consecutive intervals (Keriven et al., 2020; Kalinke & Gavioli-Akilagun, 2025; Ferrari et al., 2023). Scan-B standardizes blockwise MMD using background data, while kernel CUSUM compares candidate change locations and derives ARL and delay approximations (Li et al., 2019; Wei & Xie, 2026). Hahn CUSUM, MdFOCuS, QuantTree EWMA, and predictive-mixture CuSum provide further modeling and initialization choices (Hahn, 2022; Pishchagina et al., 2026; Frittoli et al., 2023; Halme et al., 2026). PARC tests excess over each proposal's empirical historical endpoint. We do not claim that existing methods use only fixed references or means.

Learned comparisons and representations. Sequential C2ST, contrastive online CPD, and witness tests already learn discriminative comparisons and evaluate subsequent or separate observations (Jang et al., 2022; Puchkin & Shcherbakova, 2023; Kübler et al., 2022). NN-CUSUM connects neural training to online ARL and delay guarantees (Gong et al., 2022). Li et al. (2024) learn offline change tests from labeled examples. Spectrally normalized self-supervised representations also have CPD-specific theoretical support (Bazarova et al., 2024). WiTS already tests a mean difference through a learned scalar witness (Kübler et al., 2022), so projection-dependent testing is not new. PARC uses the empirical endpoint and scale of each proposal's historical response distribution for later testing, without an initial unchanged segment or online network training.

Initialization and sequential validity. Bayesian online CPD infers run length through a predictive model and change prior (Adams & MacKay, 2007), providing an established approach without a supplied reference. Offline segmentation can use both sides of a boundary (Truong et al., 2020; Ermshaus et al., 2023), whereas our first-alarm evaluation uses only observed prefixes. CALM sets sequential thresholds by reference resampling (Cobb et al., 2022). E-C2ST provides sequential classifier-based evidence, e-detectors control ARL under specified null classes, and WATCH uses weighted-conformal martingales for adaptive deployment monitoring (Pandeva et al., 2024; Shin et al., 2024; Prinster et al., 2025). Our optional guarantee applies standard split-rank inference to whole finite-horizon trajectories (Shafer & Vovk, 2008; Angelopoulos & Bates, 2021). It requires exchangeable sequence units and does not justify the empirical archive thresholds. Appendix C.3 distinguishes implemented baselines from methods discussed here.

3 Problem definition

We observe a multivariate time series $x_1, x_2, \ldots$, with $x_t \in \mathbb{R}^d$, sequentially. Let P be its distribution and P_0 the unknown baseline law under no change. Both are distributions over complete time series, allowing temporal dependence. Write $P^{1:t}$ and $P_0^{1:t}$ for their distributions on the first t observations.

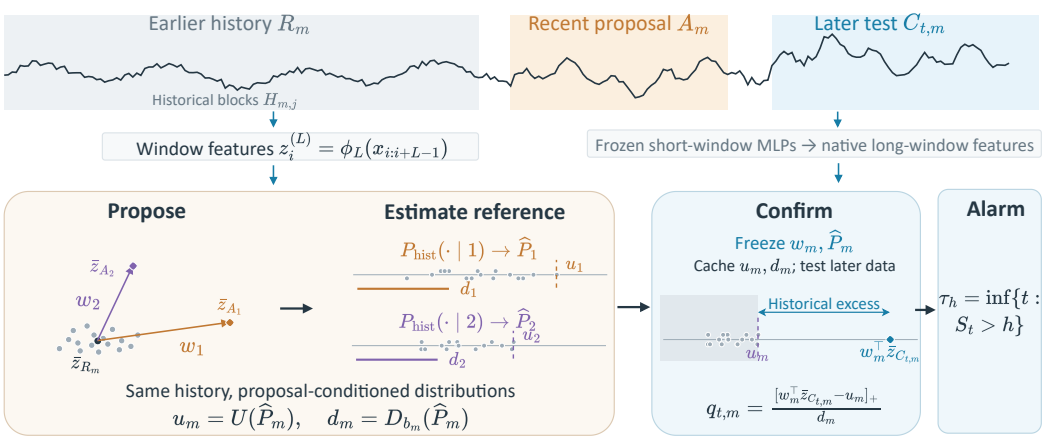

Figure 1: PARC's propose–calibrate–confirm construction. Two directions induce different empirical reference distributions on the same historical blocks. Cached endpoints and scales are frozen before later testing. Scores are pooled and filtered for persistence. Native features replace startup at 175 observations without resetting history. The geometry and time series are schematic.

The first change ν is the earliest t for which $P^{1:t} \neq P_0^{1:t}$. We use $\nu = \infty$ to denote no change. Marginal, cross-channel, and temporal changes are included.

No initial segment is supplied as unchanged, so waiting to estimate P_0 can miss early changes, while using a short history can mistake fluctuations for a shift. Our internal reference $P_{\text{hist}}(\cdot \mid m)$ describes historical block responses under proposal m, not the unknown law P_0. Its empirical estimate $\widehat{P}_m$ is constructed from observed history, which may already contain changes (Section 4). Parameters θ and threshold h are fitted before deployment, possibly using separately labeled sequences. The score S_t uses only $x_{1:t}$, not test annotations or the final record length.

The objective is prompt detection with few false alarms. The first alarm is

$$\tau_h = \inf\{t \geq 1 : S_t > h\}, \qquad \inf \varnothing = \infty. \tag{1}$$

It is a stopping time for $\mathcal{F}_t = \sigma(x_{1:t}, \theta, h)$. An alarm before ν is early. Detection succeeds when $\nu \leq \tau_h < e$, with delay $\tau_h - \nu$, where e is the evaluation endpoint. Test annotations define these evaluation targets, not detector initialization.

The optional bound on $P_0(\tau_h \leq H)$ requires a fixed horizon H and separate normal sequences jointly exchangeable with the normal test sequence. These threshold-setting units are not the internal historical blocks (Section 5). Archive alarm rates remain empirical. We study changes in the input distribution. Without response labels, we do not assume that $P(Y \mid X)$ remains unchanged, as supervised covariate shift requires.

4 Method

4.1 Proposal-conditioned historical reference

PARC estimates how much variation of each proposed type has already occurred. Past contrasts specify a direction, time, and scale. Historical blocks measured in that direction define a response distribution. Its empirical endpoint and scale determine how later observations are scored. Figure 1 shows the propose–calibrate–confirm construction. Calibration here estimates this historical reference, not a probability of change. The final alarm threshold is fitted separately. Each comparison is frozen before testing, while new proposals update the active collection.

4.2 BOUNDED FEATURES BEFORE AND AFTER STARTUP

We write $z_i^{(L)} = \phi_L(x_{i:i+L-1})$ for a feature vector available when its last raw observation arrives. The mature map uses $L = 64$ with 16 square-root normalized spectral powers and bounded mean and log-scale coordinates per channel. Normalization gives $\|z_i^{(L)}\|_2 \le 1$ (Appendix B).

For $L \in \{8, 16, 32\}$, frozen two-hidden-layer MLPs predict spectral descriptors from short windows. Synthetic training pairs end at the same time and minimize squared descriptor error, without future inputs. We compute mean and scale coordinates directly, using fixed synthetic-validation factors. The native short-feature control substitutes zero-padded spectra for the MLPs.

Using training families, we select raw input or a coordinatewise transform fixed after eight observations, without assuming those observations are unchanged. Historical and later blocks use the same preprocessing and feature map.

4.3 FROM PAST DISTRIBUTIONS TO A CHANGE PROPOSAL

Index a proposal by m, including its decision time, context L_m, and feature-block width b_m. For a block B of feature indices, let $\widehat{Q}_B = |B|^{-1} \sum_{i \in B} \delta_{z_i^{(L_m)}}$ be its empirical feature distribution, where δ_z is unit mass at z. Its mean is $\bar{z}_B = \int z \, d\widehat{Q}_B(z)$. We propose a direction through the mean contrast between the earlier and recent distributions $\widehat{Q}_{R_m}, \widehat{Q}_{A_m}$,

$$\Delta_m = \bar{z}_{A_m} - \bar{z}_{R_m}, \qquad w_m = G_m(R_m, A_m) = \frac{\eta_m \odot \Delta_m}{\max\{\|\eta_m \odot \Delta_m\|_2, \epsilon\}}, \tag{2}$$

with $\epsilon = 10^{-12}$. At startup, we set $\eta_m = \mathbf{1}$, so larger absolute feature differences receive larger weight magnitudes. The mature detector reweights differences with a frozen gate trained on synthetic partial-cycle errors and evaluated on past block statistics. This proposes a pattern, not an optimized full distribution distance. Write $f_w(z) = w^\top z$ for the scalar response, distinct from ϕ_{L_m}.

4.4 CONDITIONAL REFERENCE AND EMPIRICAL SCORE

Given the frozen history of proposal m, let $Q_{H,m}$ be the distribution of the feature mean of a uniformly selected width-b_m block within R_m. The proposal-conditioned response distribution is

$$P_{\text{hist}}(B \mid m) = \int \mathbf{1}\{f_{w_m}(z) \in B\} \, dQ_{H,m}(z). \tag{3}$$

Here B is a measurable subset of the real line. The conditioning specifies the proposal and its historical snapshot, not a law for future unchanged data. Appendix A.2 gives the sampling construction.

We approximate this distribution using $J = 16$ historical blocks $\mathcal{H}_m = (H_{m,1}, \ldots, H_{m,J})$ with evenly spaced starts. For any direction w, their empirical response distribution is

$$\widehat{P}_m^{\,w} = J^{-1} \sum_{j=1}^{J} \delta_{f_w(\bar{z}_{H_{m,j}})}. \tag{4}$$

We use $\widehat{P}_m = \widehat{P}_m^{\,w_m}$ in the implementation. For a finite empirical distribution F, define its upper endpoint $U(F) = \max \operatorname{supp}(F)$ and scale $D_b(F) = \sqrt{\operatorname{Var}_{Y \sim F}(Y)} + 1/b$. Their empirical values are

$$u_m(w) = U(\widehat{P}_m^{\,w}), \qquad d_m(w) = D_{b_m}(\widehat{P}_m^{\,w}). \tag{5}$$

Thus u_m measures historical extremeness, while d_m measures variability across projected block means, using standard deviation with divisor J. Both depend on the proposal, even for the same historical blocks. Caching $\mathcal{R}_m(w) = (u_m(w), d_m(w))$ preserves the score exactly without density fitting or rescanning history.

At each startup decision and width, retain the context maximizing the signed past score $(f_{w_m}(\bar{z}_{A_m}) - u_m(w_m))/d_m(w_m)$. Cache its direction and reference as $u_m = u_m(w_m)$, $d_m = d_m(w_m)$. Only a later test block $C_{t,m}$ contributes

$$q_{t,m} = \frac{[v_{t,m} - U(\widehat{P}_m)]_+}{D_{b_m}(\widehat{P}_m)}, \qquad v_{t,m} = f_{w_m}(\bar{z}_{C_{t,m}}). \tag{6}$$

A positive score requires exceeding every sampled historical response under the same weights. The scale reduces scores in historically variable directions, and its floor keeps constant histories well defined. This is a functional of an estimated historical distribution, not a probability of change, a p-value, or a confidence bound. We fit the final alarm threshold h separately. Later testing avoids reusing proposal data as test observations, but does not imply independence under temporal dependence. Negative feature weights allow decreases to produce positive excess. A reversal or an orthogonal change may require a new proposal.

4.5 When to use the learned startup comparison

Startup contexts are $\{8, 16, 32, 64\}$, with block widths $1, 2, 4, \ldots$. For width b, a proposal is considered every b observations. The entire raw test window starts after its direction/context decision. A proposal is retained for $8b_m$ sliding test starts. Its cached w_m, u_m, d_m are never refitted on its test observations. The active set $\mathcal{M}_t$ contains only retained comparisons whose later test blocks are available and whose lifetimes have not expired. This updates the family of conditional references, not an existing reference with the observations currently being tested. Set $V_t = \max_{m \in \mathcal{M}_t} q_{t,m}$, or zero if $\mathcal{M}_t$ is empty. Define the persistence filter

$$\mathcal{P}_2(V)_t = \min\{V_{t-1}, V_t\}. \tag{7}$$

The mature detector fixes $L = 64$, uses widths $8, 16, 32, \ldots$, and applies the same historical-excess comparison with its gated directions. Its first possible persistent score requires $r = 175$ observations. The complete score is

$$S_t = \begin{cases} \mathcal{P}_2(V^{\text{short}})_t, & t < r, \\ \mathcal{P}_2(V^{\text{mature}})_t, & t \geq r. \end{cases} \tag{8}$$

Only one stage supplies the score. The mature history is not restarted at r. Replacement limits continued short-feature use, although startup maxima can still affect the threshold. Algorithm 1 and Appendix A.3 give the procedure and indices. Appendix C.7 varies the history size, floor, lifetime, persistence, and replacement count. The count 175 marks availability, not a universally optimal transition.

5 When historical excess distinguishes change

Each proposal determines both the response measured on later data and the historical variation used to judge it. We explain when this comparison suppresses recurring fluctuations while retaining an aligned change.

Proposition 1 (Historical coverage and aligned change)**.** *Fix a proposal w and width $b > 0$. Let $F_\star$ be its response law over* all *admissible historical blocks, with support endpoints $\ell_\star, u_\star$. For a law F formed from a subset of these blocks, let $u = \max \operatorname{supp}(F)$, $\varepsilon = u_\star - u \geq 0$, $r_w = u_\star - \ell_\star$, and $d = \operatorname{sd}(F) + 1/b$. Compare a recurring block with mean response in $[\ell_\star, u_\star]$ and a later block that is such a recurrence plus a feature-mean shift δ. Writing $\gamma = w^\top \delta$, their scores satisfy*

$$0 \leq q_{\text{rec}} \leq \frac{\varepsilon}{d}, \qquad q_{\text{new}} \geq \frac{[\gamma - r_w + \varepsilon]_+}{d}. \tag{9}$$

With $\varepsilon = 0$, responses within the historical range receive zero excess. If $\gamma > r_w$, the projected shift produces positive excess. Within one scored stage, a bound above h at two consecutive eligible evaluations suffices for an alarm by the second. A useful direction captures change relative to historical variation, not merely a large recent contrast. Missing extremes or poor alignment can defeat separation. This per-proposal result is not an overall false-alarm or population power guarantee (proofs and checks in Appendices A.4 and D.1).

Representation and startup costs. Appendix A bounds score error through endpoint, mean, and scale errors (Proposition 2). Startup can raise the threshold even with unchanged mature scores (Proposition 3).

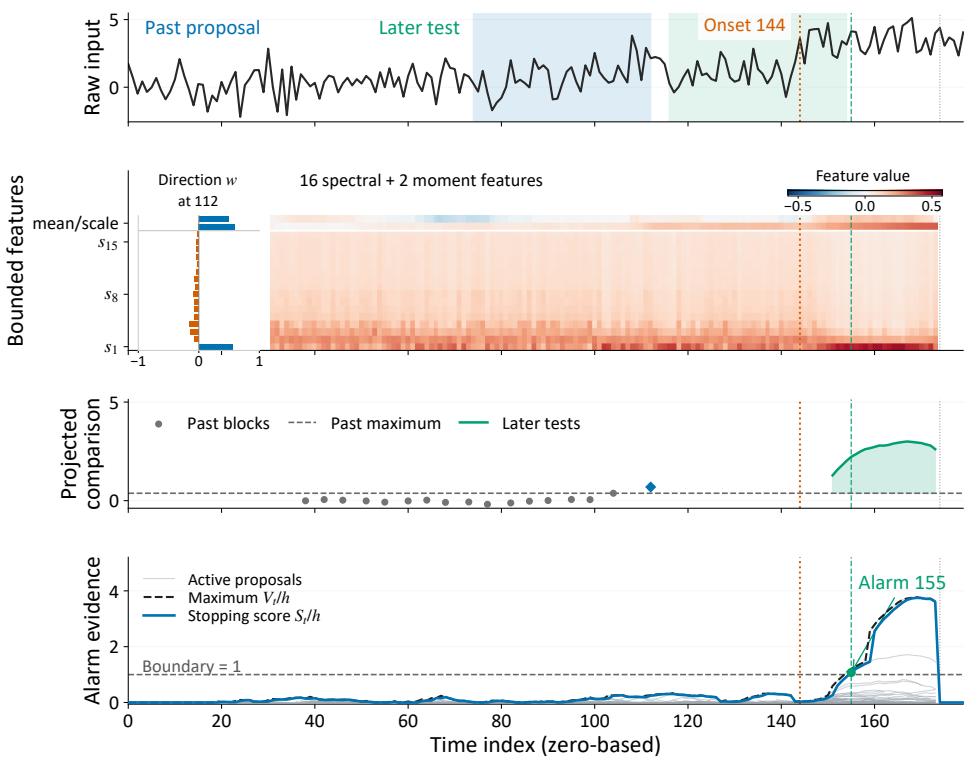

Figure 2: Later-data testing on TCPD's quality-control series. Panels show proposal/test blocks; 18 bounded features and a frozen direction; historical and later responses; and eligible scores (gray), their maximum (black dashed), and persistence score (blue). The illustrated direction is traced from the alarm, not selected by its annotation. Post-alarm scoring is illustrative (Appendix C.5).

A separate finite-horizon guarantee. For a detector fixed on disjoint design data and a fixed horizon H, the split-rank threshold in Theorem 1 gives

$$\mathbb{P}_0(\tau_h \le H \mid \text{fixed design}) \le \alpha. \tag{10}$$

This requires normal reference and test units to be exchangeable conditional on the design. Within-sequence dependence is unrestricted. The probability averages over reference and test units and does not cover label-selected benchmark thresholds or unbounded monitoring (Shafer & Vovk, 2008; Angelopoulos & Bates, 2021).

6 Experiments

6.1 Data, fitting, and evaluation

We evaluate all 75 TSSB records and all 42 records in the recovered TCPD inventory. There are 69 and 40 changed records, respectively, with six and two null records. TSSB is a segmentation collection constructed from classification time series, whereas TCPD contains diverse application series (Ermshaus et al., 2023; Van den Burg & Williams, 2020). Change annotations define the first-alarm evaluation targets. Classification labels and recommended periods are not detector inputs. For TSSB the target interval ends at the next annotation. For TCPD it ends at the record end. Appendix C describes the datasets and evaluation rules.

We evaluate every method with five outer family folds, three inner family folds, and five repeated outer partitions. Inner folds select preprocessing, configuration, and an early-alarm budget. We refit thresholds on the outer training families. Test families do not participate. Both archives informed architecture development. The family splits separate configuration fitting and evaluation, but do not remove the effects of earlier design decisions. Network weights use synthetic training data only.

Table 1: Complete non-oracle comparison, averaged over five partitions. Detection/early-alarm percentages use changed records (69 TSSB, 40 TCPD). Null-alarm percentages use small no-change sets (6 TSSB, 2 TCPD). Capped loss includes failures and is not a percentage. Bold marks our method, not winners. Counts and successful-only delays are in Appendix Table 4.

	TSSB				TCPD			
Method	Detected changes (%) ↑	Early alarms (%) ↓	Null alarms (%) ↓	Capped loss ↓	Detected changes (%) ↑	Early alarms (%) ↓	Null alarms (%) ↓	Capped loss ↓
PARC	**60.0**	**2.9**	**16.7**	**0.590**	**51.0**	**4.0**	**0.0**	**0.653**
C2ST-MLP	43.5	8.4	10.0	0.723	26.5	11.5	60.0	0.867
Energy	37.1	9.6	0.0	0.703	34.5	13.5	30.0	0.792
Hahn CUSUM	45.8	5.8	3.3	0.691	28.0	9.0	70.0	0.819
KS	38.8	12.2	0.0	0.690	31.5	8.0	20.0	0.814
MdFOCuS	43.8	3.5	3.3	0.717	36.5	14.0	30.0	0.788
Mean	29.6	7.2	10.0	0.765	28.0	15.5	50.0	0.819
NEWMA	19.7	14.5	30.0	0.872	37.0	6.5	30.0	0.768
Online RFF-MMD	48.4	10.7	6.7	0.663	34.5	10.0	30.0	0.749
PM-CuSum (M)	33.0	7.2	16.7	0.803	39.5	12.0	40.0	0.732
QT-EWMA	19.1	11.9	16.7	0.860	27.0	10.0	60.0	0.857
RBF-MMD	32.8	9.3	0.0	0.754	34.0	15.0	60.0	0.809
Variance	24.6	9.0	30.0	0.822	34.5	11.0	70.0	0.803

Deployment scoring uses neither annotations nor final record length. Changes before readiness remain failures if not detected later within the target interval.

The twelve comparison scores are Energy, RBF–MMD, KS, mean, variance, C2ST–MLP, Online RFF–MMD, NEWMA, QT–EWMA, Hahn CUSUM, Gaussian MdFOCuS, and Gaussian-mean PM-CuSum. We include available original recurrences as configuration choices. Baselines receive short-start as well as longer reference/window candidates and are not forced to wait for our mature context. These are matched implementations and operating points, not a comparison of every method under its publication's original evaluation.

We report detections, early alarms, null alarms, and successful-only delay. To include missed and early detections in a timing measure, capped delay loss is $(\tau - \nu)/(e - \nu)$ for a successful alarm before target end e, and one otherwise. Lower is better. Table 1 reports detection and early-alarm percentages over changed records and null-alarm percentages over no-change records. Counts remain available in Appendix Table 4. Repeated partitions reuse the same records and are not independent new test samples.

6.2 Complete-archive comparison

PARC detects 41.4/69 TSSB changes versus 33.4 for Online RFF-MMD, the strongest comparator by detection count, with 2.0 versus 7.4 early alarms. On TCPD it detects 20.4/40 versus 15.8 for PM-CuSum (M), with 1.6 versus 4.8 early alarms. It has the lowest capped delay loss among the evaluated scores on both archives. TSSB null alarms are 1.0/6, exceeding several baselines. TCPD has no observed null alarms on its two null records. PARC therefore improves detection coverage and capped loss, although some baselines have fewer null alarms. The small null sets do not establish a population false-alarm rate.

Appendix Figure 6 separates additional detections, lost detections, and delay on shared detections. Coverage improves more consistently than delay. Capped loss includes failures omitted by successful-only median delay.

6.3 Which part of the detector helps?

All TSSB changes occur after the mature transition, so TSSB cannot establish an early neural advantage. Its clearest component result is the cost of leaving short contexts active indefinitely. TCPD includes 29 changes before the transition: PARC detects 14.0 of these versus 11.4 for the native

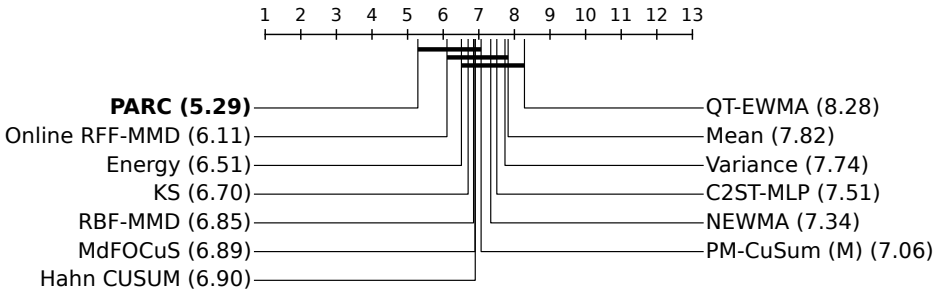

Figure 3: Family-level capped-loss ranks. Parentheses give each method's mean rank (lower is better). Predictions are averaged within record, then family. Black segments connect methods whose rank differences are not significant under the Nemenyi test. The calculation does not account for earlier architecture selection or overlap between training folds.

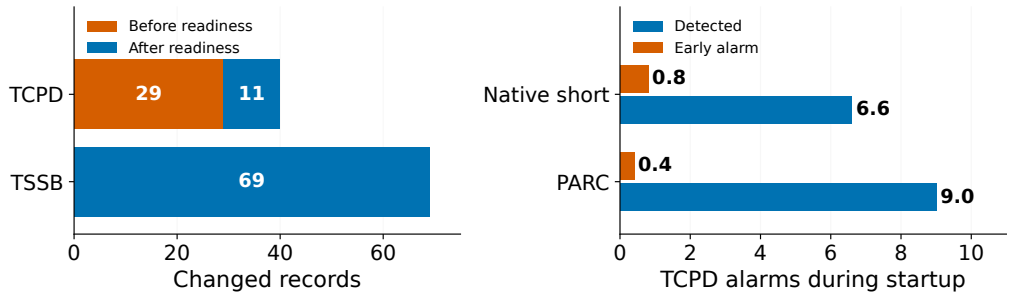

Figure 4: Changes and alarms during startup. Left: change onsets before and after the mature detector becomes available. Right: TCPD first alarms before the transition, averaged over five partitions. An early change may be detected after startup ends.

short-feature control. Actual successful alarms before the transition are 9.0 versus 6.6. These are distinct from eventual detections of early changes. This gain concerns startup detection on TCPD, not every time scale or dataset.

Keeping learned contexts active indefinitely reduces TSSB detections from 41.4 to 17.6. On TCPD it increases detections from 20.4 to 20.8, but also early alarms from 1.6 to 3.0 and adds null alarms. These results support replacing short-context comparisons after startup, although they do not establish an optimal transition time. The limited variants share the same mature score. Their different late detections also reflect earlier alarms and separately fitted thresholds, not a better mature representation.

Appendix C.6 reports three training seeds and startup gating. Across 17 configurations with matched preprocessing and fitting budgets, no configuration is best on both archives. History size and proposal lifetime matter more than persistence length (Appendix C.7). Random-direction and strongest-feature controls vary the proposal. The historical-mean control replaces the reference endpoint by its mean. These controls isolate proposal choice and the reference functional, without establishing uniform superiority of either.

6.4 Numerical checks of the theory

Numerical checks validate theory, not detection gains. Distribution-level checks cover 2,000 law pairs and 300 full-history/grid comparisons without bound violations, alongside the 2,700 feature perturbations and context reselection checks. For rank validity, 3,920 normal sequences provide 80 independent experiments under two dependent regimes, horizons 128, 256, 512, and reference sizes 9, 19, 39. Matched sampling is consistent with the marginal five-percent bound. Nine references give an infinite threshold.

Table 2: Component comparisons under the same held-out selection protocol. Each variant has its own refitted threshold. D and E denote detected changes and early alarms (mean counts); Loss is capped delay. The native control removes the short-window network. All limited variants share the same mature score.

Variant	TSSB D	E	Loss	TCPD D	E	Loss
PARC	41.4	2.0	0.590	20.4	1.6	0.653
Native short features	40.8	2.4	0.597	17.2	2.8	0.724
Fixed 64-point context	40.8	2.4	0.597	16.0	2.2	0.759
No startup retirement	17.6	4.8	0.837	20.8	3.0	0.620

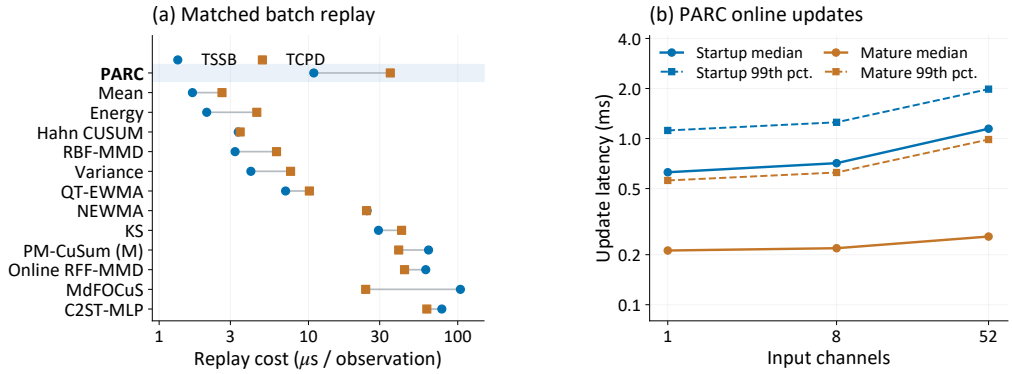

Figure 5: Batch replay cost (left) and online latency (right). Solid/dashed curves show median/99th-percentile latency. Both use one numerical thread and exclude setup. Different workloads preclude cross-panel speed comparisons. Protocols are in Appendix E.

For 512 observations and 19 references, null alarms rise from 4.4% with a matched horizon to 52.3% when the threshold uses only 128 observations. The startup-cost inequality holds in all 480 finite-threshold cases. Appendix D gives sampling details, experiment-level bootstrap intervals, and controls. Numerical checks supplement the proofs.

6.5 Inference cost

On three length-selected records per archive, PARC takes 10.9/35.3 μs per observation on TSSB/TCPD, versus 61.0/44.0 for Online RFF–MMD and 78.2/62.1 for C2ST–MLP. Mean and energy are faster. Separately, an equivalent incremental implementation has 0.26/0.99 ms median/99th-percentile mature latency for 52 channels. At 8,192 observations, state/peak memory is 123/303 MB. Memory grows with history, and latency is not worst-case bounded. Appendix E specifies both protocols.

7 Conclusion

We introduced PARC to detect change without a declared unchanged segment. Each proposal defines a historical response distribution against which we test later observations. Empirical endpoints and scales summarize past variation without density fitting. Each reference is fixed before testing. The same history can therefore provide different comparison levels for different proposed changes, rather than one fixed reference for all patterns. Across two complete archives, PARC improves detection coverage and capped loss over twelve evaluated baselines, with fewer early alarms than the strongest detection comparator on each archive. Our theory relates separation to historical coverage and direction alignment, and characterizes representation and startup threshold costs. The separate false-alarm guarantee requires exchangeable normal sequences and a fixed horizon. History containing changes, limited directions, and growing memory remain limitations. Improving historical coverage with bounded memory remains an important direction for future work.

REFERENCES

Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021. URL https://arxiv.org/abs/2107.07511.

Alexandra Bazarova, Evgenia Romanenkova, and Alexey Zaytsev. Normalizing self-supervised learning for provably reliable change point detection. In *IEEE International Conference on Data Mining*, pp. 21–30, 2024. doi: 10.1109/ICDM59182.2024.00009.

Oliver Cobb, Arnaud Van Looveren, and Janis Klaise. Sequential multivariate change detection with calibrated and memoryless false detection rates. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 226–239, 2022. URL https://proceedings.mlr.press/v151/cobb22a.html.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006. URL https://www.jmlr.org/papers/v7/demsar06a.html.

Arik Ermshaus, Patrick Schäfer, and Ulf Leser. ClaSP: Parameter-free time series segmentation. *Data Mining and Knowledge Discovery*, 37:1262–1300, 2023. doi: 10.1007/s10618-023-00923-x.

André Ferrari, Cédric Richard, Anthony Bourrier, and Ikram Bouchikhi. Online change-point detection with kernels. *Pattern Recognition*, 133:109022, 2023. doi: 10.1016/j.patcog.2022.109022.

Luca Frittoli, Diego Carrera, and Giacomo Boracchi. Nonparametric and online change detection in multivariate datastreams using QuantTree. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8328–8342, 2023. doi: 10.1109/TKDE.2022.3201635.

Tingnan Gong, Junghwan Lee, Xiuyuan Cheng, and Yao Xie. Neural network-based CUSUM for online change-point detection. *arXiv preprint arXiv:2210.17312*, 2022.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Georg Hahn. Online multivariate changepoint detection with type I error control and constant time/memory updates per series. *Statistics & Probability Letters*, 181:109258, 2022. doi: 10.1016/j.spl.2021.109258. URL https://doi.org/10.1016/j.spl.2021.109258.

Topi Halme, H. Vincent Poor, and Visa Koivunen. Adaptive sequential change detection using mixtures of predictive distributions, 2026. URL https://arxiv.org/abs/2606.05072v3. Version 3, preprint.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Sooyong Jang, Sangdon Park, Insup Lee, and Osbert Bastani. Sequential covariate shift detection using classifier two-sample tests. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9845–9880. PMLR, 2022. URL https://proceedings.mlr.press/v162/jang22a.html.

Florian Kalinke and Shakeel Gavioli-Akilagun. Optimal online change detection via random Fourier features. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0333.

Nicolas Keriven, Damien Garreau, and Iacopo Poli. NEWMA: A new method for scalable model-free online change-point detection. *IEEE Transactions on Signal Processing*, 68:3515–3528, 2020. doi: 10.1109/TSP.2020.2990597. URL https://doi.org/10.1109/TSP.2020.2990597.

Jonas M. Kübler, Wittawat Jitkrittum, Bernhard Schölkopf, and Krikamol Muandet. A witness two-sample test. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 1403–1419, 2022.

Jie Li, Paul Fearnhead, Piotr Fryzlewicz, and Tengyao Wang. Automatic change-point detection in time series via deep learning. *Journal of the Royal Statistical Society Series B*, 86(2):273–285, 2024. doi: 10.1093/jrsssb/qkae004.

Shuang Li, Yao Xie, Hanjun Dai, and Le Song. Scan B-statistic for kernel change-point detection. *Sequential Analysis*, 38(4):503–544, 2019. doi: 10.1080/07474946.2019.1686886. URL https://arxiv.org/abs/1507.01279.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100.

Victor M. Panaretos and Yoav Zemel. Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6:405–431, 2019. doi: 10.1146/annurev-statistics-030718-104938. URL https://arxiv.org/abs/1806.05500.

Teodora Pandeva, Tim Bakker, Christian A. Naesseth, and Patrick Forré. E-valuating classifier two-sample tests. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=dwFRov8xhr.

Liudmila Pishchagina, Gaetano Romano, Paul Fearnhead, Vincent Runge, and Guillem Rigaill. Online multivariate changepoint detection: Leveraging links with computational geometry. *Journal of the Royal Statistical Society: Series B*, 88(1):171–194, 2026. doi: 10.1093/jrsssb/qkaf046.

Drew Prinster, Xing Han, Anqi Liu, and Suchi Saria. WATCH: Adaptive monitoring for AI deployments via weighted-conformal martingales. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49830–49859. PMLR, 2025. URL https://proceedings.mlr.press/v267/prinster25a.html.

Nikita Puchkin and Valeriia Shcherbakova. A contrastive approach to online change point detection. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5686–5713, 2023.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. E-detectors: A nonparametric framework for sequential change detection. *New England Journal of Statistics in Data Science*, 2(2):229–260, 2024. doi: 10.51387/23-NEJSDS51.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020. doi: 10.1016/j.sigpro.2019.107299.

G. J. J. Van den Burg and C. K. I. Williams. An evaluation of change point detection algorithms. *arXiv preprint arXiv:2003.06222*, 2020.

Song Wei and Yao Xie. Online kernel CUSUM for change-point detection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(4):1251–1277, 2026. doi: 10.1093/jrsssb/qkag020.

A PROOFS AND STATISTICAL SCOPE

The analysis treats the score as a functional of an estimated proposal-conditioned historical response distribution. Its endpoint and scale describe historical extremeness and variability for the proposed pattern. We first define this distribution and its empirical approximation, then distinguish feature error at a fixed proposal from changes in the selected weights or context. Only the separate rank theorem provides a false-alarm probability guarantee under whole-sequence sampling assumptions.

A.1 DISTRIBUTIONAL TARGET AND OBSERVATION PROTOCOL

The baseline P_0 is a probability law on multivariate time series, not a single distribution from which observations must be drawn independently. For a possible deployment law P, write $P^{1:t}$ and $P_0^{1:t}$ for their joint distributions on the first t observations. The idealized first distributional change is

$$\nu(P, P_0) = \inf\{t \geq 1 : P^{1:t} \neq P_0^{1:t}\}. \tag{11}$$

Thus a change at ν leaves every pre-change prefix law unchanged, but changes the joint law by ν. Equality here means equality of probability measures, not equality between two realized sample windows. In particular, one-dimensional marginals may remain the same while temporal or cross-channel dependence changes. The definition imposes neither a known dependence order nor independence across the boundary. It also does not require a permanent post-change regime. The first change remains the target after a later reversal.

The change is defined relative to a baseline population. This definition does not make two arbitrary unknown distributions or their change time identifiable from every finite time series. Arbitrarily small changes, changes before adequate observations arrive, and short-lived changes can remain undetectable. The deterministic results in this paper give no universal power guarantee. Benchmark annotations serve as operational targets and need not identify an exact population-law change.

The task allows a time-varying baseline law, provided it is specified as a law on complete sequences. Known variation already included in that law is not, by itself, a change. For the optional rank guarantee, independently supplied normal units describe this baseline population. They are distinct from the internal reference blocks R constructed during monitoring. The same scoring procedure, including all within-sequence adaptation, is applied to each normal unit and the test unit. A horizon H is declared before monitoring. It is not the sequence's eventual length supplied as an input.

Evaluation uses the first alarm only. The endpoint e is exclusive: it is the next annotated boundary when the protocol uses one, or the boundary just after the last observed sample. On a changed record, no alarm or an alarm outside $[\nu, e)$ is a failure. On a no-change record, any alarm within the evaluation horizon is a false alarm. Unavailable early scores cannot create a detection. They do not justify excluding an early change from evaluation. Appendix C specifies how annotations provide these targets for each archive.

A.2 CONDITIONAL HISTORICAL DISTRIBUTION AND ITS ESTIMATE

A candidate index m specifies a decision time, context L_m, and width b_m. With $z_i^{(L_m)} = \phi_{L_m}(x_{i:i+L_m-1})$, define $\bar{z}_B = |B|^{-1} \sum_{i \in B} z_i^{(L_m)}$ for a block of feature-start indices B. The earlier block R_m and recent block A_m determine $w_m = G_m(R_m, A_m)$ by Eq. (2). The subscript on G_m specifies the context, width, and startup or mature gate. It does not denote a separately trained network for each proposal.

The proposal is a rule for viewing history, not an assertion that the recent block has changed. Both $\widehat{Q}_{R_m}$ and $\widehat{Q}_{A_m}$ are empirical distributions of individual feature vectors. Their mean contrast determines w_m. The historical reference below is instead a distribution of width-b_m block means, matching the type of response tested later.

A distribution defined by historical sampling. Let $\mathcal{B}_m$ contain all admissible consecutive width-b_m feature blocks in the proposal's frozen R_m. With history held fixed, sample a block uniformly from $\mathcal{B}_m$. The law of its feature mean is

$$Q_{H,m} = \frac{1}{|\mathcal{B}_m|} \sum_{H \in \mathcal{B}_m} \delta_{\bar{z}_H}.$$

Its response under proposal m has distribution $P_{\text{hist}}(\cdot \mid m) = (f_{w_m})_{\#}Q_{H,m}$, where the pushforward $(f_w)_{\#}Q$ is the distribution of $f_w(Z)$ when Z has law Q. This is Eq. (3).

The conditional interpretation can be made explicit without a model for future observations. At a fixed time s, hold the observed history and cached proposals fixed. In an auxiliary sampling experiment, select a proposal index Π with any positive probabilities over the finite eligible set, then select one of its historical blocks uniformly. For the resulting projected response Y and any measurable B,

$$\mathbb{P}_{\text{ref}}(Y \in B \mid \Pi = m, \mathcal{F}_s) = P_{\text{hist}}(B \mid m).$$

Each proposal uses its original historical snapshot, not all data available at s. The auxiliary proposal probabilities merely define a joint sampling law. They introduce no mixture weights into the detector. The detector evaluates eligible proposals and takes the maximum score. Conditioning here selects the historical sampling rule. It is not $\mathbb{P}_0(Y_{\text{future}} \in B \mid \mathcal{F}_s, \Pi = m)$ and does not assert that an unchanged future block follows this reference.

Finite empirical approximation. The implementation does not enumerate $\mathcal{B}_m$. It uses $\mathcal{H}_m = (H_{m,1}, \ldots, H_{m,J})$, with $J = 16$ evenly spaced starts. For any direction w, the estimated response distribution is

$$\widehat{P}_m^{\,w} = \frac{1}{J}\sum_{j=1}^{J} \delta_{f_w(\bar{z}_{H_{m,j}})}, \qquad f_w(z) = w^\top z, \tag{12}$$

where δ_a is a unit point mass at a. At $w = w_m$, this is $\widehat{P}_m$, the finite approximation of $P_{\text{hist}}(\cdot \mid m)$ used by the detector. Equivalently, it is the exact conditional distribution in the same auxiliary experiment when the historical draw is restricted to these J block entries. The grid is deterministic, not an iid sample. Blocks may overlap and repeated starts retain their multiplicities. Neither independence nor stationarity is required to define the estimate. We give no consistency or confidence bound for its approximation of the full historical law. In particular, a grid may miss the most extreme admissible block. Neither historical law is an estimate of the complete pre-change time-series law P_0.

A distribution functional, not a probability score. For a compactly supported probability measure F on the real line, let $\mu_F = \int y\, dF(y)$ and

$$\mathcal{T}_b(F) = \big(U(F), D_b(F)\big) = \left(\max \operatorname{supp}(F), \sqrt{\int (y - \mu_F)^2\, dF(y)} + \frac{1}{b}\right). \tag{13}$$

Then the main-text operator satisfies $\mathcal{R}_m(w) = \mathcal{T}_{b_m}(\widehat{P}_m^{\,w}) = (u_m(w), d_m(w))$. In particular, the standard deviation uses divisor J, not $J - 1$. The construction is

$$(\widehat{Q}_{R_m}, \widehat{Q}_{A_m}, \eta_m) \longrightarrow w_m, \qquad P_{\text{hist}}(\cdot \mid m) \longrightarrow \widehat{P}_m \xrightarrow{\mathcal{T}_{b_m}} (u_m, d_m).$$

The last two arrows are empirical approximation followed by summary evaluation, not density optimization. The gate η_m may also use ordered past-block statistics. It is not inferred from these two unordered empirical measures alone. The later response is the empirical expectation $v_{t,m} = \int f_{w_m}(z)\, d\widehat{Q}_{C_{t,m}}(z) = f_{w_m}(\bar{z}_{C_{t,m}})$, and the implemented score is exactly

$$T_{t,m} = (f_{w_m})_{\#}\widehat{Q}_{C_{t,m}}, \qquad q_{t,m} = \mathcal{S}_{b_m}(T_{t,m}, \widehat{P}_m).$$

For a later-response law T with mean $\mu(T) = \int y\, dT(y)$, define

$$\mathcal{S}_b(T, F) = \frac{[\mu(T) - U(F)]_+}{D_b(F)}. \tag{14}$$

Here $\mu(T_{t,m}) = v_{t,m}$. We cache $u_m = u_m(w_m)$ and $d_m = d_m(w_m)$ with the direction. These summaries are all that this score needs, but do not determine the entire reference distribution. No parametric density, smoothing bandwidth, or numerical CDF evaluation is required. The score is not a tail probability or an optimized estimate of a probability. For example, once v exceeds the largest empirical atom, the empirical upper-tail probability is zero for every such v, whereas $\mathcal{S}_b(\delta_v, F)$ still measures the magnitude of excess. Changing the proposal may change which historical block attains the maximum and the variability, even when $\mathcal{H}_m$ and ϕ_{L_m} are unchanged. Different directions need not give distinct references, for example under symmetry. This is a projection of the existing features, not a new invertible coordinate system or online retraining of ϕ_{L_m}.

Dependence on the direction. For fixed block means $h_j = \bar{z}_{H_{m,j}}$ with $\|h_j\|_2 \le 1$ and directions w, w', let $\rho = \|w' - w\|_2$. Then

$$|u_m(w') - u_m(w)| \le \rho, \qquad |d_m(w') - d_m(w)| \le \rho. \tag{15}$$

Indeed, coupling the response to each block under w with its response under w' gives $W_\infty(\widehat{P}_m^{\,w}, \widehat{P}_m^{\,w'}) \le \rho$ and $W_2(\widehat{P}_m^{\,w}, \widehat{P}_m^{\,w'}) \le \rho$. The endpoint and scale inequalities follow from Eqs. (18)–(19). The floor is unchanged. These deterministic bounds allow the maximizing historical block to change. They are not null-probability bounds. Under a common translation $h_j \mapsto h_j - a$, $u_m(w)$ shifts by $-w^\top a$, while $d_m(w)$ is unchanged. Translating the test features by the same a therefore leaves $q_{t,m}$ unchanged. The unit-ball bound applies to the features before the implementation's numerical translation.

In proofs for one comparison we omit the index m, writing $R, A, H_j, C, w, u, d, b, q$ for the corresponding indexed quantities. The cached u, d are empirical reference statistics, whereas h is the detector's final alarm threshold. The conditional sampling definition alone supplies neither selective inference after learning w nor false-alarm control for future data.

A.3 Availability and raw-time causality

We use zero-based half-open indices in this subsection to match the implementation. Counts of observations remain one-based. A lag-L feature z_i consumes raw observations $x_i, \dots, x_{i+L-1}$. A proposal at raw decision index d and feature-block width b has $k = d - b - L + 2$, reference feature indices $[0, k)$, and proposal indices $[k, k + b)$. Eligibility requires $k \ge 4b$. The proposal's last raw observation is $k+b+L-2 = d$. Its first test window begins at raw index $d+1$. For $a = 0, \dots, 8b-1$, the test contains feature starts $[d+1+a, d+1+a+b)$ and arrives at raw endpoint $d+a+b+L-1$. All test raw observations therefore arrive strictly after selection. The reference and proposal blocks may share raw observations.

At startup, decisions lie on the grid $d \in b\mathbb{N}_0$. The first eligible decision is

$$d_0 = b \left\lceil \frac{5b + L - 2}{b} \right\rceil, \qquad r_{\mathrm{grid}}(L, b) = d_0 + b + L + 1. \tag{16}$$

The second expression includes two consecutive raw endpoint scores. Thus $r_{\mathrm{grid}}(L, b) \ge 6b+2L-1$, with equality when $5b + L - 2$ is divisible by b. For $L = 8, b = 1$, the earliest persistent score is at count 21. An unavailable score is zero, not a missing evaluation record.

The mature kernel schedules proposals at feature starts $k \in \{4b, 5b, \dots\}$. Its first test ends at feature endpoint $6b + (L - 1) - 1$, or raw index $6b + 2L - 3$. That is $6b + 2L - 2$ observations. The next endpoint supplies the second persistence value, proving the count $r(L, b) = 6b + 2L - 1$. At $L = 64, b = 8$ this is 175. Subsequent proposals use the same width-spaced schedule and retain eight width-length future ranges. This count specifies when the score can be computed, not when a change will be detected.

All feature training parameters and deterministic selection rules are frozen. At a proposal decision, the conditional reference is a random probability measure whose atoms and weights are measurable from the observed history. Its endpoint and scale functionals are therefore measurable at that decision. The later response distribution uses only observations available at testing. Taking a maximum over active proposals and applying the minimum of the two most recent endpoint scores preserves adaptedness. Replacement at count 175 is deterministic. Consequently S_t is $\mathcal{F}_t$-measurable, and τ_h is a stopping time when h is fixed before the test sequence. This argument establishes causality, not independence or a martingale property.

A.4 Historical coverage and aligned change

We distinguish the full historical sampling law from its finite-grid estimate. For the fixed proposal m, let $F_\star = P_{\mathrm{hist}}(\cdot \mid m)$ be the pushforward of the uniform law over all admissible historical blocks, as defined in Appendix A.2. Let $F = \widehat{P}_m$ be its sampled block law. Since every sampled block is admissible, $\mathrm{supp}(F) \subseteq \mathrm{supp}(F_\star)$. Both supports are finite. Write their full-history endpoints as $\ell_\star, u_\star$, their range as $r_w = u_\star - \ell_\star$, and the upper-endpoint omission as $\varepsilon = u_\star - U(F) \ge 0$. These quantities depend on the frozen direction and width.

Proof of Proposition 1. If the recurring block has mean response $v_0 \in [\ell_\star, u_\star]$, then $v_0 - U(F) \le u_\star - U(F) = \varepsilon$. Taking the positive part and dividing by $d = D_b(F) > 0$ gives $0 \le q_{\mathrm{rec}} \le \varepsilon/d$. For a shifted continuation with feature mean $\bar{z}_0 + \delta$, where $w^\top \bar{z}_0 = v_0$, its response is $v_1 = v_0 + w^\top \delta = v_0 + \gamma$. Hence

$$v_1 - U(F) \ge \ell_\star + \gamma - (u_\star - \varepsilon) = \gamma - r_w + \varepsilon.$$

Monotonicity of the positive part gives the second bound. No independence between the proposal and its historical blocks is used. □

The recurring condition includes exact repetitions of historical blocks and convex combinations of their feature means. It is not asserted for every new realization of a stationary process: an unseen extreme can lie outside the observed range. The shifted condition describes a sustained change in feature means, not necessarily an additive shift of raw data. For a frequency change, for example, the relevant shift is in spectral features. It is a sufficient model of a detectable change, not a characterization of all changes in a time-series distribution. The full-history law here is not the unknown pre-change population law.

What estimating a direction contributes. The margin depends on projected signal γ relative to historical range r_w, both measured along the same estimated direction. A large recent contrast alone does not ensure a favorable margin. The following consequence makes direction accuracy explicit.

Corollary 1 (Alignment of a normalized proposal)**.** *Let $p = \eta \odot \Delta$ be the past proposal vector and suppose $\|p\|_2 \ge \epsilon > 0$, where ϵ is the normalization floor in Eq.* (2)*, so that $w = p/\|p\|_2$. If $\|p - \delta\|_2 \le a$, then*

$$w^\top \delta \ge \|\delta\|_2 - 2a, \qquad q_{\mathrm{new}} \ge \frac{[\|\delta\|_2 - 2a - r_w + \varepsilon]_+}{d}.$$

Proof. By Cauchy–Schwarz and the reverse triangle inequality, $w^\top \delta = \|p\|_2 + w^\top(\delta - p) \ge \|p\|_2 - a \ge \|\delta\|_2 - 2a$. Substitute in Proposition 1. □

Startup uses $\eta = \mathbf{1}$; the mature gate changes p. The corollary does not assert that either construction achieves small a. Proposal drift, inaccurate features, and a historical segment that already contains the change can all weaken alignment.

From a comparison to persistence. If a proposal remains eligible within the same scored stage at two consecutive evaluations and its lower bound exceeds h at both, then $V_{t-1} > h$ and $V_t > h$, so $S_t = \min(V_{t-1}, V_t) > h$. The same proposal is sufficient but not necessary, because different proposals may attain the two maxima. If no earlier alarm occurred, this gives an alarm by the second evaluation. It gives no bound on the time needed to construct an aligned proposal. Likewise, zero excess for one recurring comparison does not imply a zero pooled score unless every eligible comparison has zero excess. The coverage condition must therefore not be read as a false-alarm probability guarantee for the complete detector. In particular, missing an endpoint increases both the possible recurring excess and the shifted-score lower bound. Higher scores caused by this omission are not evidence of better separation.

A.5 Distribution metrics and score stability

Proposition 2 (Stability of the distributional score)**.** *Let $F, \widetilde{F}, T, \widetilde{T}$ be compactly supported probability laws on $\mathbb{R}$, with a common $b > 0$. Write $q = \mathcal{S}_b(T, F)$, $\widetilde{q} = \mathcal{S}_b(\widetilde{T}, \widetilde{F})$, $d = D_b(F)$, and $\widetilde{d} = D_b(\widetilde{F})$. For Wasserstein distances W_p, put $a = W_1(T, \widetilde{T}) + W_\infty(F, \widetilde{F})$ and $c = W_2(F, \widetilde{F})$. Then*

$$|\widetilde{q} - q| \le \min\left\{ \frac{a + qc}{\widetilde{d}}, \frac{a + \widetilde{q}c}{d} \right\}. \tag{17}$$

The bound controls changes in the later mean, historical endpoint, and historical scale. It does not assert that these errors are small. For a probability law F on $\mathbb{R}$, write $F(x) = F((-\infty, x])$ and

$F^{-1}(\beta) = \inf\{x : F(x) \geq \beta\}$, $0 < \beta \leq 1$. For compactly supported laws F, G and $1 \leq p < \infty$, define

$$W_p(F,G) = \inf_{\pi \in \mathcal{C}(F,G)} \left(\int |x-y|^p \, d\pi(x,y) \right)^{1/p}, \qquad W_\infty(F,G) = \inf_{\pi \in \mathcal{C}(F,G)} \operatorname{ess\,sup}_\pi |x-y|,$$

where $\mathcal{C}(F,G)$ is the set of couplings with these marginals. In one dimension the common-quantile coupling is optimal:

$$W_p(F,G)^p = \int_0^1 |F^{-1}(u) - G^{-1}(u)|^p \, du, \qquad W_\infty(F,G) = \operatorname{ess\,sup}_{0<u<1} |F^{-1}(u) - G^{-1}(u)|.$$

These are standard transport definitions, not an additional optimization performed by PARC (Panaretos & Zemel, 2019). The conditional historical distributions in Section 4 are compactly supported because the feature map and direction are bounded.

Mean, endpoint, and scale functionals. Any coupling $(Y, \widetilde{Y})$ of $T, \widetilde{T}$ gives $|\mu(T) - \mu(\widetilde{T})| \leq \mathbb{E}|Y - \widetilde{Y}|$. Taking the infimum proves the W_1 bound. For a coupling of $F, \widetilde{F}$ with $|Y - \widetilde{Y}| \leq r$ almost surely, $U(\widetilde{F}) \leq U(F) + r$ and the reverse inequality hold. Taking infima (or a sequence approaching the infimum) yields

$$|U(\widetilde{F}) - U(F)| \leq W_\infty(F, \widetilde{F}). \tag{18}$$

For any square-integrable coupling of $F, \widetilde{F}$, the reverse triangle inequality in L^2 and the contraction from centering imply

$$|\operatorname{sd}(\widetilde{Y}) - \operatorname{sd}(Y)| \leq \|(\widetilde{Y} - Y) - \mathbb{E}(\widetilde{Y} - Y)\|_2 \leq \|\widetilde{Y} - Y\|_2.$$

The common floor $1/b$ cancels. Infimizing over couplings proves

$$|D_b(\widetilde{F}) - D_b(F)| \leq W_2(F, \widetilde{F}). \tag{19}$$

Proof of Proposition 2. Put $N = [\mu(T) - U(F)]_+$ and $\widetilde{N} = [\mu(\widetilde{T}) - U(\widetilde{F})]_+$. The positive-part map is 1-Lipschitz, so $|\widetilde{N} - N| \leq a$. Both $d, \widetilde{d}$ are at least $1/b$ and $|\widetilde{d} - d| \leq c$. Adding and subtracting $N/\widetilde{d}$ gives

$$|\widetilde{q} - q| \leq \frac{|\widetilde{N} - N|}{\widetilde{d}} + N \left| \frac{1}{\widetilde{d}} - \frac{1}{d} \right| \leq \frac{a + qc}{\widetilde{d}}.$$

Adding and subtracting $\widetilde{N}/d$ instead gives $(a + \widetilde{q}c)/d$. Taking the smaller bound proves the result. □

The comparison may involve two empirical references, or a finite-block estimate and its full historical sampling law, provided their distances are evaluated or bounded. No random-sampling rate is asserted. In particular, the deterministic grid of 16 historical blocks does not supply an iid concentration theorem.

Why average approximation error is insufficient. Let $F = \delta_0$, $F_\varepsilon = (1-\varepsilon)\delta_0 + \varepsilon\delta_1$, and $T = \delta_1$, for $0 < \varepsilon < 1$. Then $W_p(F, F_\varepsilon) = \varepsilon^{1/p} \to 0$ for every finite p, but $\mathcal{S}_b(T, F) = b$ and $\mathcal{S}_b(T, F_\varepsilon) = 0$. A small amount of mass at an extreme response changes the endpoint by one. Thus the score has no continuity bound depending only on vanishing finite-p transport error over arbitrary bounded reference laws. This also describes a full historical law with one rare extreme block that a smaller grid misses. It does not assert that such a block is irrelevant. For two equal-weight empirical laws with the same J atoms, there is the finite-cardinality bound $W_\infty \leq \sqrt{J}\, W_2$ from the sorted coupling. The obstruction concerns a law-independent bound, not the absence of stability for fixed finite references.

A.6 Feature errors as a distributional corollary

For a fixed direction $\|w\|_2 \le 1$ and corresponding historical blocks, let $h_j, \widetilde{h}_j$ be block means under two feature maps of the same dimension, and let $c_0, \widetilde{c}_0$ be the later block means. The empirical references are $F = J^{-1}\sum_j \delta_{w^\top h_j}$ and $\widetilde{F} = J^{-1}\sum_j \delta_{w^\top \widetilde{h}_j}$. Pairing corresponding atoms defines a coupling, without an independence assumption. With $e_j = \|\widetilde{h}_j - h_j\|_2$ and $e_C = \|\widetilde{c}_0 - c_0\|_2$,

$$W_\infty(F, \widetilde{F}) \le \max_j e_j, \qquad W_2(F, \widetilde{F}) \le E := \left(J^{-1} \sum_j e_j^2 \right)^{1/2}.$$

Only the later mean enters the score. Replacing its response distribution by $T = \delta_{w^\top c_0}$, and similarly for $\widetilde{T}$, leaves both scores unchanged and gives $W_1(T, \widetilde{T}) \le e_C$. Proposition 2 therefore yields

$$|\widetilde{q} - q| \le \min\left\{ \frac{A + qE}{\widetilde{d}}, \frac{A + \widetilde{q}E}{d} \right\}, \qquad A = e_C + \max_j e_j. \tag{20}$$

With unit-ball feature means, $q, \widetilde{q} \le 2b$, so the looser bound $bA + 2b^2 E$ holds. A common translation of the historical and later features leaves the score unchanged. The unit-ball bound applies before the implementation's numerical translation. The direction may have been selected from approximate past features, but this corollary holds it fixed across the two scores. Synthetic training MSE alone does not establish small distribution distances on real sequences.

A.7 Changing directions, contexts, and alarm decisions

Proposition 2 compares distributions, regardless of how their directions were obtained. The feature corollary holds a direction fixed. Two runs may also select different contexts. Here both runs use the same proposal times, candidate contexts, widths, lifetimes, and transition rule, but their feature maps and learned directions may differ. The analysis compares two specified feature maps. Neither is assumed to be an ideal representation.

For one candidate, write $v = w^\top \bar{z}_C$ and $p_j = w^\top \bar{z}_{H_j}$, with hats denoting the other run, including its own direction. Put

$$A_* = |\widehat{v} - v| + \max_j |\widehat{p}_j - p_j|, \qquad E_* = \big(J^{-1}\sum_j (\widehat{p}_j - p_j)^2\big)^{1/2}.$$

Let $F = J^{-1}\sum_j \delta_{p_j}$ and $\widehat{F} = J^{-1}\sum_j \delta_{\widehat{p}_j}$. Corresponding atoms give a coupling of the two reference distributions. Using $T = \delta_v$ and $\widehat{T} = \delta_{\widehat{v}}$ is score-equivalent to keeping the later response distributions, since only their means enter. Their transport distances are bounded by A_*, E_*, so Proposition 2 gives

$$|\widehat{q} - q| \le e_q = \min\{(A_* + qE_*)/\widehat{d}, (A_* + \widehat{q}E_*)/d\}. \tag{21}$$

Indeed, the numerator difference is at most A_* and the denominator difference at most E_*. The same positive floor makes both divisions valid. For directions of norm at most one in a common feature space, $|\widehat{w}^\top \widehat{z} - w^\top \bar{z}| \le \|\widehat{z} - \bar{z}\|_2 + \|\widehat{w} - w\|_2 \|\bar{z}\|_2$. Changes in the learned direction contribute additional score error.

Corollary 2 (Selection-aware comparison stability)**.** *At one proposal decision and width, let $a_j, \widehat{a}_j$ be the two context-selection scores computed from past observations and let $\epsilon_j \ge |\widehat{a}_j - a_j|$. Write $i = \arg\max_j a_j$ and $\widehat{i} = \arg\max_j \widehat{a}_j$, with the same fixed tie-breaking, and define*

$$\mathcal{U} = \{j : a_j + \epsilon_j \ge \max_l (a_l - \epsilon_l)\}. \tag{22}$$

At each raw endpoint t, set a candidate score to zero before its test arrives or after its lifetime expires. If $e_j(t)$ bounds its paired score difference by Eq. (21)*, then*

$$|\widehat{q}_{\widehat{i}}(t) - q_i(t)| \le B(t) := \max_{j \in \mathcal{U}} \{|q_j(t) - q_i(t)| + e_j(t)\}. \tag{23}$$

Proof. The context selected in the second run satisfies $a_{\widehat{i}} + \epsilon_{\widehat{i}} \geq \widehat{a}_{\widehat{i}} = \max_l \widehat{a}_l \geq \max_l (a_l - \epsilon_l)$, so $\widehat{i} \in \mathcal{U}$, including ties. Insert $q_{\widehat{i}}(t)$ between the two selected scores, apply the triangle inequality and the candidate bound, and maximize over $\mathcal{U}$. Zero extension aligns different context arrival times on the raw-time axis without using future information to alter a proposal decision. □

Each candidate score here is the functional $\mathcal{S}_{b_m}(T_{t,m}, F_m)$; the corollary propagates distributional errors through selection of the candidate. If $\mathcal{U} = \{i\}$, the cross-context term vanishes. No uniform margin assumption is required. Near ties can change the selected context: selection scores $(1/2, 1/2 - \varepsilon)$ and $(1/2, 1/2 + \varepsilon)$ select different contexts for arbitrarily small ε, while their future scores can be $(0, 1)$. Small errors in past selection scores alone therefore need not imply small selected-score errors.

Take the maximum of $B(t)$ over all proposal slots to obtain B_t. The inequalities $|\max a - \max b| \leq \|a-b\|_\infty$ and $|\min a - \min b| \leq \|a-b\|_\infty$ give $|\widehat{V}_t - V_t| \leq B_t$ and $|\widehat{S}_t - S_t| \leq e_t := \max(B_{t-1}, B_t)$ within the active stage. Both persistence values use the same stage at replacement. Consequently, neither proposal pooling nor persistence multiplies this error by the number of comparisons. This includes the fixed-proposal replay as a special case.

For paired normal units, let $e_{\mathrm{ref}} = \max_{i,t} e_{i,t}$. Their maxima differ by at most e_{ref}. Match the two empirical laws of trajectory maxima by unit identity. Their W_∞ distance is at most e_{ref}. For a common finite order statistic $1 \leq k \leq n$, the same lower quantile of these laws therefore differs by at most e_{ref}, including ties, giving $|\widehat{h} - h| \leq e_{\mathrm{ref}}$ and hence

$$|(\widehat{S}_t - \widehat{h}) - (S_t - h)| \leq e_t + e_{\mathrm{ref}}. \tag{24}$$

Crossing indicators agree wherever $|S_t - h| > e_t + e_{\mathrm{ref}}$; if this holds through the first crossing, the first alarms agree. A shared fixed threshold has $e_{\mathrm{ref}} = 0$. Thresholds selected using change annotations need not satisfy the order-statistic bound. When $k > n$, both rank thresholds are infinite and neither detector alarms; the finite-margin inequality is not used. These are deterministic stability statements, not probabilistic detection-delay or learned-feature generalization guarantees.

A.8 Startup distributions and threshold cost

Proposition 3 (Startup cost as a quantile shift)**.** *Let Γ be any joint law of bounded nonnegative startup and mature maxima (A, B). Denote the laws of $\max(A, B)$ and B by Λ_{full} and Λ_{mat}. For $0 < \beta \leq 1$, let $h = \Lambda_{\mathrm{full}}^{-1}(\beta)$ and $h_M = \Lambda_{\mathrm{mat}}^{-1}(\beta)$ be lower quantiles. With $c_\Gamma = \operatorname{ess\,sup}_\Gamma (A - B)_+$,*

$$0 \leq h - h_M \leq c_\Gamma. \tag{25}$$

Fix a horizon and the detector design. A normal sequence induces startup and mature maxima (A, B), taking an empty stage maximum as zero. Their joint law is Γ. No independence of the two stages is assumed. The complete and mature trajectory-maximum laws are the pushforwards $\Lambda_{\mathrm{full}} = (\max)_\# \Gamma$ and $\Lambda_{\mathrm{mat}} = (\mathrm{pr}_2)_\# \Gamma$. The proposition also applies to an empirical joint law or any bounded joint law. Normality is needed for its interpretation as a normal threshold, not for the inequality.

Proof. Under the joint law Γ, let $M = \max(A, B)$. Then $B \leq M \leq B + c_\Gamma$ almost surely. Hence, for every real x,

$$\Lambda_{\mathrm{mat}}(x - c_\Gamma) \leq \Lambda_{\mathrm{full}}(x) \leq \Lambda_{\mathrm{mat}}(x).$$

Taking lower quantiles gives $\Lambda_{\mathrm{mat}}^{-1}(\beta) \leq \Lambda_{\mathrm{full}}^{-1}(\beta) \leq \Lambda_{\mathrm{mat}}^{-1}(\beta) + c_\Gamma$ for $0 < \beta \leq 1$. Compact support ensures the endpoint quantiles are finite. This proves Eq. (25). □

For the empirical joint law $\widehat{\Gamma}_n = n^{-1} \sum_i \delta_{(A_i, B_i)}$, $c_{\widehat{\Gamma}_n} = \max_i [A_i - B_i]_+$. For $1 \leq k \leq n$, the quantile at $\beta = k/n$ is exactly the kth order statistic, including ties.

For this empirical law, $h = h_M$ if and only if at least k full maxima are at most h_M. With identical mature trajectories and survival through startup, the full detector cannot cross its larger threshold earlier than mature-only. A mature score greater than $h_M + c_{\widehat{\Gamma}_n}$ also exceeds h. These are deterministic consequences for the specified trajectories, not a bound on detection probability or expected delay.

A.9 TRAJECTORY DISTRIBUTIONS AND FINITE-HORIZON VALIDITY

For fixed design $\mathcal{D}$, the complete causal detector induces the map $M_H : X \mapsto \max_{t \le H} S_t(X)$ and the normal score distribution $\Lambda_H = \mathcal{L}_0(M_H(X) \mid \mathcal{D})$, the pushforward of the conditional normal-unit law under this map. This is a distribution of whole-sequence maxima, distinct from every internal proposal-conditioned historical law.

Theorem 1 (Empirical-quantile split-rank control)**.** *Let $n \ge 1$ and $0 < \alpha < 1$. Conditional on a disjoint design collection $\mathcal{D}$, suppose normal units $X_1, \dots, X_{n+1}$, each observed through the same horizon H, are exchangeable. Fix H, preprocessing rule, network parameters, random seed, startup rule, and stopping-score procedure using $\mathcal{D}$. Apply this same procedure to each unit to obtain $M_i = M_H(X_i)$. Estimate Λ_H by $\widehat{\Lambda}_{H,n} = n^{-1} \sum_{i=1}^{n} \delta_{M_i}$. For $k = \lceil (n+1)(1-\alpha) \rceil \le n$, set $h = \widehat{\Lambda}^{-1}_{H,n}(k/n)$. With strict crossing,*

$$\mathbb{P}(M_{n+1} > h \mid \mathcal{D}) \le \frac{n+1-k}{n+1} \le \alpha. \tag{26}$$

If $k > n$, use $h = \infty$, and the probability is zero.

Proof. The measurable pushforward M_H preserves conditional exchangeability. The empirical lower quantile at k/n is the kth reference order statistic. Break ties using independent continuous random numbers. The test unit has uniform rank among the $n+1$ tie-broken maxima. Strictly exceeding the kth reference maximum implies that this rank is greater than k. There are $n+1-k$ such ranks. Without ties the events coincide. Ties can only make strict exceedance more conservative. Finally, $\{\tau_h \le H\} = \{M_{n+1} > h\}$ by construction. The infinite threshold case is immediate. □

The proof applies the standard exchangeable-rank argument (Shafer & Vovk, 2008; Angelopoulos & Bates, 2021). Freezing the procedure does not freeze the online directions across units: each unit constructs its own proposals and references using the same rule. At $\alpha = .05$, at least 19 normal units are required for a finite deterministic threshold. The result averages over reference and test units, conditional only on the separate design. It need not hold conditional on a particular realized threshold collection.

Unknown within-unit dependence is allowed, but between-unit sampling must justify exchangeability. Multiple overlapping crops from the same subject cannot be counted as independent subjects. If units contain several recordings, grouping and the maximum across recordings must be defined before threshold fitting. The guarantee is for the declared horizon and one first alarm. Repeated restarts or an unbounded horizon require additional analysis. No finite collection of empirical checks proves exchangeability of future real units.

B EXACT ARCHITECTURE AND TRAINING

B.1 FEATURES AND SYNTHETIC TRAINING

For one channel, the mature descriptor concatenates 16 square-root spectral band proportions, $\tanh(\bar{x}/3)$, and $\tanh(\log \max(s_x, 10^{-6}))$, then divides by $\sqrt{3}$. It uses the non-DC real-FFT powers of the centered 64-point window, split into 16 bands; their total is normalized with a 10^{-12} denominator guard. Multichannel descriptors are concatenated and divided by $\sqrt{d}$. The spectral norm contribution is at most $1/3$ and each scalar moment contributes at most $1/3$, proving the unit-ball bound.

Each short-window student has widths $(L+2)$–32–32–18 and tanh hidden activations. Inputs contain $\tanh((x - \bar{x})/(3 \max(s_x, 10^{-6})))$, $\tanh(\bar{x}/3)$, and $\tanh(\log \max(s_x, 10^{-6}))$. The first 16 output logits pass through softmax followed by a square root. The last two use tanh. The common $1/\sqrt{3}$ scaling bounds the output. In the deployed feature, the learned moment outputs are replaced by observed moments. Only their two coordinates are multiplied by the fixed validation factor ρ_L.

We use 1,200 synthetic training templates and 300 independent validation templates, with 32 windows per template. Harmonic signals have random periods between 8 and 512, random phases, amplitudes, and noise; additional templates are autoregressive signals. Random affine transformations and some

location/ scale changes expand the training distribution. The teacher is the native 64-point descriptor ending at the same raw time as the student's input. The student sees only its L observed points. Training minimizes mean squared error over all 18 descriptor coordinates, using Adam with learning rate .002, batch size 256, 1,200 steps, and gradient-norm clipping at five. We check validation error every 100 steps and select the optimizer seed by average synthetic validation MSE across the three contexts, before real archive evaluation. No real sequence or change annotation trains these weights.

Table 3: Primary short-window networks. Validation errors concern synthetic descriptor prediction, not detection accuracy.

Context	Parameters	Validation MSE	ρ_L
8	2002	0.00791	0.555
16	2258	0.00654	0.632
32	2770	0.00439	0.753

The factors are $\rho_L = [1 - \mathrm{MSE}_L/\mathrm{MSE}_{\mathrm{constant}}]_{[0,1]}$, approximately .555, .632, and .753. They are heuristics derived from synthetic prediction quality, not probabilities. The native short control uses a centered short window, zero-padded to 64 for its spectrum, with identical observed moments and factors. Thus it removes the network without changing the score or context selector.

The mature coordinate gate is a shared 5–8–1 tanh/sigmoid MLP with 57 parameters. Its five inputs summarize proposal magnitude, disagreement of the two reference halves, reference variance, proposal variance, and width. Specifically, with $\Delta = \bar{z}_A - \bar{z}_R$, reference-half difference v, coordinate variances s_R^2, s_A^2, and width b, the inputs are

$$\frac{1}{6}\,\mathrm{clip}_{[-12,6]}\left(\log(|\Delta| + 10^{-4}), \log(v^2 + 10^{-5}), \log(2s_R^2 + 10^{-5}), \log(s_A^2 + 10^{-5}), \log b\right). \quad (27)$$

Here v is the difference of the two reference-half means divided by $\sqrt{2}$. The gate learns to shrink a noisy contrast toward the full-cycle mean difference on synthetic periodic templates. It uses 1,200 training and 300 validation templates, 24 sampled contrasts per template, and Adam with learning rate .003, batch size 256, and 1,600 steps. Its loss is $\|\eta_\theta \odot \Delta - \Delta_{\mathrm{full}}\|_2^2$ averaged over coordinates and examples. Template periods range from 32 to 768, with partial-cycle comparisons and amplitude, phase, shape, and noise changes. The deployed 18-coordinate descriptor is also used in gate training. Startup uses $\eta = \mathbf{1}$. The learned gate is used only in the mature comparison.

B.2 Procedure and numerical conventions

Each comparison stores feature weights, not an online-trained classifier. During startup, the weights are the normalized recent-minus-historical feature difference. In the mature stage, the frozen gate modifies this difference before normalization. For a later block C, the untruncated excess is equivalently

$$w^\top(\bar{z}_C - \bar{z}_R) - \max_j w^\top(\bar{z}_{H_j} - \bar{z}_R) = w^\top \bar{z}_C - \max_j w^\top \bar{z}_{H_j}.$$

The historical average helps estimate w but cancels from this comparison. Thus a positive projected difference from that average is insufficient: the later projection must exceed every sampled historical projection. Only the sampled blocks are compared, not every possible historical window.

Here $\mathcal{M}_t^{\mathrm{short}}$ is the startup active set. Every test block contains only raw observations after its proposal's decision. The online algorithm maintains the mature history before replacement. It does not initialize the mature detector from an empty state at count 175. A new proposal can change the active comparisons but never changes an existing comparison's weights or historical statistics. Cached comparisons expire after $8b$ permitted test starts. The maximum is over complete comparison scores, not individual feature magnitudes. The empirical distribution is represented through its cached endpoint and scale for scoring. The algorithm need not store all its atoms after construction. Empty maxima are zero. Ties in context selection use the ascending fixed context order. Historical block starts in feature indices are $\lfloor (k-b)j/15 \rfloor$, $j = 0, \ldots, 15$, and the standard deviation uses divisor 16. Context-specific features are shifted by their first vector before cumulative sums. Contrasts are unchanged. Frozen-network inference uses deterministic row-wise sums to prevent batch-size changes from altering constant-input scores.

Algorithm 1 PARC with empirical proposal-conditioned references

Require: Frozen feature maps and gate, fixed preprocessing rule, threshold h, startup contexts $\{8, 16, 32, 64\}$, $J = 16$, replacement count $r = 175$

Initialize history, proposal caches, and zero endpoint scores
for each newly observed x_t **do**
  Apply the fixed causal preprocessing and update available features
  Update mature proposals and later-data scores using native features and the same reference functional $\mathcal{T}_b$
  if $t < r$ **then**
    for each due dyadic-width startup decision **do**
      for each eligible context L **do**
        Form candidate m with past $R_m, A_m, \mathcal{H}_m$ using Appendix A.3
        Propose w_m from $\widehat{Q}_{A_m}, \widehat{Q}_{R_m}, \eta_m$ by Eq. (2)
        Estimate $\widehat{P}_m$ from $\mathcal{H}_m$; compute $(u_m, d_m) = \mathcal{T}_{b_m}(\widehat{P}_m)$
        Compute the context-selection score $(f_{w_m}(\bar{z}_{A_m}) - u_m)/d_m$
      end for
      Retain the maximizing context and cache its fixed (w_m, u_m, d_m) and lifetime
    end for
    Remove expired comparisons; score completed later blocks $C_{t,m}$ by Eq. (6)
    Set $V_t^{\text{short}} = \max_{m \in \mathcal{M}_t^{\text{short}}} q_{t,m}$ (zero if empty)
    Set $S_t = \min(V_{t-1}^{\text{short}}, V_t^{\text{short}})$
  else
    Use native $L = 64$ features, widths $8, 16, \ldots$, and gated directions
    Read the maintained mature endpoint scores (do not restart at r)
    Set $S_t = \min(V_{t-1}^{\text{mature}}, V_t^{\text{mature}})$
  end if
  if $S_t > h$ **then**
    return first alarm t
  end if
end for

The raw preprocessing option leaves values unchanged. The frozen-eight option subtracts the first-eight coordinate median and divides by 1.4826 times their median absolute deviation, with standard-deviation and then unit-scale fallbacks. Its transform never updates after observation eight. This is not a claim that those observations are normal.

C Archive protocol and complete results

C.1 Datasets and target intervals

TSSB contains 75 benchmark sequences constructed from classification-series families. We retain all 69 records with change annotations and six with no target annotation. For a changed record, the first annotation is the target and the next annotation, if present, closes its evaluation interval. Construction labels identify segment boundaries. They are not response labels observed by the detector. This is first-alarm evaluation on a segmentation collection, not a claim to perform supervised classification or complete segmentation (Ermshaus et al., 2023).

The TCPD collection used here contains 42 records. Forty have targets and two do not. Targets follow the stored inventory rule: take each nonempty annotator's first change and round their median to an integer. The evaluation interval ends at the record end. This collection contains 42 rather than the 37 series in the original benchmark paper (Van den Burg & Williams, 2020). The released record inventory makes the evaluated set explicit. No method receives a target-derived warmup, period, or final-length fraction.

C.2 Fitting and nested selection

Five outer folds keep source families intact. Within outer training families, three inner folds choose a configuration and one early-alarm budget from $\{.05, .10, .20\}$. Candidate thresholds are zero, infinity, and observed pre-target maxima of training changed records or complete maxima of training null records. A threshold maximizes family-weighted detection subject to the selected early budget and zero training-null alarms. Ties favor smaller capped loss, then a larger threshold.

Inner validation uses

$$U = \text{Detect} - 2\,\text{Early} - \text{Null} - .1\,\text{CappedLoss}, \tag{28}$$

with equal weight per source family within changed/null categories. Mean inner utility selects the candidate. Deterministic ties use smaller budget and then configuration name. The selected threshold is refitted on all outer training families before scoring the held-out families. Five distinct outer partitions are used; repeats reuse the same records and are averaged, not counted as new independent observations.

The primary method selects only raw versus frozen-eight preprocessing and the threshold budget within this protocol. Its architecture and network weights are frozen. Baselines select among their declared implementation configurations under the same criterion. Both archives use the same applicable short-initialization options, with additional original configurations retained where supported. The full per-method candidate grids are included with the artifact. These alternatives are selected within training folds, not after examining the held-out record. Both archives informed architecture development. Family-level splitting does not remove the effects of those earlier design decisions.

C.3 Baseline definitions

Energy, RBF–MMD, KS, mean, and variance use distributional/window scores. C2ST–MLP is a sequential classifier comparison. Online RFF–MMD and NEWMA include their available original streaming recurrences. Hahn CUSUM and Gaussian MdFOCuS include native-start implementations. QT–EWMA uses a QuantTree-based distribution comparison; PM-CuSum (M) uses the Gaussian-mean predictive-mixture model. Bibliographic sources are given in Section 2. These twelve scores, not every method cited in the related work, define Table 1. In particular, NN-CUSUM, contrastive online CPD, Scan-B, kernel density-ratio detection, E-C2ST, CALM, WATCH, and Bayesian online CPD are discussed but not evaluated here. The empirical conclusions apply only to the evaluated implementations.

The grid varies raw or spectral inputs, fixed reference support, window width, available original recurrence choices, score normalization, and persistence. It includes reference sizes 8, 16, and 32 as well as 64, 128, and 256; shorter and native-start configurations are eligible where the implementation supports them. Configuration-specific availability is preserved. No score is granted a supplied change-free reference. Label-assisted threshold selection makes these matched benchmark operating points, not the original papers' nominal ARL or anytime guarantees.

Initialization preserves the method-specific distinction. Fixed-reference window scores estimate their reference from the declared initial observations, which may already contain a change. C2ST uses its declared reference/training support before scoring later data. NEWMA and original online RFF options initialize their streaming recurrences from the observed sequence; native Hahn and MdFOCuS options retain their own start rules. QuantTree options need enough observations to construct their reference partition, and PM-CuSum retains its predictive-model update. Unavailable scores are zero, not removed records. An initial change can contaminate a baseline's reference estimate. All methods use the same evaluation and fitting protocol, but their initialization requirements differ.

C.4 Ranks and cross-archive gains

For ranks, we first average repeated predictions within record, then records within each source family. Changed-record capped loss defines the primary rank matrix. Friedman and Nemenyi calculations use that matrix, following the multiple-dataset comparison framework of Demšar (2006). Paired Wilcoxon tests compare PARC with the baselines using the step-down correction of Holm (1979). They are descriptive calculations on development archives: their reference distributions do not account

Table 4: Mean counts and successful-only median delay in raw samples, averaged over five partitions. D, E, and N denote detected changes, early alarms, and null alarms. Capped loss is in Table 1. A small conditional delay does not imply high coverage.

	TSSB				TCPD			
Method	D	E	N	Delay	D	E	N	Delay
PARC	41.4	2.0	1.0	232.6	20.4	1.6	0.0	47.5
C2ST-MLP	30.0	5.8	0.6	334.2	10.6	4.6	1.2	125.7
Energy	25.6	6.6	0.0	82.4	13.8	5.4	0.6	60.2
Hahn CUSUM	31.6	4.0	0.2	269.1	11.2	3.6	1.4	96.7
KS	26.8	8.4	0.0	104.2	12.6	3.2	0.4	90.8
MdFOCuS	30.2	2.4	0.2	261.0	14.6	5.6	0.6	81.8
Mean	20.4	5.0	0.6	94.3	11.2	6.2	1.0	67.5
NEWMA	13.6	10.0	1.8	369.7	14.8	2.6	0.6	73.7
Online RFF-MMD	33.4	7.4	0.4	189.4	13.8	4.0	0.6	40.7
PM-CuSum (M)	22.8	5.0	1.0	471.8	15.8	4.8	0.8	56.7
QT-EWMA	13.2	8.2	1.0	197.1	10.8	4.0	1.2	117.1
RBF-MMD	22.6	6.4	0.0	109.2	13.6	6.0	1.2	72.9
Variance	17.0	6.2	1.8	148.8	13.8	4.4	1.4	82.6

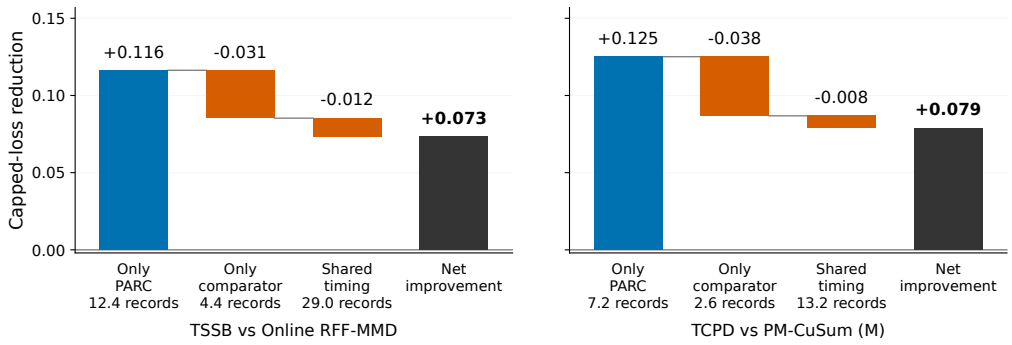

Figure 6: Capped-loss reduction against each archive's lowest-loss comparator. Additional detections, lost detections, and shared-detection timing sum to the net gain. Counts are averaged over five partitions. Positive values favor PARC. Joint misses contribute zero and are omitted.

for historical architecture search or dependence introduced by overlapping model-training folds. A sensitivity analysis includes null records with loss equal to their alarm indicator.

Figure 6 decomposes the comparison with the lowest-loss baseline in each archive. On TSSB, PARC detects 12.4 changes missed by Online RFF-MMD, while losing 4.4 detections that comparator makes. On TCPD, PARC detects 7.2 changes missed by PM-CuSum (M), while losing 2.6 detections that comparator makes. The improvement comes from detecting additional changes. Delay on changes detected by both methods is slightly worse on both archives. Component effects are examined separately in the ablation study.

Figure 6 decomposes the paired capped-loss reduction against the lowest-loss comparator on each archive. On changed record/repeat pairs, let $d = \ell_{\text{comparator}} - \ell_{\text{PARC}}$. Partition the pairs into only PARC detecting, only the comparator detecting, both detecting, and neither detecting. Each plotted contribution is the sum of d in that category divided by the total number of changed record/repeat pairs, not by the category size. The four contributions therefore sum to the aggregate loss reduction. The last category is zero because both losses are one. Null records are not part of this changed-record timing measure. Comparator choice is descriptive and does not alter any saved model, threshold, or prediction.

C.5 ILLUSTRATIVE TRAJECTORY

Figure 2 uses the quality-control record from one held-out TCPD partition. We selected the example for visual clarity from nine successful startup trajectories, not to estimate aggregate performance. The illustrated proposal uses $L = 32, b = 8$, selected at raw index 112; its test starts at 116 and the limiting persistence endpoint is 154. All indices here are zero-based. The direction bars share feature rows with the heatmap but have a weight axis, not a time axis. Historical and test projections use the same centering and scale. The final panel contains all active proposal scores, not only the illustrated direction: 224 proposals appear over the displayed startup and 35 are active at the alarm at 155. Their pointwise maximum and persistence reproduce the saved score. The original held-out threshold is unchanged. Values after the first alarm show continued scoring for illustration. Startup curves end at the replacement at index 174.

C.6 TRAINING-SEED AND COMPONENT CHECKS

Table 5: All original short-network seeds and gated/ungated startup controls. Each row uses the same mature detector and its own training-selected threshold.

Archive	Seed	Startup	Detect	Early	Null	Loss
TSSB	99431	gated	41.4	2.0	1.0	0.590
TSSB	99431	ungated	41.4	2.0	1.0	0.590
TSSB	99441	gated	40.6	2.2	1.2	0.599
TSSB	99441	ungated	40.6	2.2	1.2	0.599
TSSB	99451	gated	40.8	2.4	1.0	0.597
TSSB	99451	ungated	40.8	2.4	1.0	0.597
TCPD	99431	gated	19.2	2.0	0.0	0.678
TCPD	99431	ungated	20.4	1.6	0.0	0.653
TCPD	99441	gated	15.4	2.0	0.4	0.746
TCPD	99441	ungated	16.4	2.2	0.4	0.721
TCPD	99451	gated	17.6	2.8	0.0	0.696
TCPD	99451	ungated	18.8	2.8	0.0	0.673

The chosen network seed minimizes synthetic descriptor validation error, not archive detection loss. Its real-data advantage is nevertheless seed-dependent. The fixed-64 startup control uses only the native long context before the same replacement; learned-indefinitely keeps the context selector active throughout the sequence. Neither is counted as an additional baseline method in the twelve-score comparison.

C.7 CONTROLLED DESIGN SENSITIVITY

We evaluate 17 fixed configurations on every TSSB and TCPD record, using the same five family partitions. Raw and frozen-eight preprocessing and budgets $.05, .10, .20$ are evaluated separately, without selecting among these results. Table 6 fixes raw preprocessing and the $.05$ training early-alarm budget. Each score receives its own threshold fitted on the training families. This controlled protocol differs from Table 1, where inner folds select preprocessing and budget. A common fitting budget does not imply equal observed test alarm rates.

The study varies historical-block count $J \in \{8, 16, 32\}$, the floor γ/b with $\gamma \in \{.5, 1, 2\}$, persistence in $\{1, 2, 3\}$, lifetime in $\{4b, 8b, 16b\}$, and transition counts $175, 256, 384$. At 175 observations, the mature detector can first compute its persistent score. Switching earlier would replace startup by an unavailable score. The later values measure the effect of retaining startup comparisons longer. They improve TCPD loss but substantially worsen TSSB loss. Increasing proposal lifetime also has different effects on the two archives. Persistence changes are comparatively small, whereas the historical floor and block count can materially change outcomes. The required observation count therefore defines a reproducible transition, although its performance depends on the data distribution.

The fixed random direction and strongest-coordinate controls test direction learning without adding another detector. Replacing the historical maximum by a mean raises TSSB loss from .590 to .627 at this operating point, but the corresponding paired-family interval includes zero. The observed

Table 6: Controlled sensitivity. Detection, early-alarm and null-alarm columns are percentages, with the same denominators as Table 1. Loss is capped delay. Rows change one specified component of the fixed primary model. Thresholds are refitted by the same fixed training-budget rule.

	TSSB				TCPD			
Variant	Detect	Early	Null	Loss	Detect	Early	Null	Loss
PARC	**60.0**	**2.9**	**16.7**	**0.590**	**52.0**	**3.0**	**0.0**	**0.647**
Mean test reference	53.6	4.3	0.0	0.627	47.5	5.0	50.0	0.671
Mean selection/test	53.6	4.3	0.0	0.627	47.5	5.0	50.0	0.666
Fixed random direction	9.9	4.6	0.0	0.922	25.0	5.0	10.0	0.847
Strongest coordinate	35.9	4.3	16.7	0.769	50.0	5.0	50.0	0.647
Native short features	60.0	2.9	16.7	0.590	44.5	7.5	30.0	0.715
Fixed context 64	60.0	2.9	16.7	0.590	41.0	5.0	80.0	0.762
History 8	28.4	2.0	16.7	0.823	49.5	3.5	0.0	0.667
History 32	53.6	2.9	16.7	0.639	52.0	3.0	0.0	0.645
Floor 0.5	46.4	5.5	0.0	0.705	52.0	3.0	0.0	0.647
Floor 2	52.5	6.1	0.0	0.628	50.0	5.0	0.0	0.660
Persistence 1	60.0	2.9	16.7	0.590	52.0	3.0	0.0	0.645
Persistence 3	60.0	2.9	16.7	0.591	52.0	3.0	0.0	0.649
Lifetime 4	58.6	5.2	0.0	0.608	49.5	3.0	0.0	0.669
Lifetime 16	28.7	1.7	16.7	0.791	52.0	3.0	0.0	0.642
Retire 256	47.0	4.3	10.0	0.695	55.0	5.5	10.0	0.602
Retire 384	31.0	6.7	0.0	0.792	57.5	5.0	30.0	0.573

difference favors the maximum on TSSB, but the interval does not establish a statistically significant improvement. All alternative preprocessing/budget results are retained in the artifact. These sensitivity results were not used to select a new primary model.

D Numerical validation of the theory

We use simulation to examine the mathematical statements in Section 5, not to measure detection power or the benefit of individual components. The studies distinguish deterministic inequality checks from repeated sampling experiments. Neither replaces a proof or establishes exchangeability of future real sequences. Archive ablations remain in Section C.6.

D.1 Recurring patterns and sustained shifts

We check Proposition 1 on bounded four-dimensional feature sequences of length 128 or 512. This isolates the score's geometry from feature training. For each length, we draw 100 fresh 32-point motifs by circularly smoothing Gaussian noise with weights $.8^j$, $j = 0, \ldots, 15$, applying $\tanh$, centering each coordinate, and scaling the largest vector norm to .2. Repeating a motif creates dependent historical observations. History occupies the first $N - 64$ points, followed by a 32-point proposal block and a 32-point later block. Every scored block has width eight.

The proposal adds a fixed feature shift to the recent motif. We estimate w by normalizing its contrast with the historical mean, then freeze it before testing. Paired continuations either repeat the motif or add a sustained feature shift of norm $.04$, $.16$, or $.32$. Proposal and later shifts make angles $0°$, $60°$, $90°$, or $180°$. This explicitly tests alignment and its failure, rather than assuming that an estimated proposal always predicts the later change. Each comparison uses either all admissible historical blocks or a 16-start evenly spaced grid. We score every sliding later block and each pair of consecutive scores. Seed 20260925 fixes all draws and settings.

All 4,800 paired cases satisfy the recurrence, shifted-score, and persistence inequalities. With all historical blocks, none of the 2,400 recurring continuations has positive excess. In the aligned group, 349 of 600 shifted continuations have positive excess at every evaluation, and the sufficient lower bound certifies 342. Orthogonal and reversed directions certify none. The finite grid can miss a historical maximum: 1,026 of its 2,400 recurring continuations have some positive excess, always within ε/d. These scores are not false-alarm rates, because no statistical alarm threshold is fitted

Table 7: Short-sequence checks of Proposition 1. Each row has 600 paired recurring/shifted continuations, pooling lengths 128 and 512. Entries are counts, not detector accuracy. "Recurring > 0" counts any positive excess; "sustained > 0" requires positive excess at every later evaluation. "Certified" counts strictly positive lower bounds. All stated inequalities have zero violations.

Historical reference	Angle	Recurring > 0	Sustained > 0	Certified
All blocks	0°	0	349	342
All blocks	60°	0	167	159
All blocks	90°	0	0	0
All blocks	180°	0	0	0
16-block grid	0°	270	352	346
16-block grid	60°	252	181	169
16-block grid	90°	228	0	0
16-block grid	180°	276	0	0

here. The gap between sufficient and observed separation is expected from a worst-case bound over all recurring responses. These constructed feature sequences verify the algebra and illustrate coverage and alignment requirements. They do not establish those requirements on the archives, learned-feature accuracy, or whole-detector power. No benchmark result or model setting is changed by this study.

D.2 Score sensitivity, selection, and causality

Distribution-level checks. We check the endpoint, scale, and score inequalities on 2,000 pairs of compactly supported discrete reference laws and later-response laws, including unequal atom counts, unequal weights, and constant references. There are no violations, and 1,535 cases have nonzero score error. Another 300 comparisons replace a full historical response distribution by a 16-block evenly spaced grid. Each constructed history contains one rare extreme response. Of the 300 grids, 274 omit that endpoint. The transport bound still holds, but the example shows why the grid has no automatic endpoint-accuracy guarantee. These are deterministic approximation checks, not simulated change-detection benchmarks.

Five hundred paired-feature checks recover Eq. (20) from a coupling and confirm that the optimal-transport bound is no larger than the paired bound. Weighted joint laws give 1,000 startup-quantile checks, including endpoint quantiles. Another 1,000 checks test quantile perturbations under W_∞. Exhaustive enumeration of 464 exchangeable tied-rank cases verifies the empirical-quantile rule. All have zero violations. Four rare-mass examples reproduce the endpoint counterexample with $\varepsilon \in \{.1, .01, .001, .0001\}$ and $b = 8$: $W_2 = \sqrt{\varepsilon}$ shrinks while the score difference stays at eight. These computations check the algebra, not a population error rate.

Feature and selection checks. We evaluate the paired-feature corollary, Eq. (20), on 2,700 bounded block-mean perturbations. Each case has 16 historical blocks in 18 dimensions. Historical variation is zero, nearly zero, or positive; block widths are $b \in \{1, 4, 8, 16, 32\}$, and perturbation scales are 10^{-6}, .001, and .05. Directions are chosen from the perturbed historical blocks and held fixed for the paired score comparison, as required by the corollary. Both feature collections lie in the unit ball. Test blocks cover positive, zero, and negative nominal shifts along the direction. There are no violations of either the stated bound or its looser $bA + 2b^2E$ form. The largest observed ratio of score error to the sharper bound is .645. This is a numerical check of the inequality, not a probability bound on learned-feature error.

The selection-aware extension also needs the cross-context term. In four two-context examples, past selection scores $(.5, .5 - \varepsilon)$ and $(.5, .5 + \varepsilon)$ select different contexts with future scores $(0, 1)$, for $\varepsilon \in \{10^{-1}, 10^{-3}, 10^{-6}, 10^{-9}\}$. The selected-score difference remains one as the selection-score perturbation shrinks. Corollary 2 covers it, but a fixed-candidate distribution bound alone does not. On the 117 archive records, an additional selection-stability check compares the primary map with two learned seeds and native short features, giving 351 paired startup trajectories. It finds no violations when directions, context choices, arrivals, and expiry change. Mean bounds range

Table 8: Finite-horizon null alarms under matched whole-sequence sampling. Each entry pools 800 test evaluations across 80 independent reference/test experiments. Intervals resample experiments within each regime. Horizons and reference sizes reuse parent sequences and are not independent studies.

	$n = 19$ references		$n = 39$ references	
H	Alarms (%)	95% interval	Alarms (%)	95% interval
128	6.25	[4.38, 8.25]	4.12	[2.62, 5.75]
256	4.88	[3.25, 6.62]	5.12	[3.75, 6.50]
512	4.38	[2.75, 6.25]	3.50	[2.25, 4.88]

from .669 to 1.782, versus mean errors from .044 to .141 across archive/comparison groups. These comparisons use specified feature maps, not an unknown ideal representation.

Prefix replay and arbitrary future replacement leave the earlier scores unchanged in 240 checks across constant, noisy, and oscillatory inputs. Three thousand paired-proposal checks verify propagation through the maximum, persistence, and threshold-margin operations. Exhaustive enumeration of 644 small tied-rank cases also satisfies the strict-rank inequality, including the infinite-threshold case. These checks concern the implementation and algebra. Raw-time measurability, not the experiments, establishes causality.

D.3 Normal-sequence sampling

We generate 3,920 fresh, independent four-channel normal time series. Each has 512 observations, and its prefixes define horizons $H \in \{128, 256, 512\}$. There are no change alternatives in this study. Six independent latent components follow stationary Gaussian AR(1) models, with coefficient .8 in the steady regime and .6 in the cycling regime. For channel j, a shared component with weight .8 and an independent component with weight .6 are combined with measurement noise of standard deviation .1. Cycling adds a sinusoid of amplitude .75, period 64, and a uniformly random common phase, with fixed channel phase offsets. Latent states start in their stationary distributions. Initialization observations are not supplied to the detector.

The population baseline standard deviation is $\sigma_0 = (.8^2 + .6^2 + .1^2 + A^2/2)^{1/2}$, where A is zero or .75. Readings are quantized at $.01\sigma_0$, then converted to channel-specific units with gains $(.01, .1, 10, 100)$ and offsets $(.2, 5, 50, 1000)$. Every run uses the frozen primary detector and the same first-eight normalization described in Appendix B. The declared time unit is one observation. No physical sampling rate is assumed.

Within each regime, 40 independent experiments each supply 39 normal reference sequences and ten normal test sequences. Thresholds use the first $n \in \{9, 19, 39\}$ references and the order in Theorem 1 at $\alpha = .05$. Reference sizes are nested, and horizons share parent sequences, so those settings are not independent repetitions. For each setting, uncertainty is estimated by 10,000 bootstrap resamples of complete reference/test experiments within each regime. This preserves the dependence among test alarms sharing a threshold.

Table 8 is consistent with the marginal bound. Observed rates above five percent have intervals containing five percent. A finite Monte Carlo estimate need not lie below the population bound. Exchangeability holds by construction here, rather than being inferred from these rates. With nine reference sequences the threshold is infinite and every alarm count is zero. This is absence of a finite five-percent boundary, not useful detection. The deterministic rank rule needs at least 19 normal reference sequences.

D.4 Checks outside the theorem's assumptions

Three controls change the reference construction while keeping the detector fixed. Regime mismatch uses references from the other regime. Overlapping references treat 19 crops starting at $0, 8, \ldots, 144$ in a single parent as 19 units, while test parents remain independent; the crop length equals $H = 128$ or 256. The 3,040 resulting crop scores are not additional independent sequences. Horizon mismatch

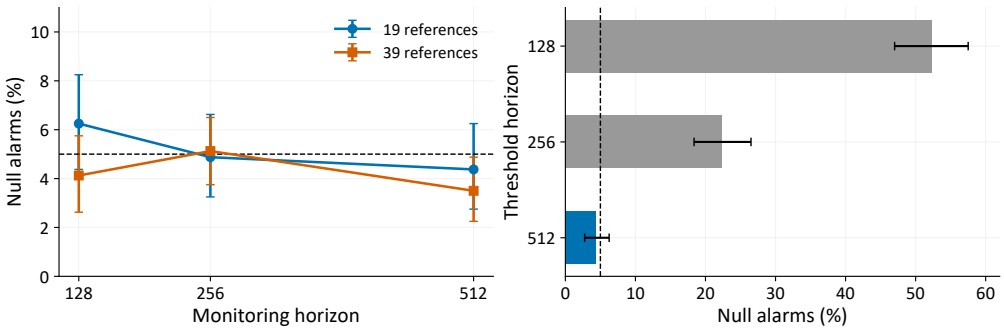

Figure 7: The finite-horizon sampling requirement. Left: matched normal-unit sampling with 19 or 39 reference sequences. Right: monitoring 512 observations with 19 references, using thresholds set at three different horizons. Whiskers are experiment-level bootstrap intervals. The dashed line is nominal five percent. Only the 512-observation threshold covers the full deployment horizon.

applies a threshold set at 128 or 256 observations to monitoring through 512. All three controls are outside the stated theorem.

At $n = 19$, horizon mismatch raises the null-alarm rate from 4.375% under the matched 512-observation threshold to 52.25% using a 128-observation threshold and 22.375% using a 256-observation threshold (Figure 7). The corresponding intervals are $[2.75, 6.25]\%$, $[47.00, 57.50]\%$, and $[18.375, 26.50]\%$. This illustrates why the theorem cannot justify indefinite monitoring with a threshold set for a shorter horizon.

The other controls are less conclusive. With 19 references, overlapping crops give 6.00% and 5.375% at horizons 128 and 256, versus matched rates of 6.25% and 4.875%. Regime-mismatched rates are 5.75%, 6.625%, and 6.125% at the three horizons. These controls show no comparably clear inflation pattern in the chosen processes. An observed rate near the target does not restore the missing sampling assumption.

D.5 Threshold cost of startup

On each normal reference collection, we separately record startup maxima A_i, mature maxima B_i, and complete maxima $\max(A_i, B_i)$. Across 480 finite-threshold cases, both bounds in Proposition 3 hold. At $H = 256$ with 19 references, startup raises the threshold in 59 of 80 experiments, with a mean increase of $.414$ score units. At $H = 512$ the increase is zero in all 80 experiments with 19 references. With 39 references, five of 80 have positive cost. Thus the cost can be positive or vanish even though the mature score is unchanged. At $H = 128$ the mature stage is unavailable, so its zero maximum is only an algebraic control. None of these results is a detection-delay or power guarantee.

E Computational cost

For one proposal, J projected historical block means give the atoms of $\widehat{P}_m$. Computing their endpoint and standard deviation costs $O(Jp)$ operations once the block means are available, with feature dimension p. A later block then needs $O(p)$ work for its projection and constant work for $\mathcal{S}_b(T, \widehat{P}_m)$ using its mean v and cached u_m, d_m. This avoids fitting a density or comparing the later block with all historical atoms again. It does not approximate the chosen empirical score. This algebraic cost statement is not a benchmark against every density-estimation method.

The current implementation computes feature cumulative sums and stores history to form reference statistics. For the mature score, dyadic widths and eight retained proposal ages require $O(p \log t)$ scoring work per endpoint after feature extraction, with fixed historical-block count 16; $p = 18d$. Startup is restricted to the first 174 observations. The exact batch replay stores cumulative feature statistics and trajectories, with memory growing with sequence length. We do not claim a constant-memory implementation.

Figure 5 compares all thirteen Table 1 methods. We select records by the 25th, 50th, and 75th percentiles of sorted raw lengths, with names breaking ties, without inspecting detection outcomes: Symbols, DodgerLoopDay, and Lightning7 for TSSB; global_co2, children_per_woman, and brent_spot for TCPD. Each method uses its saved configuration from outer partition 91001. One complete warmup precedes three repetitions with randomized method order and fresh scoring state. The plotted cost is the median across records of each record's median time divided by its length. These timing records do not constitute a full-archive average. Detection performance is reported separately on the complete archives.

Feature extraction and any within-sequence fitting are timed. Data loading, offline training, and frozen weight loading are not. Shared multi-head diagnostic routines are reduced to the requested native or direct score. Their trajectories are checked against the original calls. All 78 method–record pairs reproduce the saved first alarm, including no alarms. Short sequences may end before a baseline becomes ready, reducing its runtime without implying successful detection. This is a benchmark of the current implementations, not each method's best possible implementation.

Per-observation latency. The right panel of Figure 5 measures an incremental implementation of the unchanged detector. All 117 archive sequences under both preprocessing options reproduce the batch scores to within 2.18×10^{-14}, preserving all 2,340 checked fitted stopping decisions. Frozen weights are loaded into fresh one-thread processes. An untimed zero-input state initializes numerical kernels and is discarded before a new detector receives the timed sequence. Each update call is timed with a wall clock on a fixed Gaussian sequence of 8,192 observations, sliced to 1, 8 or 52 channels. Startup comprises the first 174 updates. The remainder use the mature stage. For each phase and channel count, the plotted median and 99th percentile are the medians of the respective run-level statistics over three fresh processes. These are timing summaries, not confidence intervals. At 52 channels, retained numeric state is 123 MB and peak process memory is approximately 303 MB. The largest recorded mature update is about 32 ms, so a small 99th percentile does not establish a worst-case latency bound. This experiment does not measure CPU utilization and is not a matched incremental-latency comparison against the other detectors.

All timing measurements use an Intel Core i9-14900KF CPU with one numerical thread.

F LIMITATIONS

The distributional stability result requires control of endpoint displacement, not merely average fit. A small empirical transport error at a finite order does not by itself bound the score uniformly over arbitrary reference laws. The conditional reference describes sampled historical responses, not an identified unchanged distribution or a conditional law for future data. The deterministic block grid may miss historical extremes, and caching its endpoint and variability does not use all distributional information. The first-eight transform and accumulated history can already contain changed data. An extreme historical projection can suppress later evidence. Bounded features control numeric magnitude, not contamination bias. A reversal or an orthogonal change requires a new informative proposal, and a finite set of directions is not universally sensitive. Replacing short contexts may miss later short-lived changes. A stronger application-transfer study and a dedicated mature-gate ablation would be needed for broader neural claims. The component comparisons do not establish universal superiority. The false-alarm guarantee requires exchangeable normal reference and test sequences and a fixed horizon; simulation cannot establish exchangeability in a new application. Detection power is not guaranteed for arbitrary time-series distributions, and memory grows with the observed history.

G AI USE STATEMENT

Generative AI tools assisted with writing and organization, literature searches, method and experiment design, implementation and synthetic-data generation, result interpretation, and formulation and review of mathematical claims and proofs. The authors are responsible for verifying this assistance and for all final claims, code, and results.