LOCAL RISK STRUCTURE FOR ADAPTIVE RETRAINING IN STREAMING TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Predictors for streaming time series must decide how much new data to collect before updating, even though the future benefit of an update is unknown. We propose RETICLE (Risk Evaluation Through Interval Comparisons of Local Errors), a data-only audit of accumulated-data quality for selecting an update size without retraining the downstream predictor at each candidate size. RETICLE fits causal local affine probes in a frozen representation and assesses their risk across bandwidths. We show that approximate quasiconvexity of the risk profile on a finite bandwidth grid is necessary for finite-sample local-learning regularity; monotone profiles are included as special cases. Using this necessary consequence as a data-quality criterion, RETICLE certifies the risk-shape condition with simultaneous confidence bounds and selects the first adequately supported checkpoint that passes. Under causal auditing assumptions, we control the probability of false shape certification simultaneously over the declared checkpoints and representations, including the selected checkpoint, without requiring independent observations. Experiments examine checkpoint selection for six downstream learners on synthetic and real time series, as well as continuous fine-tuning on three forecasting datasets with two drift detectors. RETICLE achieves the lowest test error in 12 of 18 online settings, ties for the lowest in two more, and attains an overall mean rank of 1.28 among the compared baselines.

1 INTRODUCTION

Forecasting systems in energy monitoring, environmental sensing, and industrial operations (Lim & Zohren, 2021; Benidis et al., 2022) receive multivariate observations over time and must adapt as the underlying dynamics change (Pham et al., 2023; Zhang et al., 2023; Li et al., 2025). After a change, the system must decide when to update: updating early means fitting on only a few observations from the new regime, whereas waiting longer prolongs reliance on an outdated predictor (Figure 1(a)). Because waiting is costly, the error attainable with the most data is not the benchmark for this decision; the question is whether the data collected so far support an update.

This decision is not settled by a drift alarm alone, which signals that a change has occurred but not how much data to collect. Comparing candidate sizes by trial retraining requires repeated access to the downstream learner and ties the assessment to its training procedure. We therefore ask: *How can the observed data support an update-size decision without inspecting or querying the downstream predictor?* Such a decision would decouple assessing the accumulated data from choosing and training the model that eventually uses it.

CALIPER provides the most direct precedent for this data-only objective: it selects post-drift data sizes using the effective sample size and a monotonically improving local prediction error (Fujiwara et al., 2026a). At finite sample sizes, however, narrower neighborhoods can improve local approximation while leaving fewer effective observations for fitting (Fan & Gijbels, 1996). An increase in error at fine scales can therefore be compatible with local learning, motivating a criterion that accommodates the resulting trade-off. High-dimensional histories containing many variables and time lags further motivate compact coordinates for this assessment. Repeatedly examining candidate data sizes also requires uncertainty control that covers the checkpoint ultimately selected.

We introduce RETICLE (**R**isk **E**valuation **T**hrough **I**nterval **C**omparisons of **L**ocal **E**rrors), a data-only audit for selecting a retraining data size. RETICLE uses a necessary consequence of local-

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

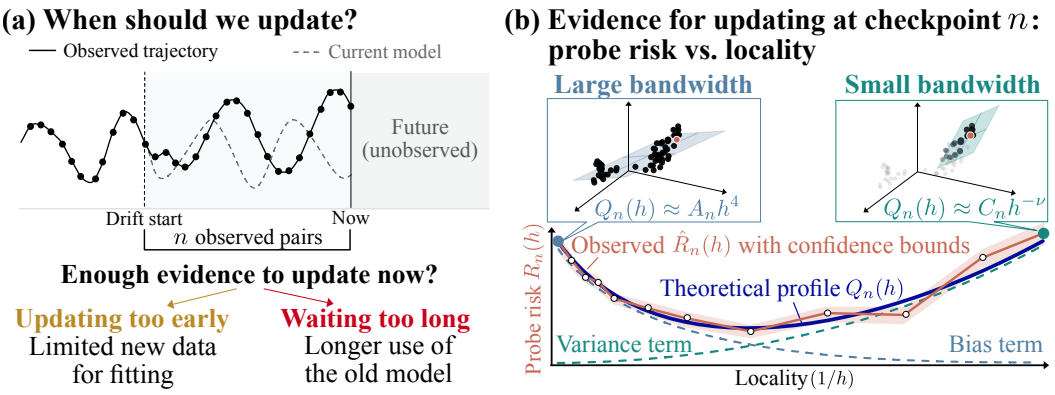

Figure 1: **Selecting an update size from local risk structure.** (a) The update-timing trade-off. (b) Local affine probes at data size n. Blue: theoretical profile $Q_n(h) = a_n + A_n h^4 + C_n h^{-\nu}$; orange: empirical risk $\widehat{R}_n(h)$ with schematic bounds. RETICLE certifies approximate quasiconvexity of $R_n(h)$ with simultaneous confidence bounds and selects the first passing checkpoint.

learning regularity as a data-quality criterion, certifying its satisfaction from probe losses without directly verifying the full regularity condition. RETICLE evaluates local affine probes in a compact representation learned separately from the downstream predictor and frozen before auditing. At each data size, it examines whether the probe risk is approximately quasiconvex on the finite bandwidth grid (Figure 1(b)). Quasiconvexity describes a nonincreasing-then-nondecreasing profile and includes monotone profiles. RETICLE uses simultaneous confidence bounds to select the first planned checkpoint with adequate local support and a certified risk-shape condition. Relative to CALIPER, RETICLE (i) audits a compact predictive representation of histories rather than the observed state space, (ii) replaces the requirement that error never increase with locality by an approximate quasiconvexity condition, which accommodates estimation cost at fine scales, and (iii) certifies this condition with simultaneous confidence bounds that remain valid at the selected checkpoint. Our contributions are threefold:

- **Learner-agnostic auditing in a predictive representation.** A representation-based audit of high-dimensional histories that, once configured, needs no downstream parameters, prediction queries, or trial retraining; one selection is shared across six learners.
- **A theoretically motivated data-quality criterion.** Under a local-learning regularity condition, the conditional probe risk lies close to a quasiconvex profile, so approximate quasiconvexity on the bandwidth grid is a necessary consequence of regularity (Proposition 1).
- **Statistical validity under checkpoint selection.** Under causal auditing assumptions allowing temporal dependence, false shape certification is controlled simultaneously over declared checkpoints and representations, including the selected checkpoint (Theorem 1).

2 Related Work

Adaptation and update-size selection. Concept-drift methods detect changes, adjust retained data, and adapt predictors (Gama et al., 2004; Bifet & Gavaldà, 2007; Gama et al., 2014; Gomes et al., 2017; Raab et al., 2020), and sequential forecasters update predictors and ensembles as observations arrive (Anava et al., 2013; Pham et al., 2023; Zhang et al., 2023; Lau et al., 2025). Others adapt to evolving regimes through model adjustment, normalization, or dynamical representations (Kim et al., 2022; Liu et al., 2022; 2023; Kim et al., 2025; Zhao & Shen, 2025; Chihara et al., 2025; Fujiwara et al., 2026b; Chihara et al., 2026). Most closely related, CALIPER selects post-drift retraining sizes using the effective sample size together with a one-step prediction error that must not increase as locality increases (Fujiwara et al., 2026a). RETICLE shares this data-only objective but audits a frozen predictive representation with an approximate quasiconvexity condition that tolerates higher error at fine scales.

Local prediction and bandwidth-dependent risk. State-space approaches to time-series prediction include reconstructed coordinates (Farmer & Sidorowich, 1987; Casdagli, 1989; Sugihara & May, 1990) and learned latent dynamics (Krishnan et al., 2017; Rangapuram et al., 2018; Rubanova et al., 2019; Fujiwara et al., 2025). The S-map method examines nonlinear state dependence through changes in forecasting performance as localization varies (Sugihara, 1994). Locally weighted and local polynomial regression provide the underlying estimation tools and analyses of design adaptation, bias, and variance (Cleveland & Devlin, 1988; Fan, 1992; Ruppert & Wand, 1994; Fan & Gijbels, 1996). Variable bandwidths allow smoothing to vary across the input space (Fan & Gijbels, 1992), while comparisons across scales enable adaptation to unknown, spatially inhomogeneous smoothness (Lepski et al., 1997). RETICLE uses the risk profile across bandwidths to assess accumulated data before a downstream update.

Sequential evaluation and statistical certification. Prequential evaluation scores predictions against realized outcomes (Dawid, 1984), while forecast-comparison tests use loss differences and conditional predictive ability (Diebold & Mariano, 1995; Giacomini & White, 2006). Martingale concentration (Freedman, 1975), confidence sequences (Howard et al., 2021), and betting-based inference (Waudby-Smith & Ramdas, 2024; Ramdas et al., 2023) support sequential uncertainty quantification. Sequential forecast comparison studies time-varying differences in expected predictive scores (Choe & Ramdas, 2024). Testable learning links acceptance of data-dependent tests to learner guarantees under distributional assumptions or distribution shift (Rubinfeld & Vasilyan, 2023; Klivans et al., 2024). RETICLE converts paired loss-difference bounds on common audit queries into a certificate for a necessary risk-shape condition that covers the declared finite audit family and the first passing checkpoint.

3 Proposed Method

At each candidate data size n, RETICLE compares local-probe risks across bandwidths h from accumulated input–response pairs in a frozen representation, and selects the first adequately supported checkpoint whose profile is certified approximately quasiconvex. The criterion is motivated by the local bias–variance trade-off: narrower neighborhoods reduce approximation error but leave fewer effective references for estimation (Figure 1(b)). Proposition 1 establishes approximate quasiconvexity as a necessary consequence of local-learning regularity, which RETICLE audits as data accumulate. For fixed c, a larger minimum local effective sample size gives a smaller tolerance η_n (Section 3.2); an adequate checkpoint passes when the upper bound $\mathrm{U}_{V,n}$ is at most η_n.

3.1 Problem definition

After a regime change or drift alarm, we select an update size for a downstream predictor. Let $\mathcal{F}_t$ denote the information available at prediction time, $\mathbf{z}_t \in \mathbb{R}^d$ the encoded history, and $\mathbf{y}_{t+1} \in \mathbb{R}^p$ the next response. Our implementation projects standardized histories onto predictive ridge-regression/SVD directions fitted on preparation data, with PCA completion when needed (Appendix K). The encoder and normalization scales are frozen during auditing; response coordinates absorb the fixed response scales. Bold lowercase and uppercase symbols denote vectors and matrices; upright type marks random Latin-letter quantities.

Assumption 1 (Structural prediction model)**.** *Let the open domain $\mathbb{D}$ contain the queried and reference states and their connecting line segments. For $0 < \beta \leq 1$, assume:*

- *(i)* A fixed latent prediction rule. *A function $\boldsymbol{f} : \mathbb{D} \to \mathbb{R}^p$, fixed at audit start, satisfies $\mathbf{y}_{t+1} = \boldsymbol{f}(\mathbf{z}_t) + \boldsymbol{\epsilon}_{t+1}$ and $\mathbb{E}[\boldsymbol{\epsilon}_{t+1} \mid \mathcal{F}_t] = \mathbf{0}$ at audit times, with finite conditional second moments.*
- *(ii)* Bounded local sensitivity and curvature. *Each response component f_a, $a = 1, \ldots, p$, is C^2 on $\mathbb{D}$, with $\|\nabla f_a(\boldsymbol{z})\| \leq G_a$ and $\|\nabla^2 f_a(\boldsymbol{z})\|_F \leq L_{2,a}$.*
- *(iii)* Locally stable curvature. *For $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{D}$, $\|\nabla^2 f_a(\boldsymbol{z}) - \nabla^2 f_a(\boldsymbol{z}')\|_F \leq L_{\beta,a}\|\boldsymbol{z} - \boldsymbol{z}'\|^\beta$.*

Finite-sample regularity. At checkpoint n, let $\mathbb{T}_n$ be the declared audit times and $\mathbb{H}_{\mathrm{f}} = \{h_1 > \cdots > h_J > 0\}$, $J \geq 3$, the finite bandwidth grid. Using the matured references $\mathbb{I}_t$, common support S_t, and bandwidth-independent weights $\mathrm{w}_{t,n} \in [0, 1]$ of Section 3.2, write the causal local

ridge probe as $\widehat{\mathbf{f}}_{t,h} = \sum_{i\in\mathbb{I}_t} \mathrm{a}_{ti,h}\mathbf{y}_{i+1}$, with coefficients from Appendix G, equation (44), and define $\mathrm{W}_n = \sum_{t\in\mathbb{T}_n} \mathrm{S}_t\mathrm{w}_{t,n} > 0$. Define reference residuals $\boldsymbol{\epsilon}_{i+1} = \mathbf{y}_{i+1} - \boldsymbol{f}(\mathbf{z}_i)$ algebraically, including earlier-regime mismatch; no centering or independence is assumed for these residuals. Put

$$\mathbf{b}_{t,h} = \sum_i \mathrm{a}_{ti,h}\boldsymbol{f}(\mathbf{z}_i) - \boldsymbol{f}(\mathbf{z}_t), \qquad \boldsymbol{\xi}_{t,h} = \sum_i \mathrm{a}_{ti,h}\boldsymbol{\epsilon}_{i+1}, \qquad \langle \boldsymbol{u}, \boldsymbol{v}\rangle_{n,w} = \frac{\sum_{t\in\mathbb{T}_n} \mathrm{S}_t\mathrm{w}_{t,n}\boldsymbol{u}_t^\top \boldsymbol{v}_t}{p\mathrm{W}_n}. \tag{1}$$

Here $\boldsymbol{u} = (\boldsymbol{u}_t)_{t\in\mathbb{T}_n}$ and $\boldsymbol{v} = (\boldsymbol{v}_t)_{t\in\mathbb{T}_n}$ stack p-vectors, as do $\mathbf{b}_h$ and $\boldsymbol{\xi}_h$; write $\|\boldsymbol{u}\|_{n,w}^2 = \langle \boldsymbol{u}, \boldsymbol{u}\rangle_{n,w}$. For fixed clipping level $M > 0$, define the removed tail

$$\kappa_n(h) = \frac{1}{\mathrm{W}_n} \sum_{t\in\mathbb{T}_n} \mathrm{S}_t\mathrm{w}_{t,n}\mathbb{E}\left[\left(p^{-1}\|\mathbf{y}_{t+1} - \widehat{\mathbf{f}}_{t,h}\|^2 - M\right)_+ \middle| \mathcal{F}_t\right]. \tag{2}$$

The following definition formalizes finite-sample regularity in approximation and residual scaling across bandwidths; smoothness alone does not ensure this behavior.

Definition 1 (Weighted local-learning regularity)**.** *For $\delta \geq 0$, let $\mathcal{G}_{n,\delta}$ be the event that there exist a bandwidth-independent sequence $\boldsymbol{b}^{(0)} = (\boldsymbol{b}_t^{(0)})_{t\in\mathbb{T}_n}$ with $\boldsymbol{b}_t^{(0)} \in \mathbb{R}^p$, constants $C_n \geq 0$, $\nu > 0$, and $\kappa_n^{(0)} \in \mathbb{R}$ for which, writing $e_{n,h} = \|\mathbf{b}_h - h^2\boldsymbol{b}^{(0)}\|_{n,w}$,*

$$\max_{h\in\mathbb{H}_\mathrm{f}} \left\{ \begin{array}{l} (2h^2\|\boldsymbol{b}^{(0)}\|_{n,w} + e_{n,h})e_{n,h} + \left|\|\boldsymbol{\xi}_h\|_{n,w}^2 - C_n h^{-\nu}\right| \\ +\, 2|\langle \mathbf{b}_h, \boldsymbol{\xi}_h\rangle_{n,w}| + |\kappa_n(h) - \kappa_n^{(0)}| \end{array} \right\} \leq \delta. \tag{3}$$

The four terms control quadratic approximation, reference-residual energy, their interaction, and clipping distortion. Regularity is relative to the true rule, reference sample, frozen representation, and probe; ν need not equal d. Under Assumption 1(ii)–(iii), Taylor bounds on $e_{n,h}$ combine with residual and clipping bounds to imply $\mathcal{G}_{n,\delta}$ when their combined bound is at most δ (Appendix G, equation (48)).

Risk-shape certification. For the clipped loss $\ell_{t,h} = \min\{p^{-1}\|\mathbf{y}_{t+1} - \widehat{\mathbf{f}}_{t,h}\|^2, M\}$, define the weighted conditional probe risk

$$\mathrm{R}_n(h) = \frac{\sum_{t\in\mathbb{T}_n} \mathrm{S}_t\mathrm{w}_{t,n}\mathbb{E}[\ell_{t,h} \mid \mathcal{F}_t]}{\mathrm{W}_n}. \tag{4}$$

Proposition 1 gives $\mathcal{G}_{n,\eta_n/2} \Rightarrow V(\mathbf{r}_n) \leq \eta_n$, where $\mathbf{r}_n = (\mathrm{R}_n(h_1), \ldots, \mathrm{R}_n(h_J))$ and V is the quasiconvexity score in equation (9). RETICLE certifies this necessary risk-shape condition; the full structural event involves the unknown $\boldsymbol{f}$ and the reference residuals.

Problem 1 (Certifying a necessary condition for local-learning regularity)**.** Given declared checkpoints $\mathbb{N} = \{n_1 < \cdots < n_K\}$, tolerances $\eta_n \geq 0$, and an error level $\alpha \in (0, 1)$, use matured responses to certify $V(\mathbf{r}_n) \leq \eta_n$ and select the first adequate passing checkpoint for updating the downstream predictor, while controlling

$$\mathbb{P}\{\exists n \in \mathbb{N} : \mathsf{PASS}_n \text{ and } V(\mathbf{r}_n) > \eta_n\} \leq \alpha.$$

3.2 RETICLE: WEIGHTED QUASICONVEXITY AUDITING

Causal local probes. The probe horizon is $H = 1$; the downstream horizon may differ. References form a growing prefix indexed from zero at a fixed origin. Checkpoint n counts matured input–response pairs, with $\mathbb{T}_n \subseteq \{0, \ldots, n-1\}$ and responses available by n. For an integer past-exclusion gap $g \geq 0$, let $\mathbb{I}_t = \{i \in \mathbb{Z}_{\geq 0} : i < t - g\}$ and $m_t := |\mathbb{I}_t| = \max(0, t - g)$. At bandwidth h, put $\mathbf{u}_{ti} = \mathbf{z}_i - \mathbf{z}_t$ and

$$\mathrm{k}_{ti,h} = \exp\left(-\frac{\|\mathbf{u}_{ti}\|^2}{2dh^2}\right), \qquad \pi_{ti,h} = \frac{\mathrm{k}_{ti,h}}{\sum_{r\in\mathbb{I}_t} \mathrm{k}_{tr,h}}. \tag{5}$$

The prediction $\widehat{\mathbf{f}}_{t,h}$ is the intercept of

$$(\widehat{\mathbf{a}}, \widehat{\mathbf{B}}) = \underset{\boldsymbol{a}\in\mathbb{R}^p,\, \boldsymbol{B}\in\mathbb{R}^{p\times d}}{\arg\min} \left\{ \sum_{i\in\mathbb{I}_t} \pi_{ti,h}\|\mathbf{y}_{i+1} - \boldsymbol{a} - \boldsymbol{B}\mathbf{u}_{ti}\|^2 + \rho\|\boldsymbol{B}\|_F^2 \right\}. \tag{6}$$

The intercept is unpenalized. The same finite grid $\mathbb{H}_\mathrm{f}$ is used for the probes, common support, and risk comparisons.

Commonly weighted audit risks. The predictable common-support indicator S_t requires all $h \in \mathbb{H}_{\mathrm{f}}$ to meet the local support conditions in Appendix I. Undefined or unsupported outputs may take arbitrary finite values because their common weight is zero. We estimate equation (4) using reference-count weights:

$$\mathrm{w}_{t,n} = \frac{m_t}{\max(1, n-1-g)} \in [0,1], \tag{7}$$

$$\widehat{\mathrm{R}}_n(h) = \frac{\sum_{t\in\mathbb{T}_n} \mathrm{S}_t \mathrm{w}_{t,n} \ell_{t,h}}{\mathrm{W}_n}. \tag{8}$$

All ratios require $\mathrm{W}_n > 0$. Common weights favor predictions with more references and preserve the comparison population, without claiming inverse-variance optimality. The target averages conditional probe risks, distinct from held-out downstream risk.

Quasiconvexity statistic and decision. For $\boldsymbol{r} \in \mathbb{R}^J$, define

$$\begin{aligned} V(\boldsymbol{r}) &= \max_{i<j<k} \min\{r_j - r_i, r_j - r_k\} = \max_{i<j<k} \big[r_j - \max\{r_i, r_k\}\big], \\ \mathbb{V}_J &= \bigcup_{v=1}^{J} \{\boldsymbol{r} : r_1 \geq \cdots \geq r_v \leq \cdots \leq r_J\}. \end{aligned} \tag{9}$$

On the ordered grid, $\mathbb{V}_J$ is the class of quasiconvex profiles, equivalently $V(\boldsymbol{r}) \leq 0$ (Boyd & Vandenberghe, 2004, Section 3.4). It includes monotone and constant profiles and flat or endpoint minima. The score is signed. Approximate quasiconvexity means $V(\boldsymbol{r}) \leq \eta$ for additive tolerance $\eta \geq 0$, with no claim between audited bandwidths. Under Assumption 2, martingale concentration and error-budget allocation yield simultaneous intervals $[\mathrm{L}_{n,ij}, \mathrm{U}_{n,ij}]$ for $\mathrm{R}_n(h_j) - \mathrm{R}_n(h_i)$ from paired losses over the declared finite family of bandwidth pairs, checkpoints, and representations (Appendix C). RETICLE converts these intervals into bounds on the quasiconvexity score:

$$\mathrm{L}_{V,n} = \max_{i<j<k} \min\{\mathrm{L}_{n,ij}, -\mathrm{U}_{n,jk}\}, \qquad \mathrm{U}_{V,n} = \max_{i<j<k} \min\{\mathrm{U}_{n,ij}, -\mathrm{L}_{n,jk}\}, \tag{10}$$

where these maxima range over $1 \leq i < j < k \leq J$ only. Let

$$\mathrm{N}_{t,h} = \frac{(\sum_{i\in\mathbb{I}_t} \mathrm{k}_{ti,h})^2}{\sum_{i\in\mathbb{I}_t} \mathrm{k}_{ti,h}^2}, \quad \mathrm{N}_{\min,n} = \min_{\substack{t\in\mathbb{T}_n : \mathrm{S}_t = 1 \\ h\in\mathbb{H}_{\mathrm{f}}}} \mathrm{N}_{t,h}, \qquad \eta_n = \frac{c}{\sqrt{\mathrm{N}_{\min,n}}}, \quad c \geq 0. \tag{11}$$

Let Adeq_n denote the reference-count, common-support, support-fraction, and representation conditions specified in Appendix I, including $\mathrm{W}_n > 0$ and a finite, valid η_n. Where η_n is undefined (outside adequate audits), we extend it to zero; such audits remain unresolved. RETICLE reports

$$\begin{cases} \mathsf{PASS}_n, & \mathsf{Adeq}_n \text{ and } \mathrm{U}_{V,n} \leq \eta_n, \\ \mathsf{VIOLATED}_n, & \mathsf{Adeq}_n \text{ and } \mathrm{L}_{V,n} > \eta_n, \\ \mathsf{UNRESOLVED}_n, & \text{otherwise.} \end{cases} \tag{12}$$

RETICLE selects the first PASS_n checkpoint in $\mathbb{N}$, if any. Simultaneous coverage preserves shape certification at this selected checkpoint (Theorem 1). Algorithm 1 in Appendix B gives the full procedure.

4 Theoretical Analysis

We establish structural necessity, simultaneous certification, and conditional precision, then state the computational scope.

4.1 Structural necessity

The necessary condition uses only Assumption 1(i) and Definition 1; the smoothness-based sufficient bounds are in Appendix G. The quasiconvexity score satisfies $\mathrm{dist}_\infty(\boldsymbol{r}, \mathbb{V}_J) = [V(\boldsymbol{r})]_+/2$ (Lemma 1 in Appendix E).

Proposition 1 (Structural necessity). *Under Assumption 1(i), let* $\mathrm{W}_n > 0$ *and, on* $\mathcal{G}_{n,\delta}$*, fix witnesses* $\boldsymbol{b}^{(0)}, C_n, \nu, \kappa_n^{(0)}$ *from Definition 1. There exist* $a_n \in \mathbb{R}$ *and* $A_n \geq 0$ *such that, with these same* C_n, ν*,*

$$\max_{h \in \mathbb{H}_{\mathrm{f}}} |\mathrm{R}_n(h) - (a_n + A_n h^4 + C_n h^{-\nu})| \leq \delta. \tag{13}$$

Consequently $V(\mathbf{r}_n) \leq 2\delta$ *and* $\mathrm{dist}_\infty(\mathbf{r}_n, \mathbb{V}_J) \leq \delta$*. In particular,*

$$\mathcal{G}_{n,\eta_n/2} \quad \Longrightarrow \quad V(\mathbf{r}_n) \leq \eta_n. \tag{14}$$

Proof sketch. The clipped risk decomposes exactly as $\mathrm{R}_n(h) = \sigma_n^2 + \|\mathbf{b}_h + \boldsymbol{\xi}_h\|_{n,w}^2 - \kappa_n(h)$, where σ_n^2 denotes the weighted conditional innovation variance. Expanding this expression and applying Definition 1 gives equation (13) with $A_n = \|\boldsymbol{b}^{(0)}\|_{n,w}^2$ and $a_n = \sigma_n^2 - \kappa_n^{(0)}$. The ideal curve is convex in $\log h$ and hence quasiconvex on the bandwidth grid. Perturbing its coordinates by at most δ changes each triple contrast by at most 2δ. The complete proof is in Appendix F; the distance identity is proved in Appendix E. □

Remark 1 (Role of the necessary condition). *Proposition 1 establishes approximate quasiconvexity as a necessary consequence of local-learning regularity. The sufficient bounds in Appendix G require control of approximation, reference-residual, and clipping terms not directly observable from probe losses. In particular,* $\mathcal{G}_{n,\delta}$ *constrains the cross term* $\langle \mathbf{b}_h, \boldsymbol{\xi}_h \rangle_{n,w}$ *between approximation error and smoothed reference residual, which need not be small under Assumption 1 alone; regularity is thus an event about the realized data, not an assumption, and RETICLE never presumes it. It therefore certifies the necessary risk-shape condition from observed probe-risk comparisons. Under Assumption 1(i), a certified violation rules out regularity at the corresponding tolerance, whereas a PASS certifies only this prerequisite for selecting an update candidate, not full regularity.*

4.2 Statistical certification

We establish simultaneous guarantees for risk-shape decisions over the declared checkpoints and representations, including the first passing planned checkpoint. These statistical guarantees require only Assumption 2; structural refutation additionally invokes Assumption 1(i).

Assumption 2 (Causal finite-family auditing). *The* K *checkpoints, their candidate audit sets in* $\{0, \ldots, n-1\}$*, the finite bandwidth grid, and a finite family of* R_{rep} *representations are declared before the corresponding audit labels are observed. Each representation and its scales are frozen before its audit. Probe predictions,* S_t*, and* $\mathrm{w}_{t,n}$ *are* $\mathcal{F}_t$*-measurable;* $\mathbf{y}_{t+1}$ *is* $\mathcal{F}_{t+1}$*-measurable. For each nonempty audit with declared start* $s_n \leq \min \mathbb{T}_n$*, the candidate schedule, endpoint, planned weights, and interval tuning parameters are* $\mathcal{F}_{s_n}$*-measurable, with tuning parameters fixed throughout the audit. The clipping level* $M > 0$ *is fixed, and the confidence budget covers the full declared family.*

Theorem 1 (Finite-family certification and structural refutation). *Under Assumption 2, the intervals of Appendix C define an event* $\mathcal{E}$ *with* $\mathbb{P}(\mathcal{E}) \geq 1 - \alpha$ *(in fact* $1 - \alpha/2$*; see Appendix D.3) on which every declared representation and checkpoint with* $\mathrm{W}_n > 0$ *satisfies*

$$\mathrm{L}_{V,n} \leq V(\mathbf{r}_n) \leq \mathrm{U}_{V,n}. \tag{15}$$

On the same event,

$$\mathsf{PASS}_n \Longrightarrow \mathrm{dist}_\infty(\mathbf{r}_n, \mathbb{V}_J) \leq \tfrac{1}{2}[\mathrm{U}_{V,n}]_+ \leq \eta_n/2, \tag{16}$$

$$\mathsf{VIOLATED}_n \Longrightarrow V(\mathbf{r}_n) > \eta_n. \tag{17}$$

Under Assumption 1(i), equation (17) also implies $\mathcal{G}_{n,\eta_n/2}^c$*. Thus, the probability of any false shape certification or false shape refutation within the declared family is at most* α*. For the first passing planned checkpoint* $\tau = \inf\{n \in \mathbb{N} : \mathsf{PASS}_n\}$*,*

$$\mathbb{P}\{\tau < \infty,\ V(\mathbf{r}_\tau) > \eta_\tau\} \leq \alpha. \tag{18}$$

Proof sketch. A weighted, self-normalized exponential inequality with predictable stopping for range candidates and a union bound over the finite family covers all declared paired conditional risk differences simultaneously, without independence across times, pairs, or overlapping windows. Coordinatewise monotonicity of the maxima and minima in equation (10) gives equation (15); all conclusions then hold pathwise, including the data-adaptive tolerance and the first passing checkpoint. The complete proof is in Appendix D. □

These conclusions hold on the same simultaneous event for any finite, nonnegative $\mathcal{F}_n$-measurable tolerance η_n.

4.3 Precision and computational scope

On $\mathcal{E}$, the reported upper bound satisfies $\mathrm{U}_{V,n} \leq V(\mathbf{r}_n) + 2\varrho_n$, where ϱ_n bounds the selected pair radii in equation (23) before intersection with the candidate range (Lemma 2 in Appendix E); an adequate checkpoint therefore passes whenever $2\varrho_n \leq \eta_n - V(\mathbf{r}_n)$. Because the λ-grid is finite and bounded away from zero, radii need not vanish, and no unconditional sample-complexity or eventual-passage guarantee is claimed; the adaptive tolerance does not estimate the structural remainder, and ESS alone does not establish $\mathcal{G}_{n,\eta_n/2}$. Converting paired intervals into bounds on V costs $O(J^2)$ per checkpoint (Proposition 2 in Appendix H).

5 Experiments

We evaluate two complementary settings for selecting an update size. The checkpoint study (Section 5.1) isolates data accumulation within a regime, motivated by learning after an abrupt change. It uses synthetic trajectories and real-data episodes with a prescribed monitoring start rather than a detected boundary. Because waiting prolongs use of the outdated predictor (Figure 1(a)), error at the largest checkpoint is not the objective; we assess selections by the marginal benefit of additional data rather than by the error attainable with the most data. The online study (Section 5.2) instead uses long, unsegmented forecasting streams to evaluate alarm-initiated data collection and repeated neural-model fine-tuning. In both studies, RETICLE audits one-step prediction losses over 14 finite bandwidths at level $\alpha = 0.05$, using a 32-dimensional predictive representation frozen within each evaluation phase. A checkpoint n counts fully observed one-step probe pairs; downstream training uses only examples whose full forecast targets are observed. Full settings are given in Appendix K.

5.1 Checkpoint selection across downstream learners

Setup. We use three synthetic systems—Lorenz–96, coupled FitzHugh–Nagumo, and Kuramoto–Sivashinsky—with 10 seeds each, and 16 real cases: four rotating-machinery runs (Saucedo-Dorantes et al., 2026), eight treadmill intervals (Krumm et al., 2022; 2026), and four WESAD baseline episodes (Schmidt et al., 2018). The largest checkpoint is 16,384 for synthetic data, 8,192 for rotating machinery and WESAD, and 4,096 for treadmill. After representation preparation, six learners (KRR, MLP, ExtraTrees, Transformer, XGBoost, and TCN) are fitted from scratch on observed histories at common checkpoints and evaluated on fixed held-out data. With $c = 0.1$, RETICLE selects the first adequate passing checkpoint once per run, without access to any learner or its errors, and this single selection is shared by all six learners. If certification reflects data adequacy for local learning, learners should show large gains per additional pair before the selection and small gains after it. Figure 2 tests this prediction using the decrease in normalized held-out MSE per additional probe pair between adjacent checkpoints, computed within each run before aggregation (Appendix K). Update timing is evaluated in Section 5.2 using the full-period forecasting error, which includes the waiting period.

Results. Figure 2 compares RETICLE's selections with the MSE improvement per additional pair across six downstream learners. RETICLE passes within the candidate range in all 30 synthetic runs and all 16 real cases, without forced selection at the largest checkpoint. Across all six dataset families, the selected checkpoints lie after the transition from large initial per-sample gains to gains that are small relative to their early-checkpoint magnitudes. The synthetic systems show the transition clearly; the real datasets show the same pattern with variability across cases often narrowing substantially, and several WESAD learners exhibit early negative gains and fluctuations that largely subside before the selected region. One selection, obtained without access to downstream learners or their errors, is consistent with this behavior across heterogeneous learners. To complement the marginal gains in Figure 2 with attained error levels, Appendix L.1 reports the selected data sizes, the normalized MSE values at the initial checkpoint $n_0 = 64$ and the selected checkpoint, and the corresponding paired percentage reductions.

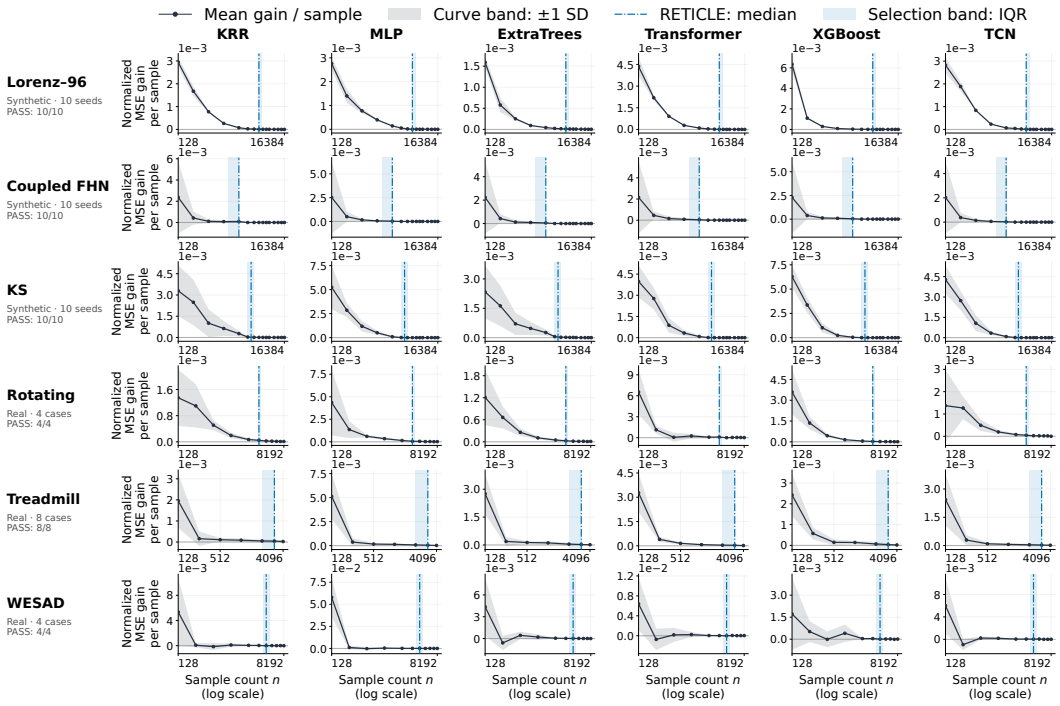

Figure 2: **Normalized MSE gain per sample and RETICLE-selected checkpoints.** Curves and gray bands: mean ± 1 SD of the decrease in normalized MSE per added one-step probe pair (actual count) between adjacent checkpoints, computed within each seed (synthetic) or case (real); positive values indicate improvement, and vertical scales differ across panels. Values are plotted at interval endpoints on a log axis from $n = 128$. Blue dash-dotted lines and bands: median and IQR of the first passing checkpoint among passing runs, shared across all six learners; PASS counts accompany dataset names.

5.2 Continuous Online Adaptation

Setup. We evaluate ETTm1 (Zhou et al., 2021), an eight-channel Jena Weather task[1], and Tetouan City electricity demand (Salam & El Hibaoui, 2018) with MLP, Transformer, and TCN forecasters and Page–Hinkley or ADWIN detection, giving 18 dataset–model–detector conditions. Each series is split chronologically into 40% training, 30% validation, and 30% testing, with joint forecasting over $H_{\text{out}} = 48$ steps. We compare four policies: RETICLE, CALIPER, a fixed waiting size, and no update, which retains the initial model. Online policy tuning uses downstream validation errors; once configured, each RETICLE audit uses only observed data and local-probe losses. Policy settings are tuned on validation per condition and then fixed; RETICLE uses the smallest eligible c whose validation MSE is within a relative 0.1% of the best eligible value (eligibility rules in Appendix K). For each forecaster, all policies start testing from the same training-fitted model. Updating policies share training-based detector calibration and one epoch of all-layer fine-tuning but keep independent detection and update states. The encoder is refitted on training plus validation data before testing and then frozen. RETICLE and CALIPER force an update if their gates have not passed by 4,096 one-step pairs. We report normalized MSE over the full test period, including waiting and update periods, with scales fixed from the initial fitting data. This score charges the cost of waiting, which the checkpoint study omits by design.

Results. Table 1 reports all 18 conditions. RETICLE attains the lowest MSE in 12 and ties for the lowest in two; validation-selected fixed waiting wins the remaining four, including three on Weather. RETICLE has the lowest overall mean rank (1.28), with mean per-condition MSE reductions of

[1]Jena Weather data are available at https://www.bgc-jena.mpg.de/wetter/weather_data.html.

Table 1: Normalized MSE over the full test period ($H_{\text{out}} = 48$; lower is better), with within-row ranks in parentheses. Best values are bold; ties receive average ranks. Fixed waiting sizes and CALIPER settings are validation-tuned under the common fine-tuning protocol. Mean ranks cover six conditions per dataset (18 overall).

Dataset	Detector	Model	RETICLE	CALIPER	Fixed	No update
ETTm1	Page–Hinkley	MLP	**1.9701 (1)**	2.5593 (3)	2.2667 (2)	5.7528 (4)
		Transformer	**1.5263 (1)**	2.1202 (3)	1.5721 (2)	3.9816 (4)
		TCN	**1.3127 (1)**	1.7922 (3)	1.3218 (2)	3.3399 (4)
	ADWIN	MLP	**3.6411 (1.5)**	4.0664 (3)	**3.6411 (1.5)**	5.7528 (4)
		Transformer	**1.6009 (1)**	1.8240 (3)	1.7767 (2)	3.9816 (4)
		TCN	1.5007 (2)	1.6063 (3)	**1.4524 (1)**	3.3399 (4)
	Mean rank (6/6)		**1.25**	3.00	1.75	4.00
Weather	Page–Hinkley	MLP	**2.7915 (1)**	3.0024 (3)	2.8297 (2)	3.7029 (4)
		Transformer	2.1008 (2)	2.1290 (3)	**2.0354 (1)**	2.7109 (4)
		TCN	**1.9984 (1)**	2.0849 (3)	2.0094 (2)	2.4751 (4)
	ADWIN	MLP	**3.0661 (1)**	3.1735 (2.5)	3.1735 (2.5)	3.7029 (4)
		Transformer	2.2278 (2)	2.2982 (3)	**2.2248 (1)**	2.7109 (4)
		TCN	2.2081 (2)	2.2171 (3)	**2.0622 (1)**	2.4751 (4)
	Mean rank (6/6)		**1.50**	2.92	1.58	4.00
Tetouan City	Page–Hinkley	MLP	**0.2917 (1)**	0.3555 (3)	0.3090 (2)	1.1593 (4)
		Transformer	**0.3854 (1)**	0.4595 (3)	0.4116 (2)	1.1919 (4)
		TCN	**0.2942 (1)**	0.3297 (3)	0.3232 (2)	0.9785 (4)
	ADWIN	MLP	**0.4753 (1)**	0.6149 (2.5)	0.6149 (2.5)	1.1593 (4)
		Transformer	**0.4847 (1)**	0.6089 (2.5)	0.6089 (2.5)	1.1919 (4)
		TCN	**0.4527 (1.5)**	**0.4527 (1.5)**	0.6225 (3)	0.9785 (4)
	Mean rank (6/6)		**1.08**	2.58	2.33	4.00
Overall mean rank (18/18)			**1.28**	2.83	1.89	4.00

6.08% relative to fixed waiting and 11.91% relative to CALIPER. Because fixed waiting is tuned on validation over sizes up to 4,096, the comparison already includes the option of waiting longer; RETICLE's advantage comes from adapting the trade-off to each episode rather than from any single waiting size.

The policies differ in how they assess the accumulated data. A single waiting size tuned on validation need not suit every test episode: fixed waiting updates without examining the current data and may therefore act on too little data in some episodes and too late in others. CALIPER adapts its waiting size but builds neighborhoods in the observed state space, where proximity need not reflect predictive similarity, and its monotonicity requirement can block gate passage when the risk profile is quasiconvex but not monotone. RETICLE addresses both limitations with a predictive representation of histories, which aligns neighborhoods with prediction rather than raw state, and a confidence-based criterion that accommodates approximately quasiconvex profiles, so gate-triggered updates adapt the waiting size to each episode even after c is fixed. The comparison evaluates the complete procedures rather than isolating individual components. Appendix L.2 details the aggregation of the per-condition MSE reductions and reports the validation-selected fixed sizes and c, mean checkpoint sizes, and counts of PASS-triggered and forced fine-tuning starts.

6 Conclusion

RETICLE selects retraining data sizes from local probe risks without accessing the downstream learner once configured. It uses approximate quasiconvexity on a finite bandwidth grid, a necessary consequence of local-learning regularity, as a data-quality criterion; causal auditing certifies this shape simultaneously over declared checkpoints and representations under temporal dependence. RETICLE attained the best overall mean rank (1.28) in online forecasting, with gains varying by dataset and detector. These results support using local risk structure as a data-quality criterion for adaptive retraining in streaming time series.

AI DISCLOSURE SECTION

In this work, we used generative AI tools to generate synthetic datasets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in writing proofs, implement methods, clean and reformat datasets, provide feedback on research methodology and experiments (including suggestions for additional analyses and ablations during manuscript revision), assist in interpreting results (reviewing and rephrasing the authors' interpretation of experimental outcomes), and assist with translation. We did not use generative AI tools to help develop theoretical models or to propose or refine hypotheses or conceptual frameworks; support for qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to draft parts of this paper, search for information, identify relevant literature, edit the paper for readability, and format references. We have reviewed all AI-assisted work: all AI-generated code was verified and tested for correctness by a first author, all AI-assisted proofs and manuscripts were checked line by line by the authors, and all AI-suggested analyses and interpretations were independently evaluated by the authors and, where adopted, verified against our own experimental records. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Oren Anava, Elad Hazan, Shie Mannor, and Ohad Shamir. Online Learning for Time Series Prediction. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 172–184. PMLR, 2013.

Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, Lorenzo Stella, François-Xavier Aubet, Laurent Callot, and Tim Januschowski. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Comput. Surv.*, 55(6), 2022. ISSN 0360-0300. doi: 10.1145/3533382.

Albert Bifet and Ricard Gavaldà. Learning from Time-Changing Data with Adaptive Windowing. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pp. 443–448, 2007. doi: 10.1137/1.9781611972771.42.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

M. Casdagli. Nonlinear Prediction of Chaotic Time Series. *Physica D: Nonlinear Phenomena*, 35 (3):335–356, 1989. doi: 10.1016/0167-2789(89)90074-2.

Naoki Chihara, Yasuko Matsubara, Ren Fujiwara, and Yasushi Sakurai. Modeling time-evolving causality over data streams. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 153–164. Association for Computing Machinery, 2025. doi: 10.1145/3690624.3709283.

Naoki Chihara, Ren Fujiwara, Yasuko Matsubara, and Yasushi Sakurai. AdaKoop: Efficient modeling of nonlinear dynamics from nonstationary data streams with Koopman operator regression. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 628–639. Association for Computing Machinery, 2026. doi: 10.1145/3770855.3817851.

Yo Joong Choe and Aaditya Ramdas. Comparing Sequential Forecasters. *Operations Research*, 72 (4):1368–1387, 2024. doi: 10.1287/opre.2021.0792. Published online in 2023; issue publication in 2024.

William S. Cleveland and Susan J. Devlin. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988. doi: 10.1080/01621459.1988.10478639.

A. P. Dawid. Present Position and Potential Developments: Some Personal Views. Statistical Theory. The Prequential Approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2): 278–290, 1984. doi: 10.2307/2981683.

Francis X. Diebold and Roberto S. Mariano. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995. doi: 10.1080/07350015.1995.10524599.

Jianqing Fan. Design-adaptive Nonparametric Regression. *Journal of the American Statistical Association*, 87(420):998–1004, 1992. doi: 10.1080/01621459.1992.10476255.

Jianqing Fan and Irene Gijbels. Variable Bandwidth and Local Linear Regression Smoothers. *The Annals of Statistics*, 20(4):2008–2036, 1992. doi: 10.1214/aos/1176348900.

Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications*. Number 66 in Monographs on Statistics and Applied Probability. Chapman and Hall, 1996. doi: 10.1201/9780203748725.

J. Doyne Farmer and John J. Sidorowich. Predicting Chaotic Time Series. *Physical Review Letters*, 59(8):845–848, 1987. doi: 10.1103/PhysRevLett.59.845.

David A. Freedman. On Tail Probabilities for Martingales. *The Annals of Probability*, 3(1):100–118, 1975. doi: 10.1214/aop/1176996452.

Ren Fujiwara, Yasuko Matsubara, and Yasushi Sakurai. Modeling latent non-linear dynamical system over time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11663–11671, 2025. doi: 10.1609/aaai.v39i11.33269.

Ren Fujiwara, Yasuko Matsubara, and Yasushi Sakurai. When to Retrain after Drift: A Data-Only Test of Post-Drift Data Size Sufficiency. In *International Conference on Learning Representations*, 2026a.

Ren Fujiwara, Yasuko Matsubara, and Yasushi Sakurai. Modeling dynamic mixtures of time-delay systems from streaming time series. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2617–2625, 2026b. doi: 10.24963/ijcai.2026/291.

João Gama, Pedro Medas, Gladys Castillo, and Pedro Rodrigues. Learning with Drift Detection. In *Advances in Artificial Intelligence – SBIA 2004*, pp. 286–295, 2004. doi: 10.1007/978-3-540-28645-5_29.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A Survey on Concept Drift Adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014. doi: 10.1145/2523813.

Raffaella Giacomini and Halbert White. Tests of Conditional Predictive Ability. *Econometrica*, 74 (6):1545–1578, 2006. doi: 10.1111/j.1468-0262.2006.00718.x.

Heitor Murilo Gomes, Albert Bifet, Jesse Read, Jean Paul Barddal, Fabrício Enembreck, Bernhard Pfahringer, Geoff Holmes, and Talel Abdessalem. Adaptive Random Forests for Evolving Data Stream Classification. *Machine Learning*, 106:1469–1495, 2017. doi: 10.1007/s10994-017-5642-8.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the Non-stationarity in Time Series Forecasting via Test-time Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17868–17876, 2025. doi: 10.1609/aaai.v39i17.33965.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift. In *International Conference on Learning Representations*, 2022.

Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Testable Learning with Distribution Shift. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 2887–2943. PMLR, 2024.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 2101–2109, 2017. doi: 10.1609/aaai.v31i1.10779.

Dominik Krumm, Daniel Koska, Jayant Wakode, Stephan Odenwald, and Christian Maiwald. Long-duration kinetic and kinematic bipedal locomotion data on a treadmill at different speed. Open Science Framework, 2022.

Dominik Krumm, Daniel Koska, Jayant Wakode, Stephan Odenwald, and Christian Maiwald. Multimodal data on bipedal locomotion during prolonged treadmill recordings at varying speeds. *Scientific Data*, 13:761, 2026. doi: 10.1038/s41597-026-07445-3.

Ying-yee Ava Lau, Zhiwen Shao, and Dit-Yan Yeung. Fast and Slow Streams for Online Time Series Forecasting Without Information Leakage. In *International Conference on Learning Representations*, 2025.

O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal Spatial Adaptation to Inhomogeneous Smoothness: An Approach Based on Kernel Estimates with Variable Bandwidth Selectors. *The Annals of Statistics*, 25(3):929–947, 1997. doi: 10.1214/aos/1069362731.

Zijian Li, Changze Zhou, Minghao Fu, Sanjay Manjunath, Fan Feng, Guangyi Chen, Yingyao Hu, Ruichu Cai, and Kun Zhang. Online time series forecasting with theoretical guarantees. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379 (2194):20200209, 02 2021. ISSN 1364-503X. doi: 10.1098/rsta.2020.0209.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9881–9893, 2022. doi: 10.52202/068431-0718.

Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. Koopa: Learning non-stationary time series dynamics with Koopman predictors. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Quang Pham, Chenghao Liu, Doyen Sahoo, and Steven C. H. Hoi. Learning Fast and Slow for Online Time Series Forecasting. In *International Conference on Learning Representations*, 2023.

Christoph Raab, Moritz Heusinger, and Frank-Michael Schleif. Reactive Soft Prototype Computing for Concept Drift Streams. *Neurocomputing*, 416:340–351, 2020. doi: 10.1016/j.neucom.2019.11.111.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

Syama Sundar Rangapuram, Matthias W. Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yulia Rubanova, Ricky T. Q. Chen, and David K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ronitt Rubinfeld and Arsen Vasilyan. Testing Distributional Assumptions of Learning Algorithms. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 1643–1656. Association for Computing Machinery, 2023. doi: 10.1145/3564246.3585117.

D. Ruppert and M. P. Wand. Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics*, 22(3):1346–1370, 1994. doi: 10.1214/aos/1176325632.

Abdulwahed Salam and Abdelaaziz El Hibaoui. Power consumption of Tetouan City. UCI Machine Learning Repository, 2018.

Juan José Saucedo-Dorantes, Miguel Delgado-Prieto, Joan Valls Pérez, and David Alejandro Elvira-Ortiz. Rotating electromechanical system dataset for condition monitoring. *Scientific Data*, 13: 873, 2026. doi: 10.1038/s41597-026-07224-0.

Philip Schmidt, Attila Reiss, Robert Dürichen, Claus Marberger, and Kristof Van Laerhoven. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, pp. 400–408. Association for Computing Machinery, 2018. doi: 10.1145/3242969.3242985.

George Sugihara. Nonlinear Forecasting for the Classification of Natural Time Series. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 348 (1688):477–495, 1994. doi: 10.1098/rsta.1994.0106.

George Sugihara and Robert M. May. Nonlinear Forecasting as a Way of Distinguishing Chaos from Measurement Error in Time Series. *Nature*, 344(6268):734–741, 1990. doi: 10.1038/344734a0.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Yi-Fan Zhang, Qingsong Wen, Xue Wang, Weiqi Chen, Liang Sun, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. OneNet: Enhancing Time Series Forecasting Models under Concept Drift by Online Ensembling. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3066.

Lifan Zhao and Yanyan Shen. Proactive Model Adaptation Against Concept Drift for Online Time Series Forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 2020–2031. Association for Computing Machinery, 2025. doi: 10.1145/3690624.3709210.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115. AAAI Press, 2021. doi: 10.1609/aaai.v35i12.17325.

APPENDIX

A NOTATION

Table 2 summarizes the main notation used in the method and theoretical analysis. Symbols used only in individual proofs are defined locally. All probe risks below concern clipped one-step prediction losses.

Table 2: Main notation for RETICLE.

Symbol	Meaning
Stream and audit setup	
$t,\ n$	Prediction time; checkpoint size (number of matured input–response pairs).
$\mathbf{z}_t,\ \mathbf{y}_{t+1}$	Latent state from the frozen encoder in $\mathbb{R}^d$ and scaled next-step response in $\mathbb{R}^p$.
$\mathcal{F}_t$	Information available when the prediction at time t is issued.
$\mathbb{N},\ K,\ R_{\mathrm{rep}}$	Declared checkpoints, their number, and the number of declared representations.
$\mathbb{T}_n,\ s_n$	Candidate audit times at checkpoint n and the declared audit start.
$\mathbb{I}_t,\ m_t,\ g$	Reference indices $\{i \in \mathbb{Z}_{\geq 0} : i < t - g\}$, their count, and the past-exclusion gap.
$\mathbb{H}_{\mathrm{f}},\ J$	Finite bandwidth grid $\{h_1 > \cdots > h_J > 0\}$ and its size.
Local probes and weighted risks	
$\widehat{\mathbf{f}}_{t,h},\ \rho$	Causal local affine prediction at bandwidth h and ridge penalty on its slopes.
$\ell_{t,h},\ M$	Clipped loss $\min\{p^{-1}\|\mathbf{y}_{t+1} - \widehat{\mathbf{f}}_{t,h}\|^2, M\}$ and clipping level.
S_t	Predictable indicator that every finite bandwidth satisfies the local support conditions.
$\mathrm{w}_{t,n},\ \mathrm{W}_n$	Common audit weight $m_t / \max(1, n-1-g)$ and total weight $\sum_{t\in\mathbb{T}_n} \mathrm{S}_t \mathrm{w}_{t,n}$.
$\mathrm{R}_n(h),\ \widehat{\mathrm{R}}_n(h)$	Weighted conditional probe risk and its estimate from realized losses, respectively.
Shape assessment and certification	
$\mathbf{r}_n$	Conditional risk profile $(\mathrm{R}_n(h_1), \ldots, \mathrm{R}_n(h_J))$.
$V(\boldsymbol{r}),\ \mathbb{V}_J$	Signed quasiconvexity score and the set of quasiconvex profiles on the bandwidth grid, including monotone profiles.
$[\mathrm{L}_{n,ij}, \mathrm{U}_{n,ij}]$	Simultaneous interval for $\mathrm{R}_n(h_j) - \mathrm{R}_n(h_i)$, where $i < j$.
$\mathrm{L}_{V,n},\ \mathrm{U}_{V,n}$	Lower and upper confidence bounds on $V(\mathbf{r}_n)$.
$\mathrm{N}_{t,h},\ \mathrm{N}_{\min,n}$	Local effective sample size and its minimum over supported audit times and finite bandwidths.
$\eta_n,\ c$	Operational shape tolerance $\eta_n = c/\sqrt{\mathrm{N}_{\min,n}}$ and its scaling coefficient.
Adeq_n	Audit adequacy: reference-count, support, and representation checks, positive total weight, and a valid tolerance.
τ	First adequate passing checkpoint: $\inf\{n \in \mathbb{N} : \mathsf{PASS}_n\}$, with $\inf \varnothing = \infty$.
$\alpha,\ \mathcal{E}$	Nominal error level for the declared audit family and the simultaneous coverage event.
Structural condition and precision	
$\boldsymbol{f},\ \boldsymbol{\epsilon}_{t+1}$	Fixed latent prediction rule and innovation in $\mathbf{y}_{t+1} = \boldsymbol{f}(\mathbf{z}_t) + \boldsymbol{\epsilon}_{t+1}$.
$\mathbf{b}_{t,h},\ \boldsymbol{\xi}_{t,h}$	Probe approximation error relative to $\boldsymbol{f}$, and smoothed reference residual.
$\mathcal{G}_{n,\delta},\ \delta$	Weighted local-learning regularity event and its structural remainder tolerance.
$Q_n(h)$	Ideal profile $a_n + A_n h^4 + C_n h^{-\nu}$, with $a_n \in \mathbb{R}$, $A_n, C_n \geq 0$, and $\nu > 0$; ν need not equal d.
$\kappa_n(h)$	Weighted conditional loss removed by clipping.
$\varrho_n,\ \Delta_n$	Bound on selected pairwise interval radii before range intersection, and margin $\Delta_n = \eta_n - V(\mathbf{r}_n)$.

B RETICLE ALGORITHM

Algorithm 1 summarizes RETICLE for one monitoring episode. The encoder and normalization scales are frozen before audit queries are issued, and each checkpoint n counts matured one-step probe pairs. The procedure returns the first adequate checkpoint whose upper confidence bound on the quasiconvexity score is within the tolerance, without accessing a downstream model.

Algorithm 1 RETICLE checkpoint selection

Require: Latent states and one-step targets $\{(\mathbf{z}_t, \mathbf{y}_{t+1})\}$
Require: Checkpoint grid $\mathbb{N}$, bandwidth grid $\mathbb{H}_{\mathrm{f}}$, and α, c
Ensure: First passing checkpoint τ, or ∞ if none passes
for $n \in \mathbb{N}$ in increasing order **do**
  Evaluate local one-step probes on the scheduled audit queries, using past reference data at every bandwidth.
  Form clipped losses on queries supported at every bandwidth.
  Compute $\mathrm{N}_{\min,n}$ and set $\eta_n \leftarrow c/\sqrt{\mathrm{N}_{\min,n}}$ if $0 < \mathrm{N}_{\min,n} < \infty$; otherwise set $\eta_n \leftarrow 0$.
  if the adequacy condition Adeq_n fails **then**
    continue
  end if
  Construct simultaneous intervals for pairwise risk differences using the common queries and weights.
  Obtain the resulting upper bound $\mathrm{U}_{V,n}$ on $V(\mathbf{r}_n)$.
  if $\mathrm{U}_{V,n} \le \eta_n$ **then**
    return n
  end if
end for
return ∞

The adequacy checks and simultaneous paired intervals are specified in Appendix I and Appendix C, respectively. The intervals share the confidence budget across the declared audit family. A violation or unresolved result at one checkpoint does not stop later audits. If no checkpoint passes, the procedure returns $\tau = \infty$.

C Exact weighted intervals used by the implementation

For one pair $1 \le i < j \le J$, let

$$\mathrm{D}_{t,ij} = \ell_{t,h_j} - \ell_{t,h_i}, \quad \mathrm{d}_{n,ij} = \mathrm{R}_n(h_j) - \mathrm{R}_n(h_i), \quad \widehat{\mathrm{d}}_{n,ij} = \mathrm{W}_n^{-1} \sum_{t \in \mathbb{T}_n} \mathrm{S}_t \mathrm{w}_{t,n} \mathrm{D}_{t,ij}. \tag{19}$$

Define the predictable forecast distance and loss-difference bound by

$$\mathrm{e}_{t,ij} = p^{-1/2} \|\widehat{\mathbf{f}}_{t,h_j} - \widehat{\mathbf{f}}_{t,h_i}\|, \qquad \zeta_{t,ij} = M - (\sqrt{M} - \mathrm{e}_{t,ij})_+^2. \tag{20}$$

The latter satisfies $|\mathrm{D}_{t,ij}| \le \zeta_{t,ij} \le M$ for every possible future response.

Before auditing, fix a finite grid $\mathbb{B} \subset (0, M]$ containing M for each declared audit family. The range and common tuning grids are

$$\begin{aligned} \mathbb{B}_{\mathrm{checkpoint}} &= M\{1/32, 1/16, 1/8, 1/4, 1/2, 1\}, \\ \mathbb{B}_{\mathrm{online}} &= M\{1/256, 1/128, 1/64, 1/32, 1/16, 1/8, 1/4, 1/2, 1\}, \\ B &= |\mathbb{B}|, \qquad \mathbb{L} = \{0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}. \end{aligned} \tag{21}$$

The checkpoint experiments use $\mathbb{B}_{\mathrm{checkpoint}}$ ($B = 6$); online adaptation uses $\mathbb{B}_{\mathrm{online}}$ ($B = 9$). Set

$$P = \binom{J}{2}, \qquad \Gamma = \log \frac{4PKR_{\mathrm{rep}}B}{\alpha}, \qquad \psi(\lambda) = -\log(1-\lambda) - \lambda. \tag{22}$$

Here B counts all declared candidates, including those that stop. The factor four reserves half of the nominal budget for auxiliary marginal-risk intervals: the interval backend retains one for $\mathrm{R}_n(h_j)$ in each pair (i, j), with repeats across pairs. These intervals enter neither the gate decision nor the reported results. Theorem 1 covers only paired differences; all $P = 91$ pairs in the main $J = 14$ grid share this allocation.

The implementation takes at most 64 of the most recent candidate calibration queries whose labels mature by the audit start, then retains those supported at every bandwidth. Let $\mathbb{C}_n$ denote these

retained queries, and let $A_{2,n} = \sum_{t\in\mathbb{T}_n} \mathrm{w}_{t,n}^2$ be the planned squared-weight sum. The implemented calibration proxy is

$$\widehat{\mathrm{Q}}_{\mathrm{pre},n,ij,b} = A_{2,n} \begin{cases} \min\{1,\ |\mathbb{C}_n|^{-1} \sum_{s\in\mathbb{C}_n} (\mathrm{D}_{s,ij}/b)^2\}, & |\mathbb{C}_n| > 0, \\ 1, & |\mathbb{C}_n| = 0. \end{cases}$$

For each (n, i, j, b) select $\lambda_{n,ij,b} \in \arg\min_{\lambda\in\mathbb{L}} (\Gamma + \psi(\lambda)\widehat{\mathrm{Q}}_{\mathrm{pre},n,ij,b})/\lambda$ before auditing, using the first grid value to break ties, and hold it fixed within the audit. The proxy uses the planned sum rather than the subsequently observed support set. It tunes interval width, and need not upper-bound the future quadratic statistic for coverage to hold.

A range candidate b is stopped at the first supported prediction whose predictable bound exceeds b. Only candidates surviving to the audit endpoint are eligible for reporting. For a survivor, compute

$$\mathrm{Q}_{n,ij,b} = \sum_{t\in\mathbb{T}_n} \mathrm{S}_t \mathrm{w}_{t,n}^2 (\mathrm{D}_{t,ij}/b)^2, \qquad \mathrm{r}_{n,ij,b} = \min\left\{2b,\ \frac{b\{\Gamma + \psi(\lambda_{n,ij,b})\mathrm{Q}_{n,ij,b}\}}{\lambda_{n,ij,b}\mathrm{W}_n}\right\}. \tag{23}$$

This is an uncentered quadratic statistic; the paired interval does not use a sample variance or an EWMA-centered residual. Choose a surviving $\widehat{\mathrm{b}}$ with the smallest radius and report

$$\mathrm{L}_{n,ij} = \max\{-\widehat{\mathrm{b}}, \widehat{\mathrm{d}}_{n,ij} - \mathrm{r}_{n,ij,\widehat{\mathrm{b}}}\}, \qquad \mathrm{U}_{n,ij} = \min\{\widehat{\mathrm{b}}, \widehat{\mathrm{d}}_{n,ij} + \mathrm{r}_{n,ij,\widehat{\mathrm{b}}}\}. \tag{24}$$

The candidate $b = M$ always survives. Selection across b uses a simultaneous range-candidate guarantee; selection across λ requires no additional union bound because it occurs before auditing.

D Proof of finite-family coverage and certification

D.1 Predictable range bound

Scale vectors by $\sqrt{p}$ and write $a = \sqrt{M}$ and $e = \|\widehat{\mathbf{f}}_i - \widehat{\mathbf{f}}_j\|/\sqrt{p}$, the quantity $\mathrm{e}_{t,ij}$ in equation (20). For a response at distance r from the closer forecast, the other distance is at most $r + e$. For $r \geq a$ both clipped squared distances equal M. For $r < a$, their difference is at most $\min\{(r+e)^2, a^2\} - r^2$. When $e < a$, this expression is maximized at $r = a - e$ and equals $a^2 - (a - e)^2$; when $e \geq a$, it is at most a^2. The same argument with forecasts interchanged proves equation (20).

D.2 One pair and one range candidate

Fix an audit, a representation, a pair, and a candidate b; suppress their indices. Write its nonempty candidate times as $t_1 < \cdots < t_T$ and condition on the pre-audit sigma-field $\mathcal{F}_{s_n}$, where $s_n \leq t_1$ is its declared start. The schedule, endpoint, and selected λ are fixed under this conditioning. Define the predictable stopped multiplier

$$\mathrm{C}_t = \mathrm{S}_t \prod_{\substack{s\in\mathbb{T}_n \\ s\leq t}} (1 - \mathrm{S}_s \mathbf{1}\{\zeta_s > b\}), \qquad \mathrm{X}_t = \mathrm{C}_t \mathrm{w}_t \mathrm{D}_t/b, \qquad \mu_t = \mathbb{E}[\mathrm{X}_t \mid \mathcal{F}_t]. \tag{25}$$

The bounds $t \leq n - 1$ and $m_t \leq \max(0, n - 1 - g)$ imply $0 \leq \mathrm{w}_t \leq 1$. Moreover, $\mathrm{C}_t \in \{0, 1\}$ and $\mathrm{C}_t \neq 0$ implies $|\mathrm{D}_t| \leq \zeta_t \leq b$, so $|\mathrm{X}_t| \leq 1$. For $|x| \leq 1$ and $0 < \lambda < 1$,

$$\log(1 \pm \lambda x) \geq \pm\lambda x - \psi(\lambda) x^2. \tag{26}$$

Indeed, in the logarithmic series the sum of the absolute values of all terms of degree at least two is bounded by $x^2 \sum_{k\geq 2} \lambda^k/k = \psi(\lambda)x^2$. Consequently,

$$\begin{aligned} &\mathbb{E}\big[\exp\{\pm\lambda(\mathrm{X}_t - \mu_t) - \psi(\lambda)\mathrm{X}_t^2\} \mid \mathcal{F}_t\big] \\ &\qquad \leq e^{\mp\lambda\mu_t}(1 \pm \lambda\mu_t) \leq 1. \end{aligned} \tag{27}$$

Since responses are one-step ($H = 1$), the factor at t_k is $\mathcal{F}_{t_k+1}$-measurable and $\mathcal{F}_{t_k+1} \subseteq \mathcal{F}_{t_{k+1}}$ because the issue times are distinct integers. Iterated conditioning therefore makes the products

nonnegative supermartingales. Conditional Markov's inequality at the declared endpoint, followed by the tower property, gives, for either sign,

$$\mathbb{P}\left\{\pm\lambda\sum_t(\mathrm{X}_t-\mu_t)>\Gamma+\psi(\lambda)\sum_t\mathrm{X}_t^2\right\}\le e^{-\Gamma}. \tag{28}$$

This self-normalized argument uses the same exponential-inequality foundation as confidence-sequence constructions (Howard et al., 2021); the stopped weights and finite-family allocation above are specified here for RETICLE.

For a surviving candidate, $\mathrm{C}_t=\mathrm{S}_t$ at every audit time, so

$$\sum_t(\mathrm{X}_t-\mu_t)=\frac{\mathrm{W}_n}{b}(\widehat{\mathrm{d}}_{n,ij}-\mathrm{d}_{n,ij}),\qquad\sum_t\mathrm{X}_t^2=\mathrm{Q}_{n,ij,b}. \tag{29}$$

Dividing by the realized positive W_n yields the untruncated radius in equation (23). Both the observed and conditional mean differences lie in $[-b,b]$ for a survivor. Capping the radius at $2b$ and intersecting the interval with $[-b,b]$ therefore preserve coverage. The proof incorporates the random support and denominator as predictable multipliers and does not condition retrospectively on a realized support set.

D.3 Simultaneous coverage and the decision

Union bounding over two tails, P finite-bandwidth pairs, K checkpoints, R_{rep} representations, and B candidates gives

$$2PKR_{\mathrm{rep}}Be^{-\Gamma}=\alpha/2\le\alpha. \tag{30}$$

Thus the paired-difference event actually has probability at least $1-\alpha/2$; the stated $1-\alpha$ guarantee is a conservative nominal guarantee consistent with the retained implementation budget. The auxiliary marginal intervals of Appendix C enter no result here. All surviving paired intervals cover simultaneously, including the selected minimum-radius candidates. On this event, each difference $\mathrm{R}_j-\mathrm{R}_i$ belongs to $[\mathrm{L}_{ij},\mathrm{U}_{ij}]$, and $\mathrm{R}_j-\mathrm{R}_k$ belongs to $[-\mathrm{U}_{jk},-\mathrm{L}_{jk}]$. Coordinatewise monotonicity of maximum and minimum proves equation (15). Equation (16) follows from equation (15), the decision rule, and Lemma 1; equation (17) follows from equation (15) and the decision rule. Under Assumption 1(i), the contrapositive of equation (14) in Proposition 1 further yields $\mathcal{G}^c_{n,\eta_n/2}$. Taking the union over checkpoints includes the first passing checkpoint and proves equation (18). The $1-\alpha$ statements are simultaneous frequentist guarantees, not posterior probabilities conditional on receiving a passing result. For a single fixed one-step audit process, Ville's inequality also controls all of its prefixes. This does not remove the K-checkpoint allocation here: different checkpoints may use different audit starts, normalization, and pre-audit tuning, and are not prefixes of a single fixed process.

E Quasiconvexity and precision

Lemma 1 (Distance to a quasiconvex profile)**.** *For every $\boldsymbol{r}\in\mathbb{R}^J$,*

$$V(\boldsymbol{r})\le 0\iff\boldsymbol{r}\in\mathbb{V}_J,\qquad\mathrm{dist}_\infty(\boldsymbol{r},\mathbb{V}_J)=\tfrac{1}{2}[V(\boldsymbol{r})]_+. \tag{31}$$

Lemma 2 (Conditional finite-sample precision)**.** *On $\mathcal{E}$, let ϱ_n bound all selected finite-bandwidth pair radii in equation (23), before intersection with the candidate range. Then*

$$\mathrm{U}_{V,n}\le V(\mathbf{r}_n)+2\varrho_n. \tag{32}$$

An adequate checkpoint passes if $\Delta_n:=\eta_n-V(\mathbf{r}_n)>0$ and $2\varrho_n\le\Delta_n$.

E.1 Proof of Lemma 1

If $\boldsymbol{r}\in\mathbb{V}_J$, each triple satisfies $r_j\le\max(r_i,r_k)$. Conversely, suppose every triple satisfies that inequality, and choose a global minimum index v. For $i<j<v$, $r_j\le\max(r_i,r_v)=r_i$; the indices after v are nondecreasing by the symmetric argument. Hence $\boldsymbol{r}\in\mathbb{V}_J$.

If $\boldsymbol{q} \in \mathbb{V}_J$ and $\|\boldsymbol{r} - \boldsymbol{q}\|_\infty \le \epsilon$, then $r_j \le \max(r_i, r_k) + 2\epsilon$. Thus $\mathrm{dist}_\infty(\boldsymbol{r}, \mathbb{V}_J) \ge [V(\boldsymbol{r})]_+/2$. For the reverse inequality, put

$$q_j = \max\left\{\min_{i\le j} r_i, \min_{k\ge j} r_k\right\}, \qquad a = [V(\boldsymbol{r})]_+. \tag{33}$$

Before a global minimum, $\boldsymbol{q}$ is the prefix minimum; after it, $\boldsymbol{q}$ is the suffix minimum. Thus $\boldsymbol{q} \in \mathbb{V}_J$. Moreover $0 \le r_j - q_j \le a$. If the difference is positive, the corresponding prefix and suffix minimizers lie strictly on either side of j, and their triple realizes $r_j - q_j$. Therefore $\boldsymbol{q} + (a/2)\mathbf{1}$ is quasiconvex and within $a/2$ of $\boldsymbol{r}$, proving equation (31). Restricting the approximating profile to $[0, M]^J$ gives the same distance for $\boldsymbol{r} \in [0, M]^J$, because coordinatewise clipping preserves quasiconvexity and cannot increase that distance.

E.2 Proof of Lemma 2

On simultaneous coverage, every finite pair obeys $|\widehat{\mathrm{d}}_{n,ij} - \mathrm{d}_{n,ij}| \le \varrho_n$. The maximum of the minima of pair contrasts is 1-Lipschitz in the sup norm on contrast coordinates. Furthermore each reported upper contrast is at most its empirical contrast plus ϱ_n. Consequently

$$\mathrm{U}_{V,n} \le V(\widehat{\mathbf{r}}_n) + \varrho_n \le V(\mathbf{r}_n) + 2\varrho_n. \tag{34}$$

The decision rule then proves the passage statement.

E.3 An explicit sufficient effective-weight condition

For each finite pair choose any surviving candidate with $\mathrm{Q}_{n,ij,b}/\mathrm{W}_n \le \overline{q}_{ij,b}$ and put

$$\mathrm{F}_{ij,b} = \frac{b\psi(\lambda_{ij,b})\overline{q}_{ij,b}}{\lambda_{ij,b}}. \tag{35}$$

Here the checkpoint n is fixed and suppressed in $\lambda_{ij,b}$. Write b_{ij} for the candidate chosen for pair (i, j) and $\lambda_{ij} := \lambda_{ij,b_{ij}}$. On the simultaneous coverage event, if Adeq_n holds, $\Delta_n \ge \Delta > 0$, and every chosen candidate satisfies $\mathrm{F}_{ij,b_{ij}} < \Delta/2$, then it is sufficient that

$$\mathrm{W}_n \ge \max_{i<j} \frac{2b_{ij}\Gamma}{\lambda_{ij}(\Delta - 2\mathrm{F}_{ij,b_{ij}})}. \tag{36}$$

Indeed its radius is at most $b_{ij}\Gamma/(\lambda_{ij}\mathrm{W}_n) + \mathrm{F}_{ij,b_{ij}} \le \Delta/2$, and the reported minimum radius is no larger. This is a pathwise sufficient condition. In particular the reference count m_t alone does not ensure it. With fixed positive λ and nonvanishing Q/W_n, the second term in the radius can remain positive. The audit effective count $\mathrm{W}_n^2/\sum_t \mathrm{S}_t\mathrm{w}_{t,n}^2$ and the local kernel ESS in equation (11) measure different concentrations of weights.

F Structural decomposition and proof

Define

$$\sigma_n^2 = \frac{\sum_{t\in\mathbb{T}_n} \mathrm{S}_t\mathrm{w}_{t,n}\mathbb{E}[\|\boldsymbol{\epsilon}_{t+1}\|^2 \mid \mathcal{F}_t]}{p\mathrm{W}_n}. \tag{37}$$

The exact identity $\widehat{\mathbf{f}}_{t,h} - \boldsymbol{f}(\mathbf{z}_t) = \mathbf{b}_{t,h} + \boldsymbol{\xi}_{t,h}$ and the conditional centering of the new audit innovation give

$$\frac{1}{p}\mathbb{E}[\|\mathbf{y}_{t+1} - \widehat{\mathbf{f}}_{t,h}\|^2 \mid \mathcal{F}_t] = \frac{1}{p}\mathbb{E}[\|\boldsymbol{\epsilon}_{t+1}\|^2 \mid \mathcal{F}_t] + \frac{1}{p}\|\mathbf{b}_{t,h} + \boldsymbol{\xi}_{t,h}\|^2. \tag{38}$$

This step uses no independence between reference residuals and reference states. Averaging with the common predictable weights and using $\min(q, M) = q - (q - M)_+$ yields

$$\mathrm{R}_n(h) = \sigma_n^2 + \|\mathbf{b}_h + \boldsymbol{\xi}_h\|_{n,w}^2 - \kappa_n(h). \tag{39}$$

Writing $\boldsymbol{r}_h = \mathbf{b}_h - h^2\boldsymbol{b}^{(0)}$, Cauchy–Schwarz gives

$$\left|\|\mathbf{b}_h\|_{n,w}^2 - h^4\|\boldsymbol{b}^{(0)}\|_{n,w}^2\right| \le 2h^2\|\boldsymbol{b}^{(0)}\|_{n,w}\|\boldsymbol{r}_h\|_{n,w} + \|\boldsymbol{r}_h\|_{n,w}^2 \tag{40}$$

$$\le (2h^2\|\boldsymbol{b}^{(0)}\|_{n,w} + e_{n,h})e_{n,h}. \tag{41}$$

Expanding equation (39) and applying equation (3) proves equation (13) with $A_n = \|\boldsymbol{b}^{(0)}\|_{n,w}^2$ and $a_n = \sigma_n^2 - \kappa_n^{(0)}$. For $Q_n(h) = a_n + A_n h^4 + C_n h^{-\nu}$,

$$\frac{d^2}{du^2} Q_n(e^u) = 16 A_n e^{4u} + \nu^2 C_n e^{-\nu u} \geq 0. \tag{42}$$

Thus every finite ordered sample of this curve belongs to $\mathbb{V}_J$. The sup-norm approximation implies distance at most δ; equation (31) then implies $V(\mathbf{r}_n) \leq 2\delta$. If $A_n, C_n > 0$, its continuous optimum is $h_* = (\nu C_n/(4A_n))^{1/(4+\nu)}$. This optimum need not lie within the finite audit grid. If either coefficient is zero, a monotone ideal curve is permitted. Both monotone directions are deliberately included: the optimum may be outside the finite grid. A vanishing leading coefficient does not identify $\boldsymbol{f}$ as globally affine.

G Local geometry and ridge in the approximation condition

Let $\mathbf{x}_{ti} = (1, \mathbf{u}_{ti}^\top)^\top$, $\boldsymbol{e}_0 = (1, 0, \ldots, 0)^\top$, and

$$\mathbf{G}_{t,h} = \sum_i \pi_{ti,h} \mathbf{x}_{ti} \mathbf{x}_{ti}^\top + \rho \operatorname{diag}(0, \boldsymbol{I}_d). \tag{43}$$

For $\rho > 0$ and a nonempty reference set, $\mathbf{G}_{t,h}$ is positive definite. The smoother coefficients are

$$\mathrm{a}_{ti,h} = \pi_{ti,h} \boldsymbol{e}_0^\top \mathbf{G}_{t,h}^{-1} \mathbf{x}_{ti}, \qquad \sum_i \mathrm{a}_{ti,h} = 1. \tag{44}$$

The coefficients may be negative. Their first centered moment $\mathbf{g}_{t,h} = \sum_i \mathrm{a}_{ti,h} \mathbf{u}_{ti}$ need not vanish because the slopes are penalized. Set $\mathbf{M}_{t,h} = \sum_i \mathrm{a}_{ti,h} \mathbf{u}_{ti} \mathbf{u}_{ti}^\top$. Under Assumption 1(ii)–(iii), for response component a, Taylor's formula gives

$$\mathrm{b}_{t,h}^{(a)} = \nabla f_a(\mathbf{z}_t)^\top \mathbf{g}_{t,h} + \tfrac{1}{2} \nabla^2 f_a(\mathbf{z}_t) : \mathbf{M}_{t,h} + \mathrm{r}_{t,h}^{(a)}, \tag{45}$$

where the Hessian condition is $\|\nabla^2 f_a(\boldsymbol{z}) - \nabla^2 f_a(\boldsymbol{z}')\|_F \leq L_{\beta,a} \|\boldsymbol{z} - \boldsymbol{z}'\|^\beta$, and

$$|\mathrm{r}_{t,h}^{(a)}| \leq \frac{L_{\beta,a}}{(1+\beta)(2+\beta)} \sum_i |\mathrm{a}_{ti,h}| \|\mathbf{u}_{ti}\|^{2+\beta}. \tag{46}$$

The coefficient follows by integrating $\int_0^1 (1-s) s^\beta \, ds = 1/((1+\beta)(2+\beta))$ along each line segment. For bandwidth-independent matrices $\boldsymbol{M}_{t,0}$ set $b_t^{(0,a)} = \frac{1}{2} \nabla^2 f_a(\mathbf{z}_t) : \boldsymbol{M}_{t,0}$. If $\|\nabla f_a\| \leq G_a$ and $\|\nabla^2 f_a\|_F \leq L_{2,a}$, then

$$\begin{aligned} |\mathrm{b}_{t,h}^{(a)} - h^2 b_t^{(0,a)}| \leq G_a \|\mathbf{g}_{t,h}\| + \frac{L_{2,a}}{2} \|\mathbf{M}_{t,h} - h^2 \boldsymbol{M}_{t,0}\|_F \\ + \frac{L_{\beta,a}}{(1+\beta)(2+\beta)} \sum_i |\mathrm{a}_{ti,h}| \|\mathbf{u}_{ti}\|^{2+\beta}. \end{aligned} \tag{47}$$

From the model to the full definition. Let $E_{n,h}$ be the square root of the weighted sum of the squared right-hand side of equation (47), over times and components, with normalization $p\mathrm{W}_n$. Then $e_{n,h} \leq E_{n,h}$. Small first moments $\mathbf{g}_{t,h}$, approximately h^2-scaled second moments, and controlled absolute higher moments make this bound small. These are conditions on the realized reference geometry and ridge coefficients. The familiar local-linear bias–variance expansion motivates such scalings (Fan & Gijbels, 1996); here the bound uses the actual smoother, including its ridge bias and signed coefficients.

For some $C_n \geq 0$, $\nu > 0$, $\kappa_n^{(0)} \in \mathbb{R}$, suppose the realized reference energy and clipping tails satisfy, for $h \in \mathbb{H}_\mathrm{f}$,

$$\left| \|\boldsymbol{\xi}_h\|_{n,w}^2 - C_n h^{-\nu} \right| \leq v_{n,h}, \qquad |\kappa_n(h) - \kappa_n^{(0)}| \leq k_{n,h}, \qquad v_{n,h}, k_{n,h} \geq 0.$$

With $B_n = \|\boldsymbol{b}^{(0)}\|_{n,w}$, a sufficient condition for $\mathcal{G}_{n,\delta}$ is

$$\max_{h \in \mathbb{H}_\mathrm{f}} \left\{ \begin{array}{c} (2h^2 B_n + E_{n,h}) E_{n,h} + v_{n,h} + k_{n,h} \\ + 2(h^2 B_n + E_{n,h}) \sqrt{C_n h^{-\nu} + v_{n,h}} \end{array} \right\} \leq \delta. \tag{48}$$

Indeed, $\|\mathbf{b}_h\|_{n,w} \le h^2 B_n + E_{n,h}$ and $\|\boldsymbol{\xi}_h\|_{n,w} \le \sqrt{C_n h^{-\nu} + v_{n,h}}$; Cauchy–Schwarz bounds the cross term by the last line, while the other terms bound their counterparts in Definition 1. If desired, B_n is bounded using $|b_t^{(0,a)}| \le L_{2,a}\|\boldsymbol{M}_{t,0}\|_F/2$ and the same weighted norm. Thus explicit function bounds and sample conditions imply the full structural condition, without separately assuming a small cross term. This sufficient bound is conservative and need not be small for every sample. It is a structural implication, not an estimator of δ or a probability bound on $\mathcal{G}_{n,\delta}$. The method instead audits its necessary risk-shape consequence.

H Computational cost

Proposition 2 (Worst-case dense local-probe cost)**.** *Let $q = d + 1$, and let m be the number of references at one query. A dense arithmetic upper bound for all J local ridge probes is*

$$O\big(md + J(mq^2 + mqp + q^3 + q^2p)\big) \tag{49}$$

arithmetic operations. Summing this expression over distinct queried times gives the cost when probe caching avoids repeated fitting. If up to n distinct queries each use their growing prefix of references, the local-probe component has worst-case cost

$$O\big(J\{n^2(q^2 + qp) + n(q^3 + q^2p)\}\big)\,. \tag{50}$$

Given cached predictions and losses, paired interval statistics for T audit points cost $O(TP(B+p))$ when the tuning grid is fixed, where $P = \binom{J}{2}$ and B is the number of range candidates. Across checkpoints, including the bounds on the quasiconvexity score, this is $O(P(B+p)\sum_{n\in\mathbb{N}}|\mathbb{T}_n| + KJ^2)$, apart from calibration scans. Computing these bounds at one checkpoint costs $O(J^2)$.

Proof. Distance evaluation costs $O(md)$. At each bandwidth, forming the weighted design and response moments takes at most $O(mq^2 + mqp)$; a dense factorization and solve with at most p right-hand sides takes $O(q^3 + q^2p)$. Summing $m \le t$ over $t \le n$ gives the second display. Each of $O(J^2)$ pairs uses T forecast distances and quadratic statistics over B candidates. For a middle bandwidth j, the score bounds can be written using left and right pairwise extrema, requiring $O(J)$ work; summing over j gives $O(J^2)$. □

Caching prevents recomputation of the same probe at later checkpoints; it does not turn growing-reference regression into a constant-time update. The quasiconvexity score alone can be evaluated in $O(J)$ with prefix and suffix minima. Neither fact implies linear-time execution of the full method in stream length. These costs exclude encoder training and additional preparation diagnostics. Checkpoint-specific audit arrays can additionally require storage proportional to $Jp\sum_{n\in\mathbb{N}}|\mathbb{T}_n|$, plus their calibration arrays; an arbitrary checkpoint schedule need not have $\sum_n |\mathbb{T}_n| = O(n_K)$.

I Support and adequacy in the main implementation

All cutoffs below are configured before auditing; configuration names refer to `GateConfig`. For the one-step protocol ($H = 1$) used throughout, let m_{ref} be the minimum reference count, t_{prep} the first allowed audit index, and f_{audit} the audit fraction. The implemented schedule is

$$s_n = \min\{n, \max(m_{\mathrm{ref}}, t_{\mathrm{prep}}, \lfloor n(1 - f_{\mathrm{audit}})\rfloor)\}, \qquad \mathbb{T}_n = \{s_n, \dots, n-1\}.$$

An empty $\mathbb{T}_n$ is inadequate. In the code, m_{ref} is `min_reference`, t_{prep} is `min_audit_index`, and f_{audit} is `audit_fraction`.

Predictable common support. For every $h \in \mathbb{H}_{\mathrm{f}}$, require

$$m_t \ge d+1, \qquad \mathrm{N}_{t,h} \ge N_{\mathrm{local,min}}, \qquad \frac{\min_{i\in\mathbb{I}_t}\|\mathbf{z}_i - \mathbf{z}_t\|}{\sqrt{d}\,h} \le r_{\mathrm{nn,max}}, \qquad \mathrm{cond}(\widetilde{\mathbf{X}}_{t,h}) \le \kappa_{\mathrm{cond}},$$

and a finite probe prediction. Their conjunction across all bandwidths defines $\mathrm{S}_t = 1$. The matrix $\widetilde{\mathbf{X}}_{t,h}$ is formed from rows $\sqrt{\mathrm{k}_{ti,h}}(1, (\mathbf{z}_i - \mathbf{z}_t)^\top)$ with $\mathrm{k}_{ti,h} > \max(\mathtt{tiny}, \omega_{\mathrm{floor}}\max_r \mathrm{k}_{tr,h})$, and each column is normalized by its Euclidean norm. Insufficient rows, numerically zero columns, or singular designs receive infinite condition number. The cutoffs are `min_local_ess`, `max_nearest_distance_ratio`, `max_condition_number`, and `weight_floor`.

Audit-level adequacy. Put $\mathrm{q}_n = \sum_{t \in \mathbb{T}_n} \mathrm{S}_t$ and define

$$\mathrm{sf}_n = \frac{\mathrm{q}_n}{|\mathbb{T}_n|}.$$

Then Adeq_n is the conjunction of

$$s_n - g \geq m_{\mathrm{ref}}, \quad \mathrm{q}_n \geq q_{\min}, \quad \mathrm{sf}_n \geq \gamma_{\min},$$
$$\mathsf{Noncollapsed}_n, \quad \mathrm{W}_n > 0, \quad 0 < \mathrm{N}_{\min,n} < \infty, \quad 0 \leq \eta_n < \infty.$$

The corresponding count and fraction keys are `min_common_queries` and `min_coverage`. The support fraction is a proportion of usable queries, distinct from statistical confidence coverage. Reference adequacy is checked after excluding the gap: when $g > 0$, a nonempty scheduled audit can fail $s_n - g \geq m_{\mathrm{ref}}$ and is reported as unresolved. The schedule alone does not guarantee adequacy.

To define $\mathsf{Noncollapsed}_n$, let $\ell_1, \ldots, \ell_d$ be the eigenvalues of the centered sample covariance of $\mathbf{z}_0, \ldots, \mathbf{z}_{s_n-1}$. It requires $s_n \geq 2$ and

$$\sum_a \ell_a \geq v_{\min}, \qquad \#\{a : \ell_a > \epsilon_{\mathrm{rank}} \max(v_{\min}, \max_r \ell_r)\} \geq \lceil r_{\mathrm{frac}} d \rceil.$$

In the code, $v_{\min}$ is `min_latent_variance`, ϵ_{rank} is `rank_relative_tolerance`, and r_{frac} is `min_latent_rank_fraction`. Theorem 1 only needs adequacy to imply positive W_n and a valid tolerance; the additional checks make the implemented decision reproducible.

J Scope and Limitations

RETICLE assesses accumulated data from a specified monitoring start without accessing the downstream learner. Its certificate concerns a necessary condition on the clipped one-step probe-risk profile in a frozen representation and over a finite bandwidth grid; no shape claim is made between audited bandwidths. It does not certify monotonic improvement with elapsed time; even a high-loss constant profile is quasiconvex. A PASS establishes neither the full structural regularity condition nor an absolute risk bound. It does not guarantee improvement after retraining or fine-tuning or identify an optimal update size for a particular learner. The guarantee does not automatically extend to newly learned representations.

The checkpoint study examines the position of the selected checkpoint relative to the observed transition from large to small gains per additional pair (Figure 2), and reports the improvement from the initial to the selected checkpoint (Table 5). The error at the largest checkpoint does not account for the cost of waiting for additional data and therefore does not by itself determine whether waiting longer is worthwhile. The online study evaluates this trade-off through a full-period score that includes waiting and update periods, comparing RETICLE with a validation-selected fixed-waiting policy.

Statistical validity requires causal auditing and covers the declared checkpoints and representations. An online alarm can define an audit start if it is a stopping time and subsequent calibration and auditing remain causal. A fresh α per drift episode gives an episode-level guarantee; stream-wide control requires a separate allocation $\sum_e \alpha_e \leq \alpha$. Budgets are not automatically shared across undeclared datasets or random seeds. No unconditional sample-size or eventual-passage guarantee is provided. An unresolved audit may reflect inadequate support or imprecise confidence bounds and does not establish that the data are unsuitable for downstream learning.

Data-only assessment still requires representation preparation and local regression. Caching avoids repeated probe fitting at the same query but does not yield constant-time updates. Growing reference sets can therefore limit scalability on long streams.

K Additional Experimental Details

Reproducibility. https://anonymous.4open.science/r/RETICLE-CE5A.

Table 3: Real-data tasks for checkpoint selection and online adaptation. Channel counts refer to variables at each time step; all listed leads are predicted jointly. Each online dataset forms one series.

Dataset	Cases/ series	Sampling	History L (samples)	Input/target channels	Leads (samples)
Checkpoint selection					
Rotating machinery	4	3,000 Hz	300	3/3	$\{1, 6, 30\}$
Treadmill	8	50 Hz	50	18/18	$\{1, 10, 25\}$
WESAD	4	25 Hz	800	1/1	$\{12, 25, 50\}$
Online adaptation					
ETTm1	1	15 min	96	7/7	$\{1, \ldots, 48\}$
Jena Weather	1	10 min	288	8/8	$\{1, \ldots, 48\}$
Tetouan City	1	10 min	144	8/3	$\{1, \ldots, 48\}$

Audit and representation settings. Local affine probes use ridge coefficient $\rho = 10^{-3}$, loss clipping $M = 4$, reference-count audit weights, and the bandwidth grid

$$\{h^{-2}\} = \{0.01, 0.1, 0.3, 0.5, 0.75, 1, 1.5, 2, 3, 4, 5, 6, 7, 8\}.$$

All 14 bandwidths enter the common-support checks and paired comparisons. Minimum reference counts, local ESS, common-query counts, and supported-query fractions are $(64, 30, 32, 0.8)$ for synthetic data and $(32, 10, 16, 0.7)$ for real and online data. The scheduled audit fraction is 0.5 and the past-exclusion gap is zero. The predictive encoder projects standardized histories onto ridge-regression/SVD directions, completed with preparation-data PCA directions when needed to obtain 32 dimensions. Each projected coordinate is centered, and all coordinates share one scale: the square root of their mean preparation-data variance, preserving relative coordinate variances. Preparation targets follow the downstream task in checkpoint studies and one-step forecasting online. Representations and their internal normalizers are frozen within each evaluation phase.

Synthetic data. Lorenz–96, coupled FitzHugh–Nagumo, and Kuramoto–Sivashinsky use seeds 3001–3010. Their state dimensions are 40, 50, and 64; observations concatenate direct state, linearly mixed, and nuisance channels, giving 120, 150, and 192 inputs, respectively. Histories contain one observation, and targets are the noisy observed state channels at the next two steps. For each system, a 20,000-observation source path (seed 91001) fixes the target and observation-noise scales; monitored and test paths contain 16,915 and 4,010 observations and are generated independently for each seed. Signal-channel observation noise has standard deviation 0.1 times the corresponding clean source-channel standard deviation. The encoder is fitted on the first 512 complete downstream examples of the monitored path. All checkpoints and learners share the test path.

Real-data cases and inputs. Table 3 summarizes the 16 real checkpoint cases and three online series. Rotating machinery comprises four runs combining conditions H/BD with operating frequencies of 15/50 Hz; each retains the 15–300 s interval at 3,000 Hz without smoothing. Treadmill comprises eight schedule-defined intervals from subjects G09/G10 and runs 01/02; three IMUs are synchronized using causal 20 ms averages of the 1 kHz measurements. WESAD uses one contiguous baseline episode (label 1) from each of S2–S5. Chest recordings are causally low-pass filtered and reduced from 700 to 25 Hz; only respiration is used for inputs and targets. These cases assess data accumulation within fixed episodes and do not provide verified drift-onset annotations.

Preparation, checkpoints, and holdouts. Every real case fits its representation on its first 500 complete downstream examples (512 for synthetic data). These are history–target pairs, not raw observations. For both synthetic and real data, checkpoint counting begins only after all preparation targets are observed. With P_{prep} preparation examples and maximum downstream lead H_{out}, the first $O = P_{\mathrm{prep}} + H_{\mathrm{out}} - 1$ one-step pairs are excluded from audit references, audit queries, and downstream training; subsequent input histories may reuse earlier observations. The first audit attempt is at 64 new pairs. For synthetic data, $m_{\mathrm{ref}} = 64$ makes this audit set empty and the result unresolved; the earliest possible selection is therefore $n = 128$. Reaching the largest checkpoint

does not force a selection; only an adequate passing checkpoint is selected. At checkpoint n, there are $\max(0, n - H_{\text{out}} + 1)$ complete downstream examples. The figure uses n, not this latter count.

The shared candidate grid is

$$\mathbb{N}_{\text{full}} = \{64, 128, 256, 512, 1024, 2048, 3072, 4096, \\ 5120, 6144, 7168, 8192, 10240, 12288, 14336, 16384\}.$$

Synthetic runs use all 16 checkpoints. Real cases use the grid up to 8,192 for rotating machinery and WESAD (12 checkpoints), and 4,096 for treadmill (8 checkpoints), excluding preparation pairs. The final 20% of each real episode is reserved for evaluation. For episode length N and history length L, set $S = \lfloor 0.8N \rfloor$. Using zero-based raw indices, evaluation origins are $S + L - 1 + jH_{\text{out}}$ for $j \geq 0$ with complete targets inside the episode; at most 600 evenly spaced origins are retained. Holdouts are fixed across checkpoints and learners, and the retained grid size enters the confidence allocation.

Downstream learners. In checkpoint experiments, learners are fitted from scratch on observed histories at each checkpoint, with input/output normalizers fitted on that checkpoint's training examples; scoring scales remain fixed. KRR uses an RBF kernel with regularization 0.01 for synthetic and 0.05 for real data. MLP has hidden widths $(64, 32)$ and at most 400 fitting iterations. Extra-Trees uses 250 trees for synthetic and 80 for real data, with minimum leaf size 2. XGBoost uses 300 trees, depth 5, and learning rate 0.05. Transformer has two layers, four heads, and width 64; TCN uses 64 channels, kernel size 3, dropout 0.1, and enough dilated blocks to cover the input history. Transformer and TCN are initially fitted for 80 epochs with batch size 128. The online study uses the same MLP, Transformer, and TCN architectures. All three use initial learning rate 10^{-3}.

Online preparation and detection. Weather uses the Jena roof-station data from 2020, with wind direction encoded as sine and cosine, giving eight channels. Tetouan uses five environmental variables and three zone loads as inputs, predicting only the loads. No calendar features are added. Forecasts with incomplete inputs or targets are excluded from scoring. The final 20% of the training interval is reserved for detector calibration and warmup; initial downstream fitting uses at most the most recent 4,096 complete examples from the preceding interval.

After all 48 targets become available, each forecast contributes its normalized MSE averaged over forecast steps and target variables. Detectors receive nonoverlapping means of 32 such losses, after clipping each loss at 4 and dividing it by 4. Page–Hinkley monitors increases with threshold 1.0, `delta`=0.005, forgetting factor 0.9999, and `min_instances`=30 block means. ADWIN uses `delta`=0.05, `clock`=1, minimum window length 5, and grace period 30. Detector parameters and block size are fixed across datasets, with a nonbinding budget of 1,000 alarms per 1,000 observed losses and experiment seed 42. Detection is paused while waiting for data or fitting an update. At model activation, the detector is reset and subsequently receives only losses from forecasts issued by the new model. Missing observations cancel an ongoing waiting period or pending update, and the detector is reset when complete observations resume.

The validation encoder uses all complete one-step pairs in training, without the downstream fitting cap. Before testing it is refitted on training plus validation data and then frozen. Downstream models and detectors are reset to their common training-derived initial states; validation updates are discarded. Downstream input/output normalizers, scoring scales, and probe-loss scales remain fixed from initial fitting data. Scoring uses a separate sample variance for each forecast-step/target coordinate, estimated from the initial fitting examples and then frozen. Normalized MSE averages squared errors divided by these variances over forecast steps, target variables, and valid origins, without clipping the reported score.

Validation selection. Candidates independently replay the full validation interval from the same initial state, using common valid origins with complete targets inside validation. Waiting sizes are $\{64, 128, 256, 512, 1024, 2048, 4096\}$ one-step pairs.

RETICLE searches $c \in \{0.1, 0.2, \ldots, 1.0\}$, requiring a complete comparable score and at least one update start. Candidates whose starts are all forced at the maximum checkpoint are excluded; mixed passed and forced starts remain eligible. Let $\bar{n}(c)$ average selected checkpoints over all started updates, including forced ones. Scanning c in ascending order, retain the first finite positive mean,

then retain candidates only if their means do not exceed the last retained mean. Rejections do not advance this reference. This is an empirical selection constraint, not a monotonicity guarantee. Among retained candidates, select the smallest c with normalized MSE at most $1.001E_*$, where E_* is their best validation MSE. If none remains, use $c = 0.1$; this includes exclusively forced starts or no updates across all candidates.

Fixed waiting minimizes validation MSE over the waiting-size grid; CALIPER selects its ESS factor from $\{2, 3, 4\}$ by the same metric. Exact ties favor the smaller waiting size and ESS factors in the order $3, 2, 4$; defaults are 4,096 and 3 if comparable scores are unavailable or no candidate with a comparable score starts an update. No selection uses test errors. Although the RETICLE audit is independent of the downstream learner, this online tuning step uses downstream validation performance.

Updates and the CALIPER baseline. After an alarm, candidate pairs use complete histories from the new episode. At checkpoint n, a contiguous episode provides $\max(0, n - 47)$ complete 48-step training examples. Each update uses these examples once in shuffled order, with batch size at most 32, learning rate 5×10^{-4} for all layers, and a fresh optimizer. MLP uses Adam with L2 regularization; Transformer and TCN use AdamW, with regularization coefficient 10^{-4}. RETICLE and CALIPER force an update at 4,096 pairs if their gates have not passed; forced starts are recorded separately from gate passes. The previous model supplies predictions during waiting and measured gate/update computation, whose elapsed time is converted to stream steps using the dataset's sampling interval. These periods are included in the full-period score.

CALIPER retains its physical-state one-step probes and monotonicity criterion, with locality grid $\{0, 0.1, 1, 2, 4, 8, 16\}$. Its ESS threshold is the selected factor times $(d_{\text{state}} + 1)$, where d_{state} is the number of physical-state variables. Readiness requires five consecutive supported queries whose cumulative error profile is monotone within 10^{-3} times its range. Local fits use no ridge penalty. Readiness is checked on each arrival and retained until a common update checkpoint. Its original full-retraining procedure is adapted to the shared fine-tuning budget and maximum checkpoint. For Tetouan, CALIPER predicts all eight physical-state variables while RETICLE audits the three demand targets; this comparison evaluates the complete procedures and does not isolate their shape constraints.

Aggregation and interpretation. For each seed or case u and learner m, let $E_{u,m}(n)$ denote held-out normalized MSE at checkpoint n: squared prediction errors are divided by fixed target-channel variances before averaging over evaluation origins, forecast leads, and target variables. Synthetic variances are estimated from the independent source-reference targets; real-data variances use observed target values through the last preparation label. Each channel's scale is shared across forecast leads and fixed across checkpoints and learners. This score is not clipped; real-data variables that are constant during preparation are excluded using a mask fixed before evaluation. For consecutive checkpoints $n_{i-1} < n_i$, the plotted quantity is

$$G_{u,m}(n_i) = \frac{E_{u,m}(n_{i-1}) - E_{u,m}(n_i)}{n_i - n_{i-1}}.$$

The denominator is the actual increase in one-step probe-pair count, even though the horizontal axis is logarithmic. The plot places each interval's value at n_i, then weights available seeds or cases equally to compute the mean and sample standard deviation. Gray bands show ± 1 sample standard deviation of these paired improvements, with bands drawn only when at least two paired values exist. The checkpoint at $n = 64$ has no preceding checkpoint, so the plot starts at $n = 128$, showing $(E_{u,m}(64) - E_{u,m}(128))/64$ when both scores are available. This display choice retains $n = 64$ in evaluation, audit selection, and the computation of the first gain. An interval is omitted if either endpoint is missing or invalid, and missing checkpoints are not bridged. Negative improvements are retained. Positive gains indicate error reductions and negative gains indicate error increases. Proximity to zero indicates little change in MSE per additional pair over the corresponding checkpoint interval; large negative values indicate deterioration. This diagnostic concerns marginal changes in error per pair, not the attained error level or the distance to the error at the largest checkpoint.

Blue selection lines and bands show the median and interquartile range of the first passing checkpoint among passing runs. Non-passing runs remain in the displayed PASS-count denominator and

Table 4: RETICLE-selected checkpoints and additional data collected after the initial checkpoint. Checkpoint sizes are median [Q1, Q3] across seeds (synthetic) or cases (real). The initial checkpoint is $n_0 = 64$; $\hat{n}$ is the first passing checkpoint, selected once per seed or case and shared across all six downstream learners. Checkpoint sizes count fully observed one-step probe pairs collected after representation preparation; PASS and Valid size report counts of seeds or cases.

Dataset	n_0	$\hat{n}$	$\hat{n} - n_0$	PASS	Valid size
Lorenz–96	64 [64, 64]	5120 [5120, 5888]	5056 [5056, 5824]	10/10	10/10
Coupled FHN	64 [64, 64]	2048 [1280, 2048]	1984 [1216, 1984]	10/10	10/10
KS	64 [64, 64]	3584 [3072, 4096]	3520 [3008, 4032]	10/10	10/10
Rotating	64 [64, 64]	3072 [3072, 3328]	3008 [3008, 3264]	4/4	4/4
WESAD	64 [64, 64]	4096 [3584, 4608]	4032 [3520, 4544]	4/4	4/4
Treadmill	64 [64, 64]	3072 [2048, 3072]	3008 [1984, 3008]	8/8	8/8

are not assigned a selection. Neither type of band is a confidence interval. These retrospective differences do not enter RETICLE's checkpoint selection.

Online ranks compare the four reported policies within each of the 18 conditions using unrounded errors, assigning average ranks to ties. The same no-update score is reused across detectors for each dataset and model. The online study measures continuous forecasting performance rather than drift-localization accuracy.

L Additional Experimental Results

L.1 Prediction Errors at Selected Checkpoints

Figure 2 measures the marginal benefit of data accumulation through the decrease in normalized MSE per additional one-step probe pair. To complement this measure with the attained error levels, Tables 4 and 5 report the selected data sizes and compare the normalized held-out MSE at the initial checkpoint $n_0 = 64$ with that at the RETICLE-selected checkpoint $\hat{n}$. This comparison quantifies the cumulative improvement from the initial small-data fit to the selected checkpoint. The error at the largest checkpoint is not used as a reference; Appendix J explains why.

Selected data sizes. The initial checkpoint n_0 is the first scheduled downstream evaluation, and $\hat{n}$ is the first checkpoint at which RETICLE passes its gate. Each seed or case contributes one selection, obtained without access to downstream learners or their errors and shared across all six learners. Table 4 reports n_0, $\hat{n}$, and the additional data collected, $\hat{n} - n_0$. These sizes count fully observed one-step probe pairs after representation preparation; downstream training uses only examples whose complete forecast targets are observed. The PASS column counts seeds or cases with a recorded gate passage, and Valid size counts verified selections with a known initial checkpoint. Both are reported relative to the total number of seeds or cases.

Prediction errors and aggregation. Let $E(n)$ denote a learner's normalized held-out MSE at checkpoint n, using the same fixed evaluation data and target-variance normalization at both checkpoints within each seed or case. For each seed or case and learner, we compute the MSE reduction as

$$100\,\frac{E(n_0) - E(\hat{n})}{E(n_0)}.$$

Positive values indicate improvement from the initial checkpoint, and negative values indicate deterioration. Table 5 reports medians and first and third quartiles across seeds or cases, with each percentage computed before aggregation. Consequently, the median reduction need not equal the reduction calculated from the two reported median MSE values. Both MSE summaries and the reduction use the same valid paired comparisons. The Valid column reports the number of complete pairs with positive initial MSE relative to the total number of seeds or cases. Downstream errors are used only for this retrospective evaluation.

All 30 synthetic runs and 16 real cases provide a passing selection and valid paired comparisons for all six learners. Median MSE reductions are positive in all 36 dataset–learner combinations, ranging

Table 5: Normalized held-out MSE at the initial checkpoint $n_0 = 64$ and the RETICLE-selected checkpoint $\hat{n}$. MSE values and reductions are median [Q1, Q3] across seeds (synthetic) or cases (real). MSE reduction is the paired percentage decrease from the initial checkpoint, computed within each seed or case before aggregation; positive values indicate improvement. The selected checkpoint is shared across all six learners within each seed or case.

Dataset	Learner	$E(n_0)$	$E(\hat{n})$	MSE reduction (%)	Valid
Lorenz–96	KRR	0.9581 [0.9172, 0.9605]	0.08341 [0.08113, 0.08571]	91.27 [90.91, 91.63]	10/10
	MLP	1.144 [1.121, 1.177]	0.169 [0.1648, 0.1745]	84.89 [84.57, 86.05]	10/10
	ExtraTrees	1.01 [0.9814, 1.021]	0.6162 [0.6154, 0.6216]	38.99 [37.12, 39.29]	10/10
	Transformer	1.121 [1.113, 1.154]	0.03788 [0.03606, 0.03818]	96.75 [96.60, 96.83]	10/10
	XGBoost	0.7491 [0.7282, 0.7876]	0.04282 [0.04145, 0.04306]	94.31 [94.16, 94.50]	10/10
	TCN	0.9766 [0.9625, 0.9902]	0.08209 [0.07935, 0.0832]	91.64 [91.57, 91.92]	10/10
Coupled FHN	KRR	0.3425 [0.3018, 0.6049]	0.04321 [0.02251, 0.2062]	87.55 [32.45, 92.92]	10/10
	MLP	0.2773 [0.2235, 0.3301]	0.02623 [0.01756, 0.05999]	89.90 [79.93, 93.12]	10/10
	ExtraTrees	0.2604 [0.2366, 0.5187]	0.03171 [0.02004, 0.1296]	86.35 [50.45, 92.00]	10/10
	Transformer	0.258 [0.2338, 0.4116]	0.02597 [0.01651, 0.1078]	88.76 [58.68, 93.19]	10/10
	XGBoost	0.238 [0.2201, 0.317]	0.02201 [0.01396, 0.1012]	90.73 [56.51, 93.27]	10/10
	TCN	0.2462 [0.2068, 0.265]	0.02421 [0.016, 0.05003]	87.98 [79.53, 92.15]	10/10
KS	KRR	1.463 [1.277, 1.59]	0.03762 [0.03291, 0.04505]	97.44 [96.61, 97.78]	10/10
	MLP	1.425 [1.201, 1.554]	0.01999 [0.01815, 0.02214]	98.48 [98.33, 98.72]	10/10
	ExtraTrees	1.223 [1.106, 1.376]	0.1346 [0.1203, 0.1514]	89.24 [88.13, 90.28]	10/10
	Transformer	1.053 [0.9744, 1.276]	0.01843 [0.01656, 0.01893]	98.39 [98.14, 98.54]	10/10
	XGBoost	1.286 [1.19, 1.359]	0.0198 [0.01884, 0.02066]	98.46 [98.37, 98.55]	10/10
	TCN	1.141 [1.063, 1.319]	0.01701 [0.0158, 0.01875]	98.49 [98.43, 98.61]	10/10
Rotating	KRR	1.035 [0.9725, 1.041]	0.3955 [0.2452, 0.5412]	61.86 [48.03, 75.09]	4/4
	MLP	1.625 [1.549, 1.719]	0.6551 [0.3989, 0.8792]	61.23 [44.26, 78.07]	4/4
	ExtraTrees	1.062 [1.011, 1.073]	0.6919 [0.5273, 0.8206]	34.96 [23.50, 48.15]	4/4
	Transformer	1.597 [1.445, 1.789]	0.7532 [0.3738, 1.128]	55.14 [28.45, 79.59]	4/4
	XGBoost	1.129 [1.035, 1.252]	0.4668 [0.2761, 0.6571]	60.05 [47.89, 73.34]	4/4
	TCN	1.117 [1.044, 1.168]	0.4911 [0.2879, 0.6824]	56.60 [38.14, 75.36]	4/4
WESAD	KRR	1.008 [0.6947, 1.208]	0.3902 [0.2574, 0.5527]	43.63 [40.26, 53.71]	4/4
	MLP	4.726 [3.82, 5.613]	0.376 [0.3041, 0.5505]	91.01 [84.97, 94.54]	4/4
	ExtraTrees	0.9786 [0.6872, 1.134]	0.2458 [0.1711, 0.4002]	61.78 [54.24, 68.68]	4/4
	Transformer	1.067 [0.7278, 1.274]	0.3184 [0.273, 0.407]	57.18 [48.42, 67.10]	4/4
	XGBoost	0.8812 [0.7093, 0.9699]	0.2312 [0.1599, 0.3808]	66.45 [59.47, 70.31]	4/4
	TCN	0.9747 [0.6606, 1.156]	0.1887 [0.1477, 0.3259]	68.61 [58.11, 76.87]	4/4
Treadmill	KRR	0.4923 [0.3683, 0.538]	0.1712 [0.1551, 0.2081]	61.18 [48.92, 70.02]	8/8
	MLP	0.9025 [0.5825, 1.004]	0.1945 [0.1805, 0.2295]	72.71 [68.11, 78.28]	8/8
	ExtraTrees	0.5143 [0.4268, 0.5714]	0.1522 [0.1394, 0.1994]	68.77 [57.45, 74.73]	8/8
	Transformer	0.5528 [0.4904, 0.6319]	0.1783 [0.1619, 0.2115]	67.38 [58.88, 73.32]	8/8
	XGBoost	0.5857 [0.5005, 0.7076]	0.168 [0.1513, 0.2206]	66.84 [61.74, 74.26]	8/8
	TCN	0.4479 [0.3738, 0.5203]	0.1581 [0.1422, 0.1857]	64.47 [54.97, 70.98]	8/8

from 34.96% to 98.49%. Thus, the shared selections yield substantial median improvements over the initial fits across the evaluated learners. Together with Figure 2, these results describe both the improvement accumulated by the selected checkpoint and the marginal benefit of collecting further data.

L.2 Update behavior during online adaptation

Table 6 complements the forecasting errors in Table 1 with RETICLE's update behavior in each of the 18 test conditions. It reports the validation-selected fixed waiting size, the validation-selected coefficient c, the numbers of fine-tuning starts triggered by gate passage or the maximum-size fallback, the mean checkpoint size, and the forced-start rate.

Counting and aggregation. Only episodes in which fine-tuning actually starts contribute to the counts. These counts can differ from the number of model activations, since a started update may not become active within the evaluated period. PASS and forced starts are distinguished by their recorded decision reasons: a PASS at the maximum checkpoint remains a PASS-triggered start. Mean n averages the numbers of fully observed one-step probe pairs collected in the corresponding episodes over all fine-tuning starts, including any forced starts. The forced-start rate is the number of forced starts divided by the total number of starts, expressed as a percentage.

Table 6: RETICLE fine-tuning starts across the 18 online test conditions. Fixed n is the waiting size (in one-step pairs) selected on validation for the fixed-waiting policy, and Selected c is the validation-selected RETICLE coefficient. PASS and Forced starts count fine-tuning starts triggered by gate passage and by the 4,096-pair fallback, Mean n is the mean checkpoint size over all starts, and Forced rate is the percentage of starts that were forced. Counts refer to fine-tuning starts, which may differ from the number of model activations.

Dataset	Detector	Model	Fixed n	Selected c	PASS starts	Forced starts	Mean n	Forced rate (%)
ETTm1	Page–Hinkley	MLP	1024	0.5	6	0	1109.3	0.0
ETTm1	Page–Hinkley	Transformer	1024	0.6	7	0	822.9	0.0
ETTm1	Page–Hinkley	TCN	1024	0.6	8	0	832.0	0.0
ETTm1	ADWIN	MLP	2048	0.4	1	0	2048.0	0.0
ETTm1	ADWIN	Transformer	1024	0.4	2	0	2048.0	0.0
ETTm1	ADWIN	TCN	512	0.4	2	0	2048.0	0.0
Weather	Page–Hinkley	MLP	2048	0.5	5	0	908.8	0.0
Weather	Page–Hinkley	Transformer	256	0.8	5	0	550.4	0.0
Weather	Page–Hinkley	TCN	1024	0.5	5	0	652.8	0.0
Weather	ADWIN	MLP	4096	0.5	3	0	469.3	0.0
Weather	ADWIN	Transformer	2048	0.5	2	0	1088.0	0.0
Weather	ADWIN	TCN	1024	0.5	2	0	2176.0	0.0
Tetouan City	Page–Hinkley	MLP	1024	0.2	3	0	1706.7	0.0
Tetouan City	Page–Hinkley	Transformer	2048	0.3	3	0	1024.0	0.0
Tetouan City	Page–Hinkley	TCN	256	0.5	3	0	426.7	0.0
Tetouan City	ADWIN	MLP	4096	0.3	1	0	1024.0	0.0
Tetouan City	ADWIN	Transformer	4096	0.2	1	0	2048.0	0.0
Tetouan City	ADWIN	TCN	512	0.1	1	0	4096.0	0.0

Observed update behavior. All 60 fine-tuning starts were triggered by gate passage: 26 on ETTm1, 22 on Weather, and 12 on Tetouan City. No test condition required a forced start. This includes Tetouan City with ADWIN and TCN, whose single start followed a PASS at $n = 4096$. Thus, all recorded fine-tuning starts in these experiments were initiated by RETICLE's audit, and the maximum-size fallback was not invoked. Validation selects the largest waiting size (4,096) in only three of the 18 conditions; in those three, RETICLE updates after far fewer pairs on average (469–2,048) and attains lower test error. Conversely, in three of the four conditions where fixed waiting attains lower error, RETICLE's mean checkpoint exceeds the fixed size. Waiting longer is therefore not uniformly better under the full-period score, which charges the waiting period.

Aggregation of forecasting improvements. The relative MSE reductions reported in Section 5.2 are computed separately for each condition as $100(1 - E_{\mathrm{RETICLE}}/E_b)$, where b is the validation-selected fixed waiting policy or CALIPER. The percentages are then averaged with equal weight across the 18 conditions using unrounded errors. The resulting mean reductions are 6.08% relative to fixed waiting and 11.91% relative to CALIPER.