FORECASTING UNDER CHANGING RESPONSE DELAYS: PREDICTION AND MEMORY WHEN RELATIONS DISAPPEAR

Anonymous authors
Paper under double-blind review

ABSTRACT

An upstream signal may remain useful when its response arrives later, but rising forecast error can also reflect a broken relation or increased noise. We study what a forecaster should predict and remember across these changes. A target-history model fitted while the relation is active can absorb the source effect it must later predict without. Evaluating the shared relation at a training-input reference improves cancellation prediction in two controlled mechanisms. We derive how the inactive prediction also changes effective delay refresh through posterior mass. Matching conditional delay weights preserves the cancellation benefit, while different return rules change its size in opposite directions. Remembering inactive delays helps old-delay recurrence and costs accuracy when a new delay returns. A fixed reference/adaptive configuration reduces noise error by 31.3% on new synthetic streams and 10.6% on semisynthetic streams driven by household measurements, relative to a fixed-scale joint tracker. It meets the synthetic criteria but fails the joint test because external delay and stable-window errors exceed their tolerances. A fixed refresh control is competitive. Changing the inactive target level reverses the reference's advantage; learning an offset mitigates this at a transient cost. These results identify a practical coupling: the forecast assigned to inactivity determines both immediate error and how much delay information survives for future use.

1 INTRODUCTION

A change in transport speed can move an upstream sensor's response from three steps later to eight steps later while leaving the response function useful. A predictor could reuse the relation by reading a different part of the observed history. But a rise in forecast error does not tell it which history to read. The response may have moved, the relation may have disappeared, or measurements may simply have become noisier. These changes call for different predictions and different ways of retaining evidence.

The distinction begins with what a state predicts. An active state can evaluate a shared response at a candidate delay. An inactive state must also produce a forecast. A natural fallback is a model fitted using only target history. Its training targets, however, still contain the source's influence. Removing the source input from a regression is a different operation from removing the relation that generated those targets. We compare this fallback with evaluating the shared relation at a source reference fixed from training. The latter specifies what the fitted relation contributes during inactivity without fitting another coefficient vector or observing the change time.

The inactive prediction also affects future memory. Its residual determines how much posterior mass becomes inactive. Returning that mass uniformly to active delays dilutes a concentrated delay preference. Thus the forecast assigned to inactivity can change actual forgetting even when transition parameters are identical. Conversely, storing an old delay through inactivity makes a prediction about how the relation will return. That memory can be helpful or costly.

We use a small forecasting model to make these interactions inspectable. One nonlinear coefficient vector is shared across candidate delays and remains frozen during evaluation. A common observation schedule separates prediction from target revelation. All variants reuse the same fitted active

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.

predictions; their differences concern the inactive mean, likelihood scale, and transition of probability. A factorization into activity and delay weights expresses the same instantaneous forecast as a joint vector. Extra labels therefore do not by themselves create useful memory or a new predictive alternative.

These interactions lead to three contributions:

1. We show how an inactive forecast couples immediate prediction and effective delay refresh through posterior mass. Target-history and training-reference comparisons reveal cancellation gains and change which fixed refresh rules meet prediction constraints.
2. Matched-trajectory and recurrence comparisons separate the roles of delay-weight evolution and retained memory. Matching conditional delay weights preserves a cancellation advantage without identifying a unique memory contribution; retaining an old delay helps when it returns but costs accuracy when a new delay appears.
3. We test these choices on new synthetic streams, measured-input semisynthetic streams, inactive-level shifts, and natural tasks. Synthetic noise and cancellation gains persist, while the household-input test fails the joint criterion. The remaining analyses identify when the reference is unsuitable.

2 RELATED WORK

Changing relations and forecasts. LIFT estimates leading variables and delays within a history window (Zhao & Shen, 2024); FACT represents interactions in time and frequency (Wang et al., 2026); COSA corrects a frozen forecast using revealed targets (Im & Kwon, 2026). State-space causal forecasting allows causal strengths and noise variances to change (Huang et al., 2019). Invariant prediction motivates transfer when a conditional relation remains stable (Rojas-Carulla et al., 2018). We examine a narrower question: after fitting a relation, which prediction represents its absence, and what does that choice do to delay memory?

State tracking and recurrence. Hidden Markov models support delay estimation (Azzouzi & Nabney, 1999; Gu et al., 2026), dynamic model averaging combines changing predictors (Raftery et al., 2010), and fixed sharing restores probability to unlikely experts (Herbster & Warmuth, 1998). Expert tracking with memory retains preferences for their later return (Bousquet & Warmuth, 2002; Zheng et al., 2019). LSTD separates long- and short-term states under unknown interventions (Cai et al., 2025), while Continuous Evolution Pool retrieves specialized forecasters when concepts recur (Zhan et al., 2026). State separation and memory reuse are therefore established ideas. Our comparisons isolate how an inactive emission and its return kernel interact, including the cost of remembering a delay that no longer applies.

Prediction and state identity. Predictive-state representations and HMM aggregation characterize when state information affects future prediction (White et al., 2000; Shalizi & Crutchfield, 2001; Littman et al., 2001). Several delays can yield similar forecasts, so low prediction error need not identify the generating delay. We measure these outcomes separately and treat the filter's likelihood weights as working quantities rather than calibrated probabilities (Gneiting & Raftery, 2007).

3 METHODOLOGY

At time t, the predictor uses source and target measurements available through $t-1$, along with training-fitted parameters. It saves a prediction of y_t, scores it when the target is revealed, and then updates for the next step. A missing target supplies neither supervision nor a state update. The setting is synchronous, two-channel, one-step prediction. The deployed predictor receives no change times, activity indicators, or true delays.

3.1 A RELATION SHARED ACROSS DELAYS

Training statistics center and scale the two variables. In the equations below, target history is standardized and z denotes the standardized source. Let $u_t = (1, y_{t-1}, \ldots, y_{t-P})^\top$ and $\phi(z) =$

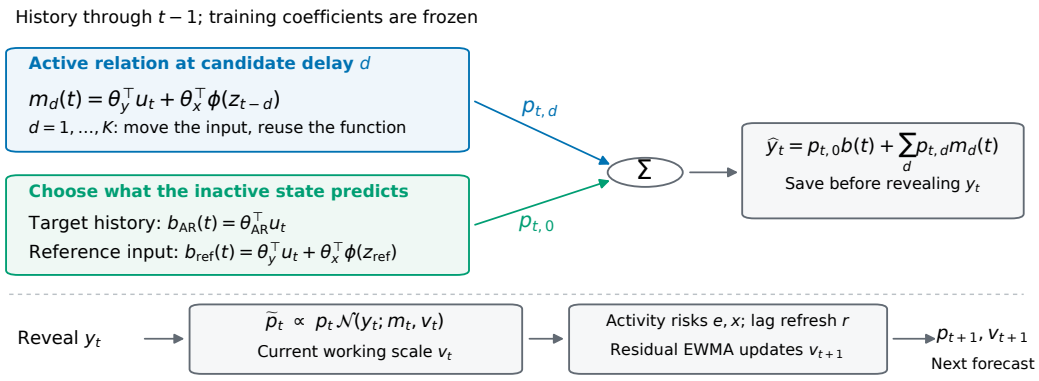


Figure 1: Predictive state semantics and observation order. Active states reuse one fitted relation at different source delays. The inactive state uses one of two explicitly defined predictions. The mixture is saved before the target is revealed. Candidate likelihoods, activity transitions, and the residual-scale update affect subsequent forecasts. Arrows show computation, not causal effects in the data-generating system.

$(z, \tanh z, z^2)^\top$. Ridge regression fits a relation for each candidate training delay. The delay with the smallest training MSE supplies one shared coefficient vector, used by every active candidate at evaluation:

$$m_d(t) = \theta_y^\top u_t + \theta_x^\top \phi(z_{t-d}), \qquad d = 1, \ldots, K. \tag{1}$$

Moving probability between candidates changes the history position while retaining the response function. We use $P = 4$, $K = 24$, ridge penalty 1, and an unpenalized intercept. The relation has eight fitted coefficients. A separately fitted target-history model adds five coefficients.

All methods clip historical inputs at eight training standard deviations. Regression targets, revealed labels, and scored errors remain unclipped. Observed training targets determine target scaling and supervised fits. Missing sensor history is forward-filled from available observations, using a training median before the first observation. This filling does not turn a missing target into a training label.

3.2 Two meanings of the inactive prediction

The target-history fallback is $b_{\mathrm{AR}}(t) = \theta_{\mathrm{AR}}^\top u_t$, fitted on the active training stream. The reference-input fallback instead uses the target-history component of the shared relation:

$$b_{\mathrm{ref}}(t) = \theta_y^\top u_t + \theta_x^\top \phi(z_{\mathrm{ref}}), \quad z_{\mathrm{ref}} = \mathrm{clip}\left(\frac{x_{\mathrm{ref}} - c_x}{s_x}, -8, 8\right). \tag{2}$$

Here c_x, s_x are the training source center and scale. The training-mean reference sets the raw source to $x_{\mathrm{ref}} = c_x$, so $z_{\mathrm{ref}} = 0$ and $\phi(z_{\mathrm{ref}}) = 0$ for this basis. A second fixed control uses raw $x_{\mathrm{ref}} = 0$. Setting the source to its mean evaluates $\phi(\mathbb{E}[X])$. It does not average nonlinear source features as $\mathbb{E}[\phi(X)]$.

The difference has a simple statistical explanation. In an active system $Y = a(H) + f(X_d) + \epsilon$ with zero conditional-mean noise,

$$\mathbb{E}[Y \mid H] = a(H) + \mathbb{E}[f(X_d) \mid H]. \tag{3}$$

If cancellation removes the entire f term and preserves a and the noise mechanism, the new conditional mean is $a(H)$. The active-data target-history conditional mean need not equal it. Our finite-order ridge AR is an approximation to that conditional mean and also has estimation and regularization error. A reference-input fallback states a different modeling assumption: inactivity uses the fitted relation at a specified source input. Its agreement with an actual cancellation is an empirical question.

3.3 WEIGHTS, LIKELIHOOD SCALE, AND ACTIVITY TRANSITIONS

Let p_t be a joint weight vector over the inactive prediction and K active candidates. The forecast is $\widehat{y}_t = p_{t,0}b(t) + \sum_{d=1}^{K} p_{t,d}m_d(t)$. After observing y_t, a common Gaussian working variance v_t gives $\widetilde{p}_{t,d} \propto p_{t,d}\exp[-(y_t - m_d(t))^2/(2v_t)]$, with $m_0(t) = b(t)$. The Gaussian normalizer cancels because the variance is shared by all candidates at that time.

We distinguish three transition parameters: e moves active mass into inactivity, x returns inactive mass to activity, and r refreshes delay weights within activity. Writing $B = \widetilde{p}_{t,0}$, $A = 1 - B$, and $q_d = \widetilde{p}_{t,d}/A$ when $A > 0$, the next prior is

$$\begin{aligned} p_{t+1,0} &= (1-x)B + eA, \\ p_{t+1,d} &= \frac{xB}{K} + (1-e)A\left[(1-r)q_d + \frac{r}{K}\right]. \end{aligned} \tag{4}$$

This makes the activity risks explicit. Uniform sharing over all $K+1$ states instead implies $e = \alpha/(K+1)$ and $x = K\alpha/(K+1)$. Increasing the number of delay candidates would then change activity behavior as well as delay resolution.

The reference-input comparison fixes $e = 0.003$, $x = 0.01$, and $r = 0.0288346$, using the same initial weights and shared active candidates as its target-history counterpart. Both update their working variance from the residual of the prediction saved before seeing the target:

$$v_{t+1} = \mathrm{clip}\left(0.95v_t + 0.05(y_t - \widehat{y}_t)^2,\ v_0,\ 16v_0\right), \tag{5}$$

where $v_0 = 2\max(\mathrm{MSE}_{\mathrm{train}}, 0.01)$. The posterior and transition use v_t. The new variance enters the next prediction. Residuals can contain a wrong mean or delay, so this is a working likelihood scale. Changing the fallback can change the realized residual and variance trajectory even though the update rule stays fixed.

3.4 STATE LABELS AND EFFECTIVE MEMORY

Separate relevance and delay weights make the same instantaneous forecast as $p = [1-g; gq]$. When $g > 0$ the map is invertible. Thus splitting the vector does not itself introduce a new predictive alternative or store additional information; differences arise from the update and the history it retains.

The full transition also shows why a nominal delay-refresh rate is an incomplete description of memory. For $A > 0$, the next active mass is $A^+ = (1-e)A + xB$. If $A^+ > 0$, its conditional delay weights satisfy

$$q^+ = (1 - r_{\mathrm{eff}})q + r_{\mathrm{eff}}\mathbf{u}, \qquad r_{\mathrm{eff}} = r + (1-r)\frac{xB}{(1-e)A + xB}, \tag{6}$$

where $\mathbf{u}$ is uniform over delays. Probability returning from the inactive state adds refresh even when $r = 0$. If all posterior mass is inactive and $x > 0$, the next conditional delay distribution is uniform; the previous conditional q is then undefined. Equation 6 describes one transition step. Observation likelihoods supply an additional, nonlinear update between transitions.

We also compare fixed fractions of the original rate $r_0 = 0.0288346$ with the rule $r_t = r_0 \min\{1, v_0/v_t\}$. The latter reduces internal sharing when the pre-reveal working variance rises; it changes neither activity risk. For two fixed candidate means separated by δ, the expected Gaussian working log likelihood ratio in favor of a correct mean is $\delta^2/(2v_t)$. This motivates slowing forgetting when evidence per observation is weaker. It does not make the variance rule optimal: scale error, mean error, and the extra return term in r_{eff} remain.

This identity connects the inactive prediction to memory: its residual changes B, which changes effective refresh. The direction and size of the forecasting effect depend on the fully specified predictor. Section 4.3 compares the complete fixed and variance-scaled panel under both inactive predictions. Similarly, preserving a delay while inactive helps only when the old delay contains information about its return. The same information can be represented by an expanded joint state.

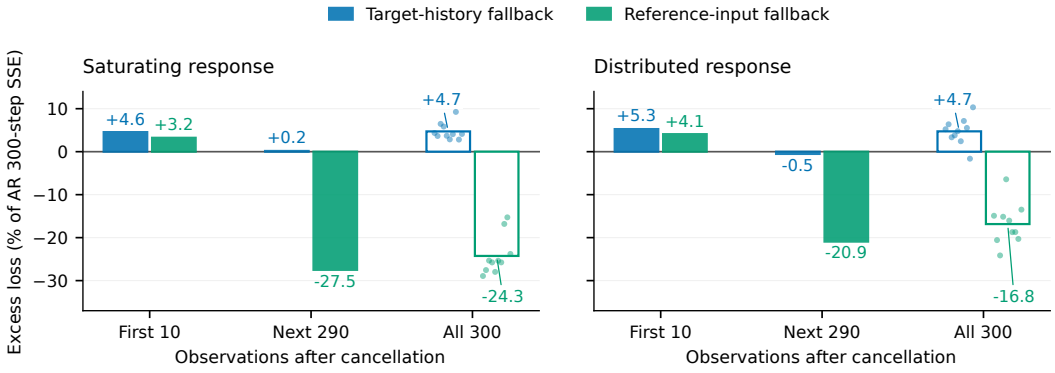


Figure 2: Early cost and later cancellation benefit. Filled bars split excess loss into the first ten observations and the next 290; outlines give their sum, with all ten independent-pair totals shown as dots. Each contribution uses that pair's full 300-point AR SSE denominator, so it is not short-window relative MSE. Both filters have the same activity risks, refresh and EWMA rules.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

The controlled panel uses two nonlinear response mechanisms with target autoregression and Gaussian noise: a saturating quadratic response to an AR(1) source and a distributed sine–tanh response to an AR(2) source. The shared model approximates the first response and omits part of the second. Cancellation removes the entire external term; a delay change moves its input position. Each mechanism has ten independent training/evaluation pairs, each evaluated under all eight conditions: stable, delay, noise, both orders of delay and noise, simultaneous delay and noise, cancellation, and cancellation after noise.

These comparisons reuse a 160-stream development panel. Ratios are formed within each pair before averaging, and uncertainty resamples independent training/evaluation pairs with their conditions kept together. Joint with fixed scale is the stable/delay/noise reference; standalone AR is also a cancellation reference. Appendix B gives generators, metrics, complete comparison families and descriptive interval definitions.

4.2 PREDICTING RELATION LOSS

At fixed risks and refresh, replacing the target-history fallback with the training-mean reference lowers cancellation MSE from about 1.047 times AR in both mechanisms to 0.757 and 0.832. Cancellation after a noise increase also improves, to 0.794 and 0.775 times AR. Directly paired noise error is 0.99978 times the target-history version: the replacement preserves its noise benefit rather than adding a material one. The raw-zero reference behaves similarly, while always using the training-mean prediction raises stable error to 18.60 and 9.26 times Joint (Appendix C). The active candidates remain essential.

The improvement mainly follows the initial transient (Figure 2). The first ten observations with the reference contribute $+3.23$ and $+4.08$ percentage points of full-window AR loss. The next 290 contribute -27.49 and -20.92 points, yielding total reductions of 24.26% and 16.84%. A saved-prediction decomposition using the true post-cancellation mean finds that squared conditional-mean bias falls from 36.98% to 8.70% of the same AR loss in the saturating system, and from 39.41% to 15.58% in the distributed system. The latter retains model mismatch, but this decomposition does not assign the remaining bias uniquely to the omitted response: estimation and state weighting also contribute. The true mean is used only for evaluation.

Table 1: Inactive prediction changes which refresh rules suffice. Noise is a within-stream ratio to the target-history filter at r_0, averaged across four noise conditions and both mechanisms. Cancellation is relative to standalone AR (saturating/distributed); delay is the worst registered condition mean relative to Joint. Pass requires all 44 preset constraints. C marks cancellation failure. All 14 crossings, including lower fixed rates and their delay failures, appear in Table 4.

Inactive prediction	Refresh	Noise	Cancel S/D	Delay	Constraints
Target history	r_0	1.0000	1.047/1.047	1.011	Pass
Target history	$r_0/4$	0.9573	1.056/1.054	1.013	C
Target history	Variance-scaled	0.9568	1.048/1.048	1.010	Pass
Reference	r_0	0.9998	0.757/0.832	1.013	Pass
Reference	$r_0/4$	0.9569	0.767/0.837	1.015	Pass
Reference	Variance-scaled	0.9564	0.758/0.832	1.011	Pass

4.3 Inactive predictions and refresh policies

We cross both inactive predictions with seven preset policies: $r/r_0 \in \{1, 1/2, 1/4, 1/8, 1/16, 0\}$ and variance scaling $r_t = r_0 \min\{1, v_0/v_t\}$. Each filter retains its own pre-reveal variance. The 44 constraints bound stable, pre-event and delay loss by 1.02 times Joint, cancellation loss by 1.05 times both AR and Joint, and extra capped recovery by ten observations, while requiring at least 2% noise improvement over Joint. They apply across both mechanisms and the registered conditions (Appendix B).

With the target-history fallback, variance scaling reduces directly paired noise error to 0.9568 times the original explicit-risk filter, with a descriptive 95% interval of $[0.9544, 0.9594]$. Fixed quarter-rate sharing reaches almost the same noise error, 0.9573, but misses the cancellation tolerance in both mechanisms. Changing the fallback changes this feasible set: reference half-rate and quarter-rate filters satisfy all 44 constraints. Adaptive noise error is 0.9995 times quarter-rate error, with interval $[0.9973, 1.0015]$. This interval neither establishes equivalence nor shows adaptive refresh is necessary with the reference fallback. Lower fixed rates eventually fail the delay constraint; all 14 comparisons are retained in the appendix.

These comparisons show which policies meet the stated tolerances, rather than a unique causal pathway. In particular, the preset pure-noise interaction between fallback replacement and variance scaling is -0.00022 in units of the adaptive target-history loss, with an interval spanning zero. The important change is that a different inactive mean admits useful fixed refresh rates previously excluded by cancellation.

4.4 Matching and retaining delay memory

4.4.1 Matched delay trajectories

The inactive mean changes both posterior mass and each filter's residual scale. To isolate part of this coupling, a diagnostic shares the original target-history filter's pre-reveal variance between both inactive means. It then compares three return rules at fixed quarter-rate or at the variance-scaled rate derived from that donor. Uniform return (U) is the original rule. Common total refresh (C) chooses feasible internal rates so both filters have the same r_{eff}. Self-return (S) sends inactive mass to the filter's refreshed conditional delay vector. Under C and S, the two inactive means have identical conditional delay trajectories; their total activity mass and full predictions can still differ. Appendix E specifies the interventions and all comparisons.

The reference's cancellation advantage persists under both matches. With the shared variance-scaled policy, the reference-minus-history loss differences for cancellation after noise are -0.2020 under U, -0.1578 under C and -0.2149 under S, normalized by the same stream's original adaptive target-history loss. The C and S changes have opposite signs, with paired intervals excluding zero. Pure cancellation shows the same directional contrast, while noise-window interactions include zero. Thus changes in conditional delay memory affect the size of the benefit, but are not its necessary explanation under these interventions. Sharing a donor scale changes the experiment; these contrasts do not decompose the original filters' gains into causal percentages.

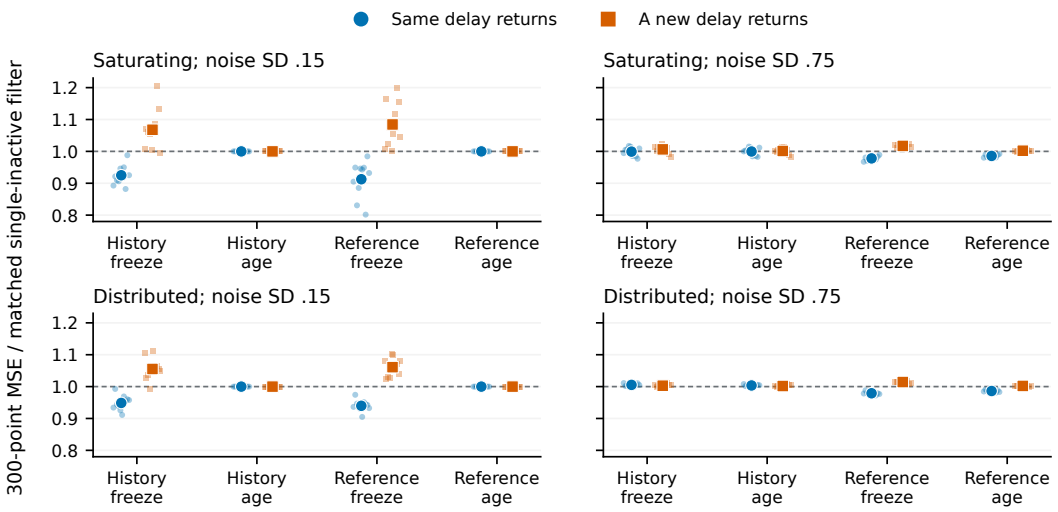


Figure 3: Remembering an inactive delay helps its return but can obstruct relocation. All four expanded filters are compared with their respective single-inactive-state variance-scaled reference. Large markers show mean within-pair 300-point loss ratios; faint markers show all ten pairs. The two mechanisms and both noise levels retain same-delay and new-delay returns. Freeze does not age mass already inactive; new entries can still alter its delay distribution. Age also refreshes existing inactive mass. These are 80 separately generated diagnostic streams, not additional candidate-selection observations.

4.4.2 Recurrence and Relocation

An expanded joint filter keeps K inactive states I_d alongside its K active states. All I_d predict the same inactive mean; their indices only affect the delay distribution returned to activity. We compare leaving already inactive mass unaged with applying the existing refresh operator to it, for both inactive means. New active-to-inactive entries can change either version's conditional inactive distribution. Initial inactive mass is spread uniformly, so the model is not given the old delay. The activity risks, active predictions and pre-reveal scale rule remain fixed.

The recurrence diagnostic crosses both mechanisms, ten paired repetitions, two noise levels and returns at either the old delay or a new delay. Reference-freeze improves 300-point same-delay loss by at least 2% in all four mechanism/noise cells, with paired difference intervals below zero. At low noise its early 40-point loss ratios are 0.687 and 0.749; when a new delay returns, they become 1.330 and 1.262. The corresponding 300-point new-delay costs are 8.44% and 6.10%, exceeding the preset 2% tolerance. High-noise new-delay costs are smaller but remain positive in both paired difference intervals. Aging avoids the large relocation cost but fails the 2% same-delay benefit criterion in all four cells. No variant meets both full boundary criteria. The complete core controls and all recurrence comparisons appear in Appendix F.

The gain is therefore conditional on recurrence structure. Identical inactive emissions supply no direct likelihood evidence about which delay should return. Retaining a delay helps if the old value remains informative; a new value makes that same memory an obstacle. An expanded joint state can express this memory, so the result does not establish an inherent advantage of factorized state labels.

4.5 Held-out Transfer

We fixed the reference/adaptive configuration before a two-arm test: 320 new synthetic streams and 576 streams driven by measured household electric-power inputs (Hebrail & Berard, 2006). Both arms retain all 64 comparison methods. The external arm generates response targets from the two controlled mechanisms; it tests transfer to measured inputs, not forecasting of the household's measured target. Its twelve monthly blocks come from one household, with three noise repetitions per

Table 2: Two-arm confirmation of the fixed reference/adaptive filter. Joint outcome: **FAIL**. Noise is the mean paired MSE ratio to JointSoft with a descriptive 95% interval. Delay and Stable are maximum cell ratios to JointSoft; Stable includes every pre-change cell. Cancel is the maximum ratio to AR over both cancellation conditions and mechanisms. The final column counts all 44 predeclared criteria met. Controlled intervals resample paired replicates within mechanism; household-input intervals resample shared month blocks within one household. Lower loss ratios are better.

Arm	Inactive prediction / refresh	Noise / JointSoft	Delay	Stable	Cancel / AR	Met
Controlled	History / fixed r_0	0.716 [0.706, 0.726]	1.015	1.004	1.092	**43/44**
	History / adaptive	0.687 [0.677, 0.698]	1.014	1.003	1.093	**43/44**
	Reference / fixed quarter	0.686 [0.676, 0.696]	1.013	0.981	0.827	44/44
	Reference / adaptive	0.687 [0.677, 0.697]	1.014	1.004	0.818	44/44
Household input	History / fixed r_0	0.900 [0.880, 0.922]	1.013	1.003	1.165	**43/44**
	History / adaptive	0.885 [0.863, 0.907]	1.011	1.001	1.156	**43/44**
	Reference / fixed quarter	0.892 [0.869, 0.916]	1.101	1.043	0.882	**41/44**
	Reference / adaptive	0.894 [0.870, 0.918]	1.113	1.096	0.880	**41/44**

Table 3: The reference needs a credible inactive level. After cancellation, ℓ is the target steady-state offset; the table reports ℓ/s_y, where s_y is training target SD. Entries are mean paired 300-point loss ratios to AR, shown as saturating/distributed. Both inactive-state filters use variance-scaled refresh. RLS adapts the shared-feature coefficients. All signs and both noise levels are retained; each mechanism has ten pairs.

Noise SD	ℓ/s_y	Reference / AR	History / AR	Shared RLS / AR
.15	-1	3.045/1.915	1.125/1.055	1.522/1.467
.15	0	0.775/0.829	1.069/1.045	4.548/3.357
.15	$+1$	4.914/1.934	1.073/1.058	2.792/1.483
.75	-1	1.117/0.902	1.042/0.949	1.071/0.959
.75	0	0.793/0.786	0.972/0.886	1.074/0.952
.75	$+1$	1.113/0.908	1.005/0.972	1.082/0.971

block. Intervals resample independent training/evaluation pairs within each synthetic mechanism, or shared month blocks across mechanisms and noise repetitions externally (Appendix D).

On new synthetic pairs, reference/adaptive meets all 44 criteria, with mean paired noise error 0.687 times Joint and interval $[0.677, 0.697]$. Its worst delay-window ratio is 1.014; pure-cancellation ratios to AR are 0.818 and 0.811. The household-driven arm retains a noise advantage, 0.894 times Joint with interval $[0.870, 0.918]$, and cancellation ratios 0.775 and 0.880 to AR. Yet distributed-response delay loss rises to 1.113 times Joint and stable-window loss to 1.096, exceeding their 1.02 tolerances. The delay-first condition fails on the same 300-point prefix before noise changes, so these two failed cells are dependent. The joint criterion is not met. None of the 64 methods meets the entire external criterion set.

The controls explain why a noise benefit is insufficient. History/adaptive keeps the external stable and delay ratios near one but incurs a 1.156 cancellation ratio to AR. Reference with fixed quarter-rate refresh has slightly lower noise loss than reference/adaptive and satisfies the synthetic criteria; it still fails the same three external cells. Their directly paired noise comparison favors fixed refresh by about 0.18% externally, and does not establish an adaptive advantage on the synthetic arm (Appendix D). Thus changing the inactive prediction changes the pattern of errors, while neither refresh policy removes the transfer cost.

A post-hoc decomposition locates most of the reference filter's excess loss over the history/adaptive control in squared distance from the true conditional mean. In May, 47.83% of source-history values lie outside the training range and 14.33% exceed the model's clipping limits. This association identifies an input-range boundary; it does not show that clipping caused the failure, and all twelve months remain in the test.

4.6 INACTIVE TARGET LEVELS

The cancellation benefit assumes that the reference describes the remaining response. To test that assumption, a paired 120-stream diagnostic sets the inactive steady-state target offset to zero or ± 1 training target SD. Sources and target innovations are identical across levels. The reference still evaluates the shared model at the training input mean and retains its target-history term; it is not a constant target-mean predictor.

At zero offset, its 300-point error is 0.775–0.829 times AR. At nonzero offsets and low noise, that ratio rises to 1.915–4.914, while the target-history fallback gives 1.055–1.125 (Table 3). Thus the same reference that helps remove a source contribution can misrepresent a change in the level of the remaining process. Larger noise partly masks this mismatch; it does not resolve the underlying choice of inactive prediction.

A follow-up diagnostic learns one inactive-prediction intercept, comparing posterior-weighted and ungated updates at three fixed rates. All six reduce 300-point loss in every shifted-level cell, but all still lose to AR over the first ten points. Slow updates can remain worse after 300 points; faster updates erode some zero-offset benefits. Posterior weighting is not necessary: one matched ungated control has less than half its loss. Appendix H gives the complete six-rule comparison and coefficient-adaptation controls. These reused development streams show that a fixed reference's failure is repairable, with a learning cost; they provide no new household confirmation or rule for choosing an update.

4.7 PREDICTION RECOVERY AND DELAY IDENTIFICATION

A 1,800-stream diagnostic varies source correlation, response amplitude, noise and delay step at stationary source variance one. For the original explicit-risk filter, 40-point oracle conditional information explains held-out delay-concentration time about 60% better than step count, but explains prediction recovery little better. It also fails the stronger comparisons against physical variables and a linear AR(1) proxy. Predicting adequately and identifying a delay are distinct outcomes; this result supplies neither a universal recovery law nor an observable online information measure. Appendix G gives the fixed grid, held-out splits and failed comparisons.

4.8 NATURAL FORECASTING TASKS

The natural tasks predict PT08.S1 from four different source sensors in UCI Air Quality, and T1 from T2 in Appliances Energy Prediction (Vito, 2008; Candanedo, 2017). They use chronological $60/20/20$ splits, training-fitted scaling, 24-point histories and the same prediction/revelation order. These intervals were exposed in the original study; they contain no verified cancellation or delay labels. Both fallback versions here use constant refresh r_0, testing the inactive-mean replacement without the later variance-scaled refresh rule. The reference improves on the target-history fallback by about 0.20% on S3 and loses by 0.46%–3.83% on the other four. It beats standalone AR on three air tasks but loses on S4 and Appliances. The best previously evaluated neural mean is lower on every task; all four neural controls and their three-initialization variability and all five task results appear in Appendix I. The four air tasks share a target and time interval, so they represent one system, not four independent datasets.

5 CONCLUSION

An inactive state is a predictive assumption, and retained memory is an assumption about recurrence. A useful reference can improve cancellation without making adaptive refresh essential; a useful old delay can become a liability when the relation returns elsewhere. The exact transition identity explains why the two choices interact, while the matched-memory and recurrence comparisons delimit what that identity predicts about loss. Practical use therefore requires validating the inactive forecast and the expected return pattern against task-specific alternatives. The held-out test makes this concrete: synthetic gains survive, but measured input dynamics expose stable and relocation costs that fail the joint criterion. A fixed reference also fails when the inactive process settles at another level; learning an inactive offset can then help after an initial learning cost. The useful outcome is a testable choice of inactive prediction and memory, with its transfer failures made explicit.

AI USE STATEMENT

AI agents assisted in literature retrieval, research and experimental design, theoretical derivations and proof drafting, implementation, synthetic-data generation, experiment execution, analysis, figures and writing. Numerical checks and reviews by separate agents are computational verification, not human peer review. The authors remain responsible for the final work.

REPRODUCIBILITY STATEMENT

The anonymous supplement provides fixed configurations, data-retrieval procedures and hashes, model sources, and complete saved numerical inputs for the reported panels. Display builders reconstruct figures and tables; three replay entries cover the controlled confirmation, household confirmation and natural tasks. Additional diagnostics support saved-data reconstruction. Full numerical replay is a separate verification step. Per-stream raw archives, neural checkpoints and third-party repositories are excluded from the compact supplement.

ETHICS STATEMENT

The study uses controlled numerical systems and public sensor data. It infers no personal attributes and makes no physical intervention. Forecast dependence is statistical rather than causal; the results establish no operational safety guarantee.

REFERENCES

Mehdi Azzouzi and Ian T. Nabney. Time delay estimation with hidden Markov models. In *9th International Conference on Artificial Neural Networks: ICANN '99*, volume 1, pp. 473–478, 1999. doi: 10.1049/cp:19991154. URL https://research.aston.ac.uk/en/publications/time-delay-estimation-with-hidden-markov-models/.

Olivier Bousquet and Manfred K. Warmuth. Tracking a small set of experts by mixing past posteriors. *Journal of Machine Learning Research*, 3:363–396, 2002. URL https://www.jmlr.org/papers/v3/bousquet02b.html.

Ruichu Cai, Haiqin Huang, Zhifan Jiang, Zijian Li, Changze Zhou, Yuequn Liu, Yuming Liu, and Zhifeng Hao. Disentangling long-short term state under unknown interventions for online time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15641–15649, 2025. doi: 10.1609/aaai.v39i15.33717. URL https://ojs.aaai.org/index.php/AAAI/article/view/33717.

Luis Candanedo. Appliances Energy Prediction. UCI Machine Learning Repository, 2017. URL https://doi.org/10.24432/c5vc8g. Dataset.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Ya Gu, Lujia Wang, Yonghong Tan, and Shiyi Yin. Expectation maximization estimation for multirate state-space models with time-varying delays. *Journal of the Franklin Institute*, 363(6):108526, 2026. doi: 10.1016/j.jfranklin.2026.108526. URL https://doi.org/10.1016/j.jfranklin.2026.108526.

Georges Hebrail and Alice Berard. Individual household electric power consumption. UCI Machine Learning Repository, 2006. URL https://archive.ics.uci.edu/dataset/235/individualhouseholdelectricpowerconsumption. Dataset.

Mark Herbster and Manfred K. Warmuth. Tracking the Best Expert. *Machine Learning*, 32(2):151–178, 1998. doi: 10.1023/A:1007424614876. URL https://doi.org/10.1023/A:1007424614876.

Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2901–2910, 2019. URL https://proceedings.mlr.press/v97/huang19g.html.

Jeonghwan Im and Hyuk-Yoon Kwon. COSA: Context-aware Output-Space Adapter for Test-Time Adaptation in Time Series Forecasting. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 24845–24874, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/2a8ce71baac4c89bf9ff479d8240c7d9-Paper-Conference.pdf.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, pp. 1555–1561, 2001. URL https://proceedings.neurips.cc/paper/2001/file/1e4d36177d71bbb3558e43af9577d70e-Paper.pdf.

Adrian E. Raftery, Miroslav Kárný, and Pavel Ettler. Online Prediction Under Model Uncertainty via Dynamic Model Averaging: Application to a Cold Rolling Mill. *Technometrics*, 52(1):52–66, 2010. doi: 10.1198/tech.2009.08104. URL https://doi.org/10.1198/tech.2009.08104.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018. URL https://jmlr.org/papers/v19/16-432.html.

Cosma Rohilla Shalizi and James P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3–4):817–879, 2001. doi: 10.1023/A:1010388907793. URL https://link.springer.com/article/10.1023/A:1010388907793.

Saverio Vito. Air Quality. UCI Machine Learning Repository, 2008. URL https://doi.org/10.24432/c59k5f. Dataset.

Huiqiang Wang, Jieming Shi, and Qing Li. FACT: Fine-grained Across-variable Convolution for Multivariate Time Series Forecasting. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 4967–4996, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/090d636f951ac0bb18fd5610c95c6a76-Paper-Conference.pdf.

L. B. White, R. Mahony, and G. D. Brushe. Lumpable hidden Markov models: Model reduction and reduced complexity filtering. *IEEE Transactions on Automatic Control*, 45(12):2297–2306, 2000. doi: 10.1109/9.895565. URL https://doi.org/10.1109/9.895565.

Tianxiang Zhan, Ming Jin, Yuanpeng He, Yuxuan Liang, and Shirui Pan. Continuous evolution pool: Taming recurring concept drift in online time series forecasting. *arXiv preprint arXiv:2506.14790v3*, 2026. URL https://arxiv.org/abs/2506.14790v3. Version 3, revised August 18, 2026.

Lifan Zhao and Yanyan Shen. Rethinking Channel Dependence for Multivariate Time Series Forecasting: Learning from Leading Indicators. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 41433–41451, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/b52b07a239a7afa155ca25cf17a55074-Paper-Conference.pdf.

Kai Zheng, Haipeng Luo, Ilias Diakonikolas, and Liwei Wang. Equipping experts/bandits with long-term memory. In *Advances in Neural Information Processing Systems*, volume 32, pp. 5927–5937, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/39ae2ed11b14a4ccb41d35e9d1ba5d11-Abstract.html.

A STATE MEANINGS, INFORMATION, AND MEMORY

What active observations determine about cancellation. Consider an unanchored model class in which both decompositions $a(H) + f(X)$ and $[a(H) + c] + [f(X) - c]$ are permitted, for a nonzero constant c. Their active conditional means agree at every observed (H, X). If cancellation is defined as deleting the entire source term in each decomposition, their first cancelled conditional means are $a(H)$ and $a(H) + c$. The same available history therefore leads any predictor to the same value p in the two worlds, while

$$\max\{(p - a(H))^2, (p - a(H) - c)^2\} \geq c^2/4. \tag{7}$$

The inequality follows because two points separated by $|c|$ cannot both lie less than $|c|/2$ from p. This concerns the first prediction before cancellation is observed. Subsequent targets can distinguish the worlds. It applies to the stated unanchored function class. Our fitted basis uses $\phi(0) = 0$ and a separate target-history intercept, so it already fixes one decomposition convention. The counterexample does not assert parameter nonidentifiability within that fixed basis. The remaining empirical issue is whether the convention describes the inactive regime.

Working evidence and true information. At a prediction step, fix two history-measurable candidate means μ_i, μ_j and a common positive working variance v. Their Gaussian log likelihood ratio is

$$\ell = \frac{(Y - \mu_j)^2 - (Y - \mu_i)^2}{2v} = \frac{\delta}{v}\left(Y - \frac{\mu_i + \mu_j}{2}\right), \qquad \delta = \mu_i - \mu_j. \tag{8}$$

If the true conditional mean is μ_i and conditional variance is σ^2, zero-mean residuals give $\mathbb{E}[\ell \mid H] = \delta^2/(2v)$ and $\mathrm{Var}(\ell \mid H) = \delta^2\sigma^2/v^2$. These moment identities do not require Gaussian residuals. The first is the true conditional KL between the two candidate worlds only when their working densities are the true Gaussian conditional distributions, including $v = \sigma^2$. With arbitrary true mean m, the expected working ratio is $\delta(2m - \mu_i - \mu_j)/(2v)$ and can be negative.

For the oracle diagnostic, candidate means are evaluated on the same actual target history and source path, with a common true innovation variance. The autoregressive contribution then cancels in their difference. Summing $\delta_t^2/(2\sigma_t^2)$ over a fixed window gives realized conditional information along that path. Its appropriate expectation appears in the chain-rule KL for the observation sequence. An observed path sum is not itself the unconditional KL. Input correlation changes the sequence of mean separations, which a delay step count alone does not specify.

Effective delay refresh. Summing the active entries of Equation 4 gives $A^+ = (1 - e)A + xB$. Dividing each active entry by A^+ proves Equation 6. When $A > 0$, $e < 1$, and $r < 1$, the contraction toward the uniform distribution obeys

$$-\log(1 - r_{\text{eff}}) = -\log(1 - r) + \log\left(1 + \frac{xB}{(1 - e)A}\right). \tag{9}$$

The first term is internal delay sharing; the second is uniform mass returning from inactivity. A rule that changes r according to the working variance controls the first term, but not automatically the second. When $A = 0$ and $x > 0$, the next conditional active distribution is uniform and the old conditional distribution is undefined. If $A^+ = 0$, the new conditional distribution is also undefined. These identities concern the transition operator, with likelihood updates treated separately.

When retaining an inactive delay changes prediction. An expanded state can retain inactive indices I_d alongside active indices A_d. Suppose inactive states have the same emission distribution, and the covariate sequence defining future emissions is fixed and common to the full and compressed predictors, independent of discarded history. Aggregating all I_d preserves every future observation law when transitions from any two inactive indices have identical total probability into each retained block. Summing hidden paths within blocks then gives the same observation law and aggregated filtering posterior (White et al., 2000). With a common reactivation probability ρ and return-delay kernel R, identical rows of R suffice. Common return delay alone is insufficient if reactivation timing still depends on the discarded index.

For square-integrable Y, full history H, and retained information $C = C(H)$, the squared-loss cost of compression is

$$\mathbb{E}[(Y - \mathbb{E}[Y \mid C])^2] - \mathbb{E}[(Y - \mathbb{E}[Y \mid H])^2] = \mathbb{E}[(\mathbb{E}[Y \mid H] - \mathbb{E}[Y \mid C])^2]. \tag{10}$$

Expanding the square and conditioning on H removes the cross term. If full history identifies an inactive delay D, the compressed predictor retains probabilities w_d, and reactivation gives mean $r_d = \sum_j R(d,j)\mu_j$, the conditional cost before observing activation is $\rho^2 \operatorname{Var}_{D\sim w}(r_D)$. An ordinary joint model on $2K$ states can retain the same information as its factorization into activity-specific delay distributions.

B CONTROLLED DATA, FITTING, AND EVALUATION

B.1 WHAT CHANGES IN EACH STREAM

All controlled innovations below are independent standard Gaussians. The two response mechanisms are

$$x_t = .65x_{t-1} + .65\eta_t, \qquad y_t = .35y_{t-1} + a_t\{.9\tanh(1.2x_{t-d_t}) + .2x_{t-d_t}^2\} + \sigma_t\epsilon_t, \quad (11)$$

$$x_t = .8x_{t-1} - .3x_{t-2} + .65\eta_t, \quad y_t = .25y_{t-1} + a_t\{.75\sin(x_{t-d_t}) + .3\tanh(x_{t-d_t-2})\} + \sigma_t\epsilon_t. \quad (12)$$

The first is called saturating and the second distributed. In the latter, d_t is an anchor parameter: the relation also uses d_t+2. It is not a claim that the response has one unique physical lag. Setting $a_t = 0$ deletes the complete external term and leaves target autoregression unchanged. This intervention defines the cancellation studied here.

Training and evaluation each retain 1,800 observations after a 200-point burn-in. Training has $a = 1$, $d = 3$ and $\sigma = .15$. Each of ten independent repetitions per mechanism uses the same training fit and evaluation innovations for all eight conditions. The first event is at 500; a second event is at 900. New delays follow the repeated sequence $(9, 17, 9, 17)$ across repetitions; increased noise SD follows $(.45, .45, .60, .60)$. Delay-only, noise-only and cancellation change at 500. Ordered noise–delay and delay–noise use 500 then 900; simultaneous changes use 500. Noise–cancellation increases noise at 500 and cancels at 900. Stable streams keep their initial settings. Training and evaluation seeds are independent; conditions within a pair are not.

B.2 SHARED FITS AND COMPARISON FAMILIES

All deployable filters use measurements through $t-1$ to predict y_t. Histories are standardized with training statistics and clipped to $[-8, 8]$; targets and errors are not clipped. After excluding the first 24 history points, ridge regression fits each delay using four target lags, an intercept and $(z, \tanh z, z^2)$. The minimum-training-MSE delay supplies the eight shared active coefficients and working scale. The target-history fallback fits five coefficients separately. Initial inactive mass is .1; active conditional weights put .999 mass on the training-selected delay and .001 uniformly over all 24 delays. No evaluation state initializes the model.

The controlled comparison retains the following method families. Joint mixes 25 predictions using a Gaussian training scale and .03 uniform sharing. Separate updates conditional delay weights and a relevance gate; the gate's mixing parameter is .02. A transition-matched joint control separates representational changes from transition changes. Scale controls cross the original and matched transition laws with fixed variance multipliers $4, 16$, residual EWMA rates $.01, .05, .2$, Student degrees of freedom $3, 5, 10$, and three-scale HMM switching rates $.005, .02, .1$. The HMM supports variance multipliers $1, 4, 16$. Student controls fix squared scale to v_0, so their distribution variance is $\nu v_0/(\nu-2)$; they are not variance-matched Gaussian controls.

Standalone AR, nonlinear RLS on the full source/target history, and RLS on the shared feature representation provide mean-adaptation comparisons. Both RLS controls use forgetting .995. Explicit activity-risk controls cross entry $\{.0003, .0012, .003\}$ with exit $\{.001, .003, .01, .0288\}$ under the .05 EWMA. Scalar amplitude controls use the same delayed features and revealed targets, including an unconstrained same-information comparator. The reference and refresh comparisons retain these preceding controls; the complete factorial panel has 64 methods. Expanding inactive memory adds four, for 68 on the core panel. The full per-stream tables preserve every method rather than selecting a favorable subset for an average. KnownDelay receives the state only in the evaluator and uses the fitted active mean or AR fallback. It is a state-informed reference, not the true conditional mean.

Table 4: Complete fallback-by-refresh development comparison. Noise ratios are formed within a stream relative to the original explicit-risk AR fallback, then averaged across four noise conditions and both mechanisms. Cancellation uses standalone AR; S/D denote saturating/distributed. Delay is the worst mechanism/condition mean relative to Joint. Pass means all preset constraints hold; C and D mark cancellation and delay failures. These are descriptive results from the reused development panel.

Fallback	Refresh	Noise /AR-r_0	Cancel S /AR	Cancel D /AR	Delay /Joint	Constraints
AR	r_0	1.0000	1.047	1.047	1.011	Pass
AR	$r_0/2$	0.9722	1.053	1.052	1.006	C
AR	$r_0/4$	0.9573	1.056	1.054	1.013	C
AR	$r_0/8$	0.9491	1.059	1.056	1.027	C+D
AR	$r_0/16$	0.9444	1.060	1.057	1.042	C+D
AR	0	0.9361	1.065	1.057	1.121	C+D
AR	Variance-scaled	0.9568	1.048	1.048	1.010	Pass
Reference	r_0	0.9998	0.757	0.832	1.013	Pass
Reference	$r_0/2$	0.9720	0.763	0.835	1.007	Pass
Reference	$r_0/4$	0.9569	0.767	0.837	1.015	Pass
Reference	$r_0/8$	0.9484	0.770	0.838	1.029	D
Reference	$r_0/16$	0.9434	0.771	0.839	1.045	D
Reference	0	0.9351	0.772	0.840	1.134	D
Reference	Variance-scaled	0.9564	0.758	0.832	1.011	Pass

B.3 Loss windows, recovery, and uncertainty

The primary event loss is MSE over 300 observations beginning at the relevant event. Cumulative squared error over that window is 300 times MSE; its paired ratio is therefore the same. Pre-event error uses the scored observations before 500; stable anchor error uses 500–799. Noise error averages the four noise conditions within a repetition before averaging repetitions and weighting mechanisms equally. Every loss ratio divides by the same stream's reference loss.

Prediction recovery asks when a trailing 40-point MSE is at most 1.25 times the fitted KnownDelay reference's 300-point event MSE plus .005 in standardized units. Three consecutive eligible windows are required. The recorded time is capped at 400 if recovery is not reached; it is not treated as an observed uncensored recovery. Exact anchor concentration instead requires conditional delay mass at the new anchor to be at least .5 for ten consecutive points. Points with active mass at most 10^{-12} do not satisfy concentration.

The 44 core constraints comprise 22 per mechanism: mean noise ratio at most .98 relative to Joint; four delay-window ratios at most 1.02; four corresponding mean extra capped-recovery times at most ten; eight pre-event ratios and one stable-window ratio at most 1.02; and two cancellation conditions compared with both Joint and AR, each ratio at most 1.05. A favorable mean cannot offset a failed cell. These are preset decision tolerances, not significance tests. Intervals resample the ten independent repetition clusters within each mechanism, keeping all their conditions and methods together, and weight mechanisms equally. We use 10,000 bootstrap draws and percentile 95% intervals. Repeated use of the development panel makes these descriptive intervals; additional mechanisms or diagnostic streams do not turn them into independent confirmation.

C Complete inactive-mean and refresh controls

Both source references are evaluated after training standardization and before nonlinear features; neither averages features. Raw-input zero gives cancellation/AR ratios .741/.810 across the two mechanisms, versus .757/.832 for the training mean. Using the mean reference alone lowers them to .701/.774, but its worst stable and delay ratios rise to 18.597 and 15.923 times Joint. This control retains the shared target-history coefficients but omits active-state selection, exposing why selection matters. Table 4 retains all fourteen crossings: reference half-rate and quarter-rate pass, whereas the lowest-noise fixed rates fail delay constraints.

D HELD-OUT SYNTHETIC AND MEASURED-INPUT TESTS

The reference with variance-scaled refresh was fixed before either test arm. Both retain the same 64-method matrix, active training procedure, activity risks and comparison criteria used in development. Outcomes do not change the selected configuration.

The synthetic arm uses twenty new training/evaluation pairs per mechanism, giving $2\times20\times8 = 320$ streams. Training retains 1,800 observations and evaluation 2,200 after independent 200-point burn-in periods. Training keeps delay 3 and noise SD .15. Events move to 650 and 1,100; repetitions cycle through the nine combinations of new delay $\{7, 13, 21\}$ and noise SD $\{.375, .525, .75\}$. All conditions within a pair share source and standard target innovations. These pairs were not used to choose the configuration.

The measured input is `Global_active_power` from the UCI household dataset (Hebrail & Berard, 2006). We fix twelve nonoverlapping 4,600-minute blocks starting at midnight on the first day of each month of 2008. In each block, the first 2,000 source observations support training and the next 2,600 support evaluation; separate 200-point target burn-ins leave 1,800/2,400 observations. Missing inputs are filled from the past, with the most recent preceding measurement used at a block start; no future input supplies an earlier value. Input normalization is fitted on the training segment. The target follows Equations (11)–(12) with this measured source. Its units are generated-response units, not kilowatts.

Each block crosses both mechanisms, three independent noise repetitions and all eight conditions, giving 576 streams. Events are at 650 and 1,100. Across months, new delay and increased noise cycle together through $(7, .375), (13, .525), (21, .75)$. This preset pairing tests heterogeneous conditions; it does not separate effects of season, delay amplitude and noise amplitude. Each month's conditions share source and standardized target innovations within a noise repetition.

The same 44 criteria are applied to each arm. Synthetic intervals resample twenty repetition clusters independently within each mechanism, retaining all their conditions and methods. Externally, method/reference ratios are first formed within each month/noise repetition, then averaged over the three repetitions. A shared bootstrap draw resamples the twelve months with all their mechanisms and conditions together. Both analyses use 10,000 draws and equal mechanism weighting. Nonoverlapping blocks from one household do not establish independence across households or eliminate seasonal dependence. These are descriptive intervals without simultaneous coverage adjustment for all methods and criteria.

The six synthetic methods meeting all criteria and all external failures remain in the complete public tables. The fixed candidate meets the synthetic criterion and misses the external criterion; no alternative is selected afterward. Quarter-rate reference noise loss is slightly lower in both arms. The adaptive/fixed paired ratios are 1.00182 $[0.99931, 1.00434]$ synthetically and 1.00184 $[1.00015, 1.00363]$ externally. Full paired comparisons and every unmet criterion remain in the numerical supplement. The latter ratio favors fixed refresh by a small amount, while both reference versions incur external stable and delay costs. Conditional delay summaries are interpreted only when active mass is positive; zero active mass leaves the joint prediction and return transition well-defined, but supplies no conditional delay identity.

E MATCHING CONDITIONAL MEMORY WITHOUT MATCHING PREDICTIONS

The diagnostic reuses all 160 development streams and training fits. Twelve filters cross two inactive means, three return rules and two refresh policies; seven deployed controls and KnownDelay are also saved. The twelve share the original target-history filter's pre-reveal variance. Their rate is $r_0/4$ or $r_0 \min(1, v_0/v_t)$; the donor retains constant r_0. These interventions differ from filters that update variance from their own predictions.

Write the post-revelation active mass as A and inactive mass as B, and let $h = xB/[(1-e)A+xB]$. For $K_r q = (1-r)q + r\mathbf{u}$, uniform return U gives $q^+ = (1-h)K_r q + h\mathbf{u}$. For the two C filters, choose the common total share

$$h_{\max} = \max(h_{\text{history}}, h_{\text{ref}}), \qquad \tau = \max\{h_{\max}, r + (1-r)h_{\max}\}, \qquad r_i = \frac{\tau - h_i}{1 - h_i}. \quad (13)$$

Table 5: Shared-scale matching, equally weighted across both mechanisms. Each Δ is (reference loss minus target-history loss), divided by the same stream's variance-scaled target-history loss, then averaged within the registered window and mechanism. Intervals resample whole training/evaluation pairs. Negative Δ favors the reference. The numerical supplement retains every mechanism-specific interval. The two matching interventions change the cancellation difference in opposite directions; noise interactions include zero.

Policy/window	Δ_U	Δ_C	Δ_S	95% CI: $\Delta_U - \Delta_C$	95% CI: $\Delta_U - \Delta_S$
Quarter/noise	-0.0005	-0.0005	-0.0004	[-0.0003, +0.0005]	[-0.0004, +0.0002]
Quarter/cancel	-0.2416	-0.2401	-0.2488	[-0.0020, -0.0009]	[+0.0059, +0.0087]
Quarter/N+cancel	-0.1986	-0.1565	-0.2102	[-0.0483, -0.0360]	[+0.0088, +0.0146]
Scaled/noise	-0.0004	-0.0005	-0.0004	[-0.0002, +0.0005]	[-0.0002, +0.0003]
Scaled/cancel	-0.2413	-0.2397	-0.2483	[-0.0020, -0.0011]	[+0.0053, +0.0091]
Scaled/N+cancel	-0.2020	-0.1578	-0.2149	[-0.0506, -0.0380]	[+0.0098, +0.0161]

When $h_i = 1$, $\tau = 1$ and uniform return determines $q^+ = \mathbf{u}$; we set its immaterial internal rate to zero. Infeasible or nonfinite τ is rejected, not clipped into apparent agreement. C uses the paired filters' revealed-label posteriors to set their next priors. It is a coupled diagnostic, not a free independent filter with unchanged computational cost. S instead returns inactive mass to $K_r q$, giving $q^+ = K_r q$ directly. All three preserve the original activity-risk transition for total mass.

Paired filters share initial conditional weights, active likelihoods and working variance; the inactive likelihood cancels from the conditional active Bayes update. C and S preserve this equality at the next prior, so induction gives equal conditional delay trajectories. Activity mass, inactive emissions and forecasts can still differ. All 640 matched stream/policy/kernel pairs agree within 2×10^{-15} for posterior and next-prior weights. Missing labels cause no update.

For each return rule, let Δ_K be reference-minus-history loss, normalized by the same stream's original adaptive target-history loss. Table 5 gives both policies, equally weighted across mechanisms. The noise row averages the four noise-containing conditions; it is not an isolated noise-only experiment. All 24 C/S cancellation intervals favor the reference. U-minus-C cancellation differences are negative, while U-minus-S differences are positive. None of eighteen registered cells meets the criterion of nonzero, same-direction changes under both matches. These interventions do not identify a universal mediation effect.

F Inactive memory and the recurrence boundary

The four expanded filters retain 24 active and 24 inactive states. Let a and i be post-revelation group vectors without within-group renormalization, and extend the refresh operator to arbitrary mass: $K_r z = (1-r)z + r(\sum_d z_d)\mathbf{u}$. The two dynamics are

$$\text{freeze: } a^+ = (1-e)K_r a + xi, \qquad i^+ = (1-x)i + ea, \tag{14}$$

$$\text{age: } a^+ = (1-e)K_r a + xK_r i, \qquad i^+ = (1-x)K_r i + ea. \tag{15}$$

Mass entering inactivity preserves its index before active refresh. Freeze does not age mass already inactive, but new entries can still change the conditional inactive delay distribution. The initial inactive mass .1 is distributed uniformly; the active prior is the original training prior. Both dynamics use the original activity risks and their own pre-reveal variance-scaled rate. All inactive states have the same prediction and likelihood, so an inactive observation does not directly distinguish their delay indices.

The four additions run alongside all original 64 methods on the complete core panel. History-freeze fails the pure-cancellation tolerance (1.053 and 1.052 times AR). The other three satisfy all 44 constraints, but none provides the preset 2% additional noise improvement over the adaptive history filter. The recurrence experiment directly tests when retaining inactive identity helps and when it costs accuracy.

The additional 80 streams have 2,400 retained observations, cancellation at 650 and revival at 1,100. Each of twenty mechanism/repetition groups shares source and standard target innovations across old-delay/new-delay returns and noise SD .15/.75. The old delay is 3; the new delay is the registered

Table 6: All recurrence conditions in a compact view. Each entry gives the mean paired 300-point MSE ratio for saturating/distributed mechanisms, relative to the adaptive single-inactive-state filter with the same inactive prediction. Core indicates the original 44-constraint result. Every pair and both 40/300-point difference intervals remain in the numerical supplement; no mechanism or noise condition is omitted here.

Inactive mean / dynamics	Same, .15	Same, .75	New, .15	New, .75	Core
History / freeze	0.925/0.949	0.999/1.005	1.068/1.055	1.006/1.003	Fail
History / age	1.000/1.000	1.000/1.004	1.000/1.000	1.001/1.002	Pass
Reference / freeze	0.913/0.940	0.978/0.979	1.084/1.061	1.017/1.014	Pass
Reference / age	1.000/1.000	0.986/0.986	1.000/1.000	1.002/1.002	Pass

Table 7: Oracle information explains delay concentration better than step count, but gives no comparable explanation of prediction recovery. Entries are held-out log-time MSE of the information explanation divided by each fixed comparator; lower is better. N/H denote new repetitions and held-out correlation $\rho = .95$. The physical comparator uses step, amplitude, noise and correlation; the proxy uses the linear AR(1) information formula. Both stronger claims fail.

Outcome	Step N	Step H	Physical N	Physical H	Proxy N	Proxy H
Prediction recovery	0.987	0.992	0.982	0.985	0.998	0.990
Delay concentration	0.399	0.403	1.161	0.799	1.030	1.041

9 or 17 for that repetition. The training fit is identical to the corresponding core fit. The boundary reports ten fixed methods: AR, KnownDelay, the two original fallback filters, the two adaptive fallback filters and the four expansions. The four expansions compare with the adaptive single-inactive-state filter using the same fallback.

The primary boundary window is the first 300 observations after revival; 40-point results describe the early cost. A same-delay cell passes if its paired mean ratio is at most .98 and its loss-difference interval is strictly negative. A new-delay cell passes if the ratio is at most 1.02 and mean extra capped recovery is at most ten. A claim across the registered boundary requires all four mechanism/noise cells, without averaging away failures, and all core constraints. Bootstrap draws keep the four conditions of each repetition together. Reference-freeze passes all four same-delay cells but only two new-delay cells; history- freeze passes two of each and also fails core cancellation. Both aging versions pass all new-delay cells but none of the same-delay benefit cells. No expansion satisfies both complete boundary claims.

G Information explains identification more than forecast recovery

Sources follow a stationary AR(1) with population variance one: $x_0 \sim \mathcal{N}(0,1)$ and $x_t = \rho x_{t-1} + \sqrt{1-\rho^2}\eta_t$. The 1,800-stream grid crosses $\rho \in \{-.5, 0, .5, .8, .95\}$, amplitude $c \in \{.5, 1, 1.5\}$, noise SD $\sigma \in \{.15, .45, .75\}$, delay step $\Delta \in \{1, 4, 10, 18\}$, both response functions and five repetitions. Amplitude multiplies the full external response. The 150 mechanism/correlation/amplitude/repetition groups share training and innovations across noise/step variants. Training retains 1,800 points with noise .15 and delay 3; evaluation retains 2,400 after 512 burn-in, changes delay at 650 and holds noise constant.

With the same actual target history, evaluate the true conditional means using the new and old delay. Define the fixed-window oracle information

$$I_{40} = \sum_{t=650}^{689} \frac{(\mu_{\mathrm{new,t}} - \mu_{\mathrm{old,t}})^2}{2\sigma^2}. \tag{16}$$

This evaluator-only quantity uses neither the realized recovery time nor any predictor output. Methods receive no true means, variance or event time. The primary tracker is the original explicit-risk filter; the six controls are fixed-scale Joint, residual EWMA, AR, KnownDelay and two amplitude controls that reuse the primary delay weights.

OLS fits $\log(1+T)$ on the first four correlations and repetitions 0–2 (864 streams); validation uses new repetitions 3–4 (576) and $\rho = .95$ (360). Step and information use $\log\Delta$ and $\log(I_{40}+10^{-8})$, adding their square for recovery but not concentration. The physical comparator uses $\log\Delta, \log c, \log\sigma, \rho$ and the same delay-square term. The proxy substitutes $40c^2(1-\rho^{\Delta})/\sigma^2$ for information. Outputs are clipped to legal time ranges. Cross-mechanism transfer does not refit a target-mechanism intercept.

The weak criterion requires 10% lower held-out log-time MSE than step on both panels, negative paired intervals and no mechanism degraded by over 5%. Only concentration passes. The stronger criterion also requires 5% improvement over both physical and proxy explanations within each mechanism and transfer direction; both outcomes fail. Concentration information/physical error is 1.161 on new repetitions, with a strictly positive difference interval. At high correlation, concentration times are capped in 10/180 saturating and 31/180 distributed streams; prediction recovery is never capped. All cells, coefficients and failed comparisons remain in the numerical supplement.

H THE INACTIVE-LEVEL ASSUMPTION AND COEFFICIENT ADAPTATION

Fourteen fixed methods comprise AR, fixed-scale Joint, residual EWMA, both original inactive means, both variance-scaled versions, quarter-rate reference, four expanded-memory filters and two RLS controls. They reuse twenty training arrays, with one fit per mechanism/repetition. New evaluation seeds produce 1,800 retained points and cancellation at 650. Each group crosses noise .15/.75 and $\ell \in \{-s_y, 0, s_y\}$, where s_y is training target population SD. Levels share targets before cancellation and innovations afterward.

For true target AR coefficient a, cancellation replaces the source term with innovation intercept $(1-a)\ell$. Relative to the zero-level path, the target shift after $k+1$ observations is $\ell(1-a^{k+1})$. Predictors receive no true coefficient, shift or event time. No feature clipping occurs; larger shifts remain untested. Intervals resample ten repetition groups within each mechanism, keeping all six conditions together.

The shared-feature RLS updates the fitted shared nonlinear map; the full RLS control updates a predictor using all lag features. Both use only revealed targets. We retain losses over the first ten, the next 290, the full 300 and all 1,150 post-cancellation observations. A first cumulative-loss crossing need not persist. A persistent crossing is the first observation at which the RLS-minus-reference cumulative loss is strictly negative and remains so through the fixed horizon. Shared-feature RLS crosses on all 80 nonzero-level streams; full RLS crosses on 76 of 80 nonzero-level streams; its four censored cases occur at high noise in the distributed mechanism. Neither RLS control persistently crosses on the forty zero-level streams. All methods, loss windows, first and persistent crossings and censored cases remain in the numerical supplement.

Learning only the inactive intercept. After observing this failure, a diagnostic reuses the 160 core and 120 level streams, retaining all 68/14 preceding controls. Six additions keep the active predictions fixed and change the inactive prediction to $b_{\mathrm{ref}}(t)+c_t$, with $c_0 = 0$. After predicting and revealing the label, update

$$c_{t+1} = c_t + \eta w_t(\widetilde{y}_t - b_{\mathrm{ref}}(t) - c_t), \qquad \eta \in \{.01, .05, .2\}. \tag{17}$$

Here $\widetilde{y}_t$ is the standardized revealed target. Posterior weighting uses the inactive probability after likelihood evaluation with the old c_t, before transition; ungated updating uses $w_t = 1$. The new intercept is used only at the next prediction. Neither rule observes the true level, event time or target AR coefficient, and missing labels cause no update.

All six pass the original 44 core guards without an additional 2% noise improvement over the adaptive history fallback. They reduce every shifted-level cell's 300-point mean loss relative to the fixed reference, but every first-ten-point cell remains worse than AR. The first five rules persistently overtake the reference on 80/80 shifted-level streams; ungated .2 crosses on 79/80, retaining one censored case. Faster learning can sacrifice zero-offset gains. For saturating, low-noise, positive-shift data at $\eta = .01$, posterior/ungated paired loss is 2.024 (descriptive 95% interval [1.775, 2.267]). Thus responsibility weighting can slow adaptation and is not generally necessary. The full numerical supplement retains all cells, intervals, crossings and zero-offset costs. This post-confirmation diagnostic neither selects a new candidate nor tests the updates on household inputs or recurrence.

Table 8: Learning the inactive offset repairs much of the shifted-level mismatch, with a transient cost. All six rules are shown. Ranges cover all eight nonzero-level mechanism/noise/sign cells, each with ten paired streams. Loss ratios below one favor the update. Fixed ref. means the zero-intercept reference with variance-scaled refresh. The core noise column compares with history-fallback adaptive refresh; all six pass 44 core guards but achieve no 2% improvement. No update is selected. These are reused development data, not an additional confirmation.

Weight	η	Core noise ratio	First 10 / AR	300 / AR	300 / fixed ref.
Posterior	.01	0.999541	1.184–4.473	0.739–1.702	0.349–0.877
Posterior	.05	0.999529	1.184–4.455	0.500–0.943	0.193–0.866
Posterior	.2	0.999553	1.174–4.402	0.464–0.853	0.164–0.916
Ungated	.01	0.999594	1.182–3.652	0.619–1.123	0.180–0.862
Ungated	.05	0.999411	1.177–3.708	0.405–0.798	0.119–0.860
Ungated	.2	1.000263	1.177–3.013	0.326–0.855	0.110–0.939

Table 9: Standardized MSE on all five previously exposed natural tasks. The fallback variants use the same activity risks and EWMA rule. The lowest of four existing neural means is shown (three initializations each). The air-quality tasks share one target and time interval.

Task	AR	AR fallback	Ref. fallback	Neural control	MSE
Air S2→S1	0.1787	0.1641	0.1649	FACT	0.1424
Air S3→S1	0.1787	0.1634	0.1630	FACT	0.1402
Air S4→S1	0.1787	0.1802	0.1870	FACT	0.1427
Air S5→S1	0.1787	0.1627	0.1637	COSA	0.1433
Appliances T2→T1	0.001105	0.001132	0.001175	LightMTS	0.000981

I NATURAL TASKS AND COMPLETE NEURAL COMPARISONS

Missing natural-task history uses past-only filling, initialized by a training median; missing target labels are excluded from fitting, scoring and updating. Both inactive means retain constant r_0 and own-residual EWMA .05. Table 10 retains every neural control and reports initialization variability. The four air tasks share target and timestamps; they are not independent datasets.

Table 10: All four neural controls on every natural task. Entries give standardized MSE mean $\pm$ sample SD across three initializations. Appliances values are displayed in units of 10^{-3} for readability. The air tasks share their target and time interval.

Task	LightMTS	DLinear–LIFT	FACT	COSA
Air S2→S1	0.1439 ± 0.0031	0.1552 ± 0.0016	0.1424 ± 0.0049	0.1436 ± 0.0030
Air S3→S1	0.1423 ± 0.0013	0.1571 ± 0.0039	0.1402 ± 0.0049	0.1420 ± 0.0013
Air S4→S1	0.1467 ± 0.0027	0.1568 ± 0.0014	0.1427 ± 0.0083	0.1463 ± 0.0024
Air S5→S1	0.1435 ± 0.0029	0.1573 ± 0.0006	0.1439 ± 0.0096	0.1433 ± 0.0030
Appliances ($\times 10^{-3}$)	0.9810 ± 0.0061	1.2747 ± 0.2441	0.9877 ± 0.0443	0.9817 ± 0.0061