Journal of Science and Technology – Data Science and Artificial Intelligence
Volume 36, Issue 1, Oct 2026, 009-017

Post-Deployment Monitoring of Prediction-Error Deterioration in Machine Learning Models Using a Two-Channel Exponentially Weighted Moving Average Chart

Ba Quynh Pham[1], Huu Du Nguyen[2], Anh Son Ta[2], Thanh Hoa Le[3,*]

[1]*Nguyen Truong To Secondary School, Ha Noi, Vietnam*
[2]*Hanoi University of Science and Technology, Ha Noi, Vietnam*
[3]*Faculty of Basic Sciences, Vietnam Maritime University, Hai Phong, Vietnam*
[*]*Corresponding author email: lethanhhoa@vimaru.edu.vn*

Abstract

This study develops a forward-cross-fitted two-channel EWMA chart for post-deployment monitoring of prediction-error deterioration in continuous-response machine-learning models. A temporally valid Phase-I residual reference is first constructed, after which the resulting errors are transformed into complementary signed and magnitude channels. These channels are combined through an equal-share rule with separately calibrated finite-horizon limits. The proposed method is evaluated through a unified Monte Carlo study covering six deterioration mechanisms with comparable increases in mean squared prediction error. The results show that forward cross-fitting substantially reduces false-alarm inflation relative to an in-sample residual reference, while the signed and magnitude channels respond to different forms of deterioration. The equal-share construction also improves directional sensitivity relative to a common-limit rule while retaining information from both channels. In addition, false-alarm calibration is shown to be particularly affected by serial dependence, bootstrap block length, and monitoring horizon. In contrast, the mismatch between the forward-cross-fitted reference and the final deployed model has only a modest effect in the controlled setting. To further examine the practical behavior of the chart, we consider two real-data applications. The ethanol sensor study illustrates reference sensitivity and channel attribution, while the Metro traffic study provides a controlled stress test on actual post-deployment residual sequences from a different regression domain.

Keywords: EWMA, control chart, cross-fitting, prediction-error monitoring, statistical process monitoring.

1. Introduction

Machine-learning (ML) models are increasingly deployed in systems that operate continuously. After deployment, changes in sensors, operating conditions, populations, or physical mechanisms may degrade predictive performance [1]. Residual charts observe the prediction-error process directly. A change in that process can be caused by a change in the conditional predictor–response relationship, noise inflation, sensor degradation, covariate shift into poorly learned regions, or other model mismatch. In this study, *prediction-error deterioration* denotes the monitored quantity, while *predictive-relationship change* denotes one possible mechanism.

Residual monitoring connects naturally to statistical process control (SPC). EWMA charts provide persistent-change sensitivity [2], while joint monitoring of location and dispersion has been studied through one-chart EWMA and Max-EWMA constructions [3, 4]. Profile monitoring similarly treats a response–predictor relationship as the monitored process characteristic [5]. These classical results are important here because the proposed chart also combines directional and magnitude evidence, but the monitored observations are prediction residuals from a black-box ML predictor instead of direct process measurements. Phase-I estimation is another relevant issue: estimated control limits can materially alter Phase-II behavior [6], and bootstrap adjustments have been proposed to improve conditional chart performance [7].

In ML monitoring, Ross *et al.* [8] adapt EWMA ideas to classifier drift, ADWIN compares adaptive windows [9], and Zhang *et al.* [10] monitor Fisher score vectors with multivariate EWMA (MEWMA). Wu and Apley [11] further study bootstrap control limits for the score-based setting. These approaches require either classifier-specific quantities or a model-specific probabilistic score. Pyeon *et al.* [12] address localized regression drift under label scarcity, whereas Wang *et al.* [13] combine outlier handling, dynamic residual monitoring, drift typing, and adaptation. Our setting assumes that responses eventually become available and that the deployment predictor remains fixed during monitoring.

The first design issue is the Phase-I residual reference. Training residuals from a flexible learner may be too optimistic for prospective monitoring. Cross-fitting separates fitting from evaluation in

p-ISSN 3093-3331
e-ISSN 3093-3323
https://doi.org/10.51316/jst.194.dsai.2026.36.1.2
Received: Aug 20, 2026; Revised: Sep 28, 2026;
Accepted: Oct 2, 2026; Online: Oct 15, 2026

complex change-point problems [14]; for ordered data, the corresponding reference should also respect time. The second issue is that signed residuals and residual magnitude respond to different forms of deterioration. Finally, the two resulting EWMAs need not have the same finite-horizon null distribution even after input standardization, so a common control limit can favor one channel. We therefore use forward cross-fitting, robust channel construction, and an *equal-share* joint rule with separate channel limits. The common marginal exceedance probability is selected by bootstrap so that the overall finite-horizon alarm probability reaches its design target. The chronological forward-CF construction follows rolling-origin evaluation principles [15].

No new theoretical property is claimed for the resulting chart or its components. The contribution is the integration of three elements for fixed-model regression monitoring: (i) a chronological forward-CF residual reference, (ii) channel-specific equal-share finite-horizon calibration for complementary signed and magnitude EWMAs, and (iii) a controlled evaluation that separates reference construction, channel specialization, calibration error, and temporary-model/final-model mismatch. The real-data analysis combines an unlabeled gas-sensor diagnostic with a paired stress test on actual post-Phase-I Metro traffic residual windows.

2. Proposed Monitoring Method

2.1. Problem Formulation

Let $\mathscr{D}_0 = \{(X_t, Y_t) : t = 1, \ldots, n_0\}$ be the stable historical dataset, where $X_t \in \mathbb{R}^p$ is the predictor vector and $Y_t \in \mathbb{R}$ is a continuous response. A learning algorithm $\mathscr{A}$ fitted to all historical data produces the deployment predictor $\widehat{f}_0 = \mathscr{A}(\mathscr{D}_0)$. When the true response corresponding to a deployed prediction becomes available, the residual is

$$e_t = Y_t - \widehat{f}_0(X_t),\ t > n_0 \tag{1}$$

The direct monitoring target is persistent departure of this residual process from its Phase-I behavior. An alarm does not, by itself, identify whether the cause is conditional-mean drift, increased noise, sensor degradation, covariate shift, or another source of model mismatch.

2.2. Forward-Cross-Fitted Reference Residuals

If the final predictor is evaluated on the same observations used for fitting, the in-sample residuals

$$e_t^{\mathrm{IN}} = Y_t - \widehat{f}_0(X_t),\ t \le n_0 \tag{2}$$

can be overly optimistic. We instead divide Phase I chronologically into an initial block R_0 and subsequent blocks $R_1, \ldots, R_K$. For each R_k, a temporary predictor is fitted only on earlier observations,

$$\widehat{f}_k = \mathscr{A}(R_0 \cup \cdots \cup R_{k-1}) \tag{3}$$

and the reference residuals are

$$e_t^{\mathrm{CF}} = Y_t - \widehat{f}_k(X_t),\ t \in R_k \tag{4}$$

The residuals from $R_1, \ldots, R_K$ are concatenated in chronological order. The labeled observations, not the residuals, are used at each later fitting step. After the reference sequence is formed, $\widehat{f}_0$ is refitted on all Phase-I observations. Thus the reference is a prospective proxy and is not an exact draw from the final-model residual distribution.

2.3. Robust Standardization and Residual Channels

Let $\mathscr{E}_0$ denote the forward-cross-fitted reference residuals. Their robust location and scale are

$$\begin{aligned} \widehat{\mu}_e &= \mathrm{median}(\mathscr{E}_0) \\ \widehat{\sigma}_e &= 1.4826\ \mathrm{median}\left|e_t^{\mathrm{CF}} - \widehat{\mu}_e\right| \end{aligned} \tag{5}$$

where the factor 1.4826 gives the normal-consistent median absolute deviation (MAD) scaling [16]. A deployment residual is standardized as

$$r_t = \frac{e_t - \widehat{\mu}_e}{\widehat{\sigma}_e} \tag{6}$$

The signed and magnitude transforms are

$$\begin{aligned} a_t^{(0)} &= \psi_c(r_t) = \max\{-c, \min(r_t, c)\}, \\ b_t^{(0)} &= \min(r_t^2, c^2) \end{aligned} \tag{7}$$

The signed transform is clipped to the interval $[-c, c]$, whereas the magnitude transform is nonnegative and bounded above by c^2. Let (μ_a, s_a) and (μ_b, s_b) be the means and standard deviations of these transforms on the reference sequence. The standardized channel inputs are

$$a_t = \frac{a_t^{(0)} - \mu_a}{s_a},\ b_t = \frac{b_t^{(0)} - \mu_b}{s_b} \tag{8}$$

The signed channel targets persistent over-or underprediction, whereas the magnitude channel can respond when positive and negative errors grow without a stable signed bias.

2.4. Equal-Share Two-Channel Exponentially Weighted Moving Average and Finite-Horizon Calibration

The two EWMAs are

$$A_t = \lambda a_t + (1-\lambda)A_{t-1},\ B_t = \lambda b_t + (1-\lambda)B_{t-1} \tag{9}$$

with $A_0 = B_0 = 0$. Because $|A_t|$ is two-sided and $B_t^+ = \max(B_t, 0)$ is one-sided and right-skewed, input standardization does not imply equal finite-horizon tail behavior. We therefore calibrate separate channel limits.

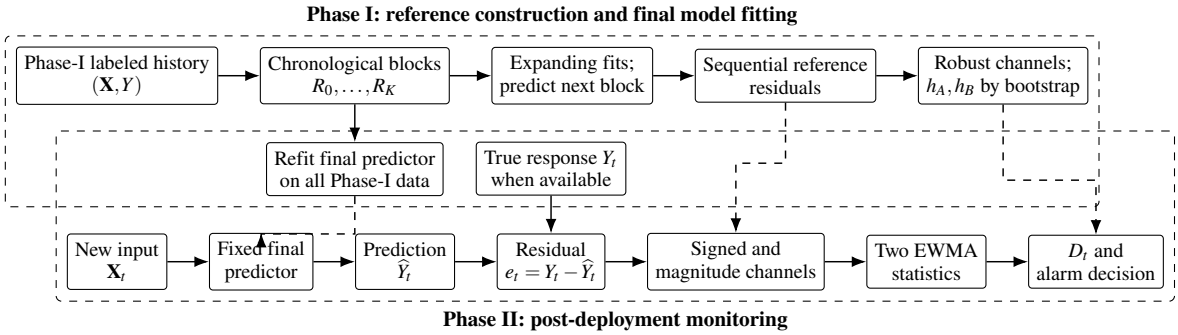

Fig. 1. Phase-I/Phase-II workflow. Forward cross-fitting constructs the ordered residual reference; the final predictor is then refitted on all Phase-I data. The two deployment channels use separately calibrated finite-horizon limits

For a bootstrap path, define

$$M_A = \max_{1\le t\le H} |A_t^*|,\ M_B = \max_{1\le t\le H} (B_t^*)_+ \tag{10}$$

For a candidate channel-wise exceedance level α_{ch}, let

$$h_A(\alpha_{\text{ch}}) = Q_{1-\alpha_{\text{ch}}}(M_A),\ h_B(\alpha_{\text{ch}}) = Q_{1-\alpha_{\text{ch}}}(M_B) \tag{11}$$

The same α_{ch} is used for both channels, and a one-dimensional search selects α_{ch}^* so that the empirical joint finite-horizon probability

$$\begin{aligned}\widehat{P}_0\Big(&\max_{1\le t\le H} |A_t^*| > h_A(\alpha_{\text{ch}}^*)\\ &\text{or } \max_{1\le t\le H} (B_t^*)_+ > h_B(\alpha_{\text{ch}}^*)\Big) \approx \alpha\end{aligned} \tag{12}$$

The proposed statistic is therefore

$$D_t = \max\left\{\frac{|A_t|}{h_A}, \frac{B_t^+}{h_B}\right\},\ \text{alarm if } D_t > 1 \tag{13}$$

Because the alarm rule uses a strict inequality, $D_t = 1$ is not treated as an alarm. We call this the equal-share rule because both channels receive the same marginal exceedance probability, while their numerical limits remain channel specific. Circular blocks are used when serial dependence is present. Each bootstrap path contains a warm-up segment of length $m = |\mathscr{E}_0|$ followed by a monitoring segment of length H; the robust residual standardization and channel centering/scaling are re-estimated from the warm-up segment. For comparison, a common-limit maximum and a Bonferroni construction with marginal level $\alpha/2$ are included as ablations.

3. Simulation Study

All principal Monte Carlo comparisons use $R = 1000$ replications with common random numbers across methods, $N_{\text{boot}} = 1000$ bootstrap paths, and separate in-control samples for false-alarm calibration ($R_{\text{cal}} = 2000$) and verification ($R_{\text{ver}} = 1000$). The same baseline configuration is reused in the one-factor-at-a-time (OFAT) calibration study. Throughout this section, PFA denotes the finite-horizon false-alarm probability.

3.1. Data-Generating Process and Matched Deterioration Severity

Predictors follow stationary AR(1) processes,

$$X_{j,t} = \rho_j X_{j,t-1} + \eta_{j,t},\ j = 1,2 \tag{14}$$

where $\eta_{j,t} \sim N(0, \sigma_{x,j}^2)$, and

$$\begin{aligned}Y_t &= m(X_t) + \varepsilon_t\\ m(X_t) &= \beta_1 \sin X_{1,t} + \beta_2 X_{2,t}^2 + \beta_3 X_{1,t} X_{2,t}\end{aligned} \tag{15}$$

with $\varepsilon_t = \phi\varepsilon_{t-1} + \zeta_t$ and $\zeta_t \sim N(0, \sigma_\zeta^2)$. Six post-change mechanisms are considered: abrupt and gradual location shifts, a centered interaction change, a change in the sinusoidal coefficient, noise-scale inflation, and a combined location–scale change.

To avoid comparing mechanisms that merely use similarly sized coefficients, all changes are indexed by the same target increase in mean squared error (MSE), $q = \Delta\text{MSE}$. Let $v_j = \sigma_{x,j}^2/(1-\rho_j^2)$ and $v_\varepsilon = \sigma_\zeta^2/(1-\phi^2)$. The abrupt shift is $\delta_L = \sqrt{q}$; the gradual shift reaches the same final value over L_G observations; the interaction change is $\delta_R = \sqrt{q/(v_1 v_2)}$. The sinusoidal-coefficient change uses $\delta_C = \sqrt{q/[\sin^2(X_1)]}$, with $[\sin^2(X_1)] = \{1 - e^{-2v_1}\}/2$, and the noise multiplier is $\gamma_V = \sqrt{1 + q/v_\varepsilon}$. For the mixed mechanism, half of q is assigned to location and half to scale: $\delta_{LS} = \sqrt{q/2}$ and $\gamma_{LS} = \sqrt{1 + (q/2)/v_\varepsilon}$. In the study, we use $q \in \{0.5, 1.0, 1.5\}$. At $q = 1$, the corresponding numerical values are approximately 1.000, 0.763, 1.446, 1.370, and $(0.707, 1.199)$.

3.2. Ablations, Comparators, and Performance Measures

The residual variants isolate reference construction, robust clipping, and channel choice (Table 2). IN-EWMA and CF-EWMA are signed single-channel charts without clipping; they differ only in the residual reference. FCF-R and FCF-M are the clipped signed-only and magnitude-only counterparts. FCF-DS uses both channels with the equal-share rule in (13). Page-Hinkley [17] and ADWIN [9] are external stream-monitoring benchmarks.

Table 1. Unified simulation settings

Component	Setting
Predictor persistence	$\rho_1 = 0.60,\ \rho_2 = 0.30$
Noise persistence	$\phi = 0.35$
Stable response	$(\beta_1, \beta_2, \beta_3) = (1.5, 0.8, 0.5)$
Innovation scales	$\sigma_{x,1} = \sigma_{x,2} = \sigma_\zeta = 1$
Phase-I length	$n_0 = 1200$
Forward cross-fitting	400 initial; four blocks of 200
Monitoring	$H = 300,\ \tau = 51,\ L_G = 50$
Severity	$\Delta\text{MSE} \in \{0.5, 1, 1.5\}$
Base learner	XGBoost: 60 trees, depth 3; learning rate 0.07
FCF parameters	$\lambda = 0.10,\ c = 2.5,\ \alpha = 0.05$
Bootstrap	circular $L_B = 10$, $N_{\text{boot}} = 1000$
Monte Carlo	$R = 1000$
PFA calibration/verification	$R_{\text{cal}} = 2000$, $R_{\text{ver}} = 1000$

Table 2. Residual-monitoring ablations

Method	Reference	Clip	Channel(s)
IN-EWMA	in-sample	no	signed
CF-EWMA	forward-CF	no	signed
FCF-R	forward-CF	yes	signed
FCF-M	forward-CF	yes	magnitude
FCF-DS	forward-CF	yes	signed + magnitude

A recent residual-based regression detector, EWMAD-DT of Wang *et al.* [13], was evaluated for calibration compatibility. The published detector uses the dynamic threshold $\Theta_t = \xi\theta_t$ with $\xi \in (0,1)$. In the fixed-predictor finite-horizon setting considered here, the largest admissible multiplier ($\xi \approx 1$) produced a verified PFA of 0.425. Achieving a PFA near 0.05 required $\xi = 5.034$, which lies outside the published parameter range and therefore represents a different operating regime. Table 3 reports this calibration-compatibility diagnostic.

Table 3. EWMAD-DT calibration diagnostic over $H = 300$

Setting	ξ	Ver. PFA	95% CI
Published-range cap	1.000	0.425	[0.395, 0.456]
Out-of-range adaptation	5.034	0.062	[0.049, 0.079]

At $\xi = 5.034$, all 62 in-control alarms occurred at $t = 2$–4, showing that this out-of-range operating point is dominated by start-up behavior. Wang *et al.* reset the running residual mean and related reset-state variables after a drift, but do not state a warm-start or explicit initial value for θ. In our fixed-model adaptation we therefore used the zero-state initialization $S_0 = \bar{R}_0 = \theta_0 = 0$, consistent with the reset convention for the running statistics. This makes $\Theta_t = \xi\theta_t$ very small at the beginning. After the transient, the large multiplier together with the adaptive threshold makes the detector effectively insensitive in our setting. The exploratory $\Delta\text{MSE} = 1$ runs consequently produced 0/1000 post-change detections for each mechanism; the exact one-sided 95% binomial upper bound is 0.0030. We therefore exclude this extrapolated point from matched-PFA power comparisons and retain it only as a calibration-compatibility diagnostic.

Let T denote the first alarm time. We use PFA, detection probability P_D, and conditional detection delay (CDD), defined as

$$\text{PFA} = P(T \le H \mid H_0),\ P_D = P(\tau < T \le \tau + 100) \quad (16)$$

and

$$\text{CDD} = [T - \tau \mid \tau < T \le \tau + 100] \quad (17)$$

For post-change comparisons, each method is calibrated on 2000 independent in-control replications and verified on 1000 new in-control replications. Common random numbers are used across methods and mechanisms. By convention, P_D counts alarms strictly after the onset index τ and within the next 100 monitoring observations, matching the implementation used for all methods. Conditional detection delay (CDD) is reported together with P_D in Table 5; CDD is interpreted only among runs detected within this post-change window.

3.3. Native False-Alarm Behavior

Table 4 reports native finite-horizon PFA before the independent PFA-matching step. "Independent errors" means $\phi = 0$ while retaining the AR predictor processes. Forward cross-fitting markedly reduces the inflation created by in-sample residuals, but the one-level bootstrap does not provide exact nominal control for FCF-DS.

Table 4. Native empirical PFA over $H = 300$ ($R = 1000$). Monte Carlo standard errors (MCSEs) are 0.002–0.01

Method	Independent errors	AR(1) errors	AR(1)+2% contamination
IN-EWMA	0.196	0.180	0.176
CF-EWMA	0.093	0.087	0.126
FCF-R	0.079	0.090	0.100
FCF-M	0.084	0.076	0.070
FCF-DS	0.095	0.097	0.094
Page-Hinkley	0.053	0.061	0.053
ADWIN	0.005	0.040	0.097

3.4. Detection After Independent False-Alarm Probability Calibration

Table 5 reports the $q = \Delta\text{MSE} = 1$ comparison. CDD is reported conditionally on detection. When P_D is low, the corresponding CDD is based on a small selected subset of runs and may be influenced by chance-timed alarms; such values are therefore interpreted cautiously and are not used for direct cross-method ranking. The verified PFAs are not identical: they range from 0.051 to 0.067 for the residual EWMA variants, with MCSE about 0.007–0.008. We therefore do not interpret small P_D differences as method superiority. The main conclusion is channel specialization: FCF-R is strongest for directional shifts and the mixed change, whereas FCF-M is strongest for centered interaction, nonlinear-coefficient, and noise-scale deterioration. FCF-DS does not dominate either specialist; it reduces

Table 5. Detection at $\Delta\text{MSE} = 1$ ($R = 1000$). Panel A gives verified PFA and P_D (P_D MCSEs are at most 0.016); Panel B gives conditional detection delay (CDD), in monitoring observations, among detections within $(\tau, \tau + 100]$.

Panel A: verified PFA and detection probability P_D

Method	Ver. PFA	Abrupt	Gradual	Interaction	Coef.	Noise	Mix
IN-EWMA	0.067	0.702	0.547	0.276	0.197	0.086	0.459
CF-EWMA	0.066	0.681	0.544	0.280	0.191	0.095	0.457
FCF-R	0.051	0.785	0.644	0.120	0.229	0.116	0.509
FCF-M	0.055	0.398	0.293	0.553	0.457	0.295	0.370
FCF-DS	0.064	0.724	0.573	0.480	0.426	0.250	0.488
Page-Hinkley	0.035	0.388	0.302	0.021	0.028	0.018	0.160
ADWIN	0.038	0.471	0.439	0.059	0.072	0.030	0.180

Panel B: conditional detection delay (CDD)

Method	Abrupt	Gradual	Interaction	Coef.	Noise	Mix
IN-EWMA	45.3	65.1	49.9	48.0	49.9	48.5
CF-EWMA	43.3	63.9	48.4	46.5	49.4	47.6
FCF-R	43.0	64.9	52.2	50.1	50.2	48.3
FCF-M	47.5	65.9	44.6	45.9	47.4	48.6
FCF-DS	43.7	65.0	45.3	46.3	47.4	48.1
Page-Hinkley	71.8	80.6	74.2	77.6	79.0	73.0
ADWIN	54.5	71.5	53.3	63.5	56.5	57.1

the penalty when the informative residual characteristic is unknown. Across the three tested severities, its approximate worst deficits relative to the better specialist are 0.059, 0.073, and 0.083, compared with 0.161, 0.203, and 0.182 for the common-limit maximum.

Table 6. Sensitivity of FCF-DS to matched deterioration severity ($R = 1000$).

Mechanism	$\Delta\text{MSE} = 0.5$	$\Delta\text{MSE} = 1.0$	$\Delta\text{MSE} = 1.5$
Abrupt location	0.343	0.724	0.893
Gradual location	0.246	0.573	0.813
Centered interaction	0.298	0.480	0.642
Nonlinear coefficient	0.200	0.426	0.683
Noise-scale inflation	0.101	0.250	0.457
Location+scale	0.205	0.488	0.710

3.5. Combination-Rule and Tuning-Parameter Sensitivity

The channel-combination ablation evaluates whether a common numerical limit disproportionately penalizes the signed channel (Table 7). Equal-share raises abrupt-shift P_D from 0.582 to 0.724 while verified PFA rises only from 0.055 to 0.064. As a scale check, the independent baseline in Table 8 differs from Table 5 by 0.018 in PFA (0.046 versus 0.064) but by only 0.046 in abrupt P_D (0.678 versus 0.724); thus the 0.142 common-limit/equal-share gap is substantially larger than the difference associated with that PFA shift, although the two experiments use independent streams and this is not a formal PFA adjustment. This improvement comes with a trade-off: for centered interaction, the common-limit rule performs better (0.523 versus 0.480) despite its lower PFA. Bonferroni is nearly identical to equal-share. We retain equal-share because it estimates the common channel-wise exceedance level jointly rather than fixing it at $\alpha/2$.

Sensitivity to λ and c is assessed after separate PFA calibration for each setting (Table 8). The sensitivity experiment uses an independent calibration and replication stream, so the baseline row $\lambda = 0.10, c = 2.5$ is an independent estimate and is not expected to

Table 7. Dual-channel combination rules at $\Delta\text{MSE} = 1$; each rule is independently calibrated

Rule	PFA	Abrupt	Gradual	Interaction	Coef.	Noise	Mix
Common limit	0.055	0.582	0.452	0.523	0.450	0.281	0.411
Equal-share	0.064	0.724	0.573	0.480	0.426	0.250	0.488
Bonferroni	0.063	0.722	0.574	0.477	0.427	0.247	0.486

reproduce Table 5 exactly. Verified PFA ranges from 0.037 to 0.048. The smaller $\lambda = 0.05$ is more responsive to the sustained changes considered. The clipping threshold changes the direction–magnitude balance: $c = 2.0$ favors abrupt location change, whereas larger c retains more information for the centered interaction. Because $c = 3.0$ also has the lowest verified PFA (0.037), its lower abrupt P_D cannot be attributed to clipping alone.

Table 8. Equal-share sensitivity to λ and c at $\Delta\text{MSE} = 1$

Factor	Value	PFA	Abr.	Inter.	Noise
λ	0.05	0.042	0.863	0.555	0.317
λ	0.10	0.046	0.678	0.465	0.235
λ	0.20	0.048	0.416	0.353	0.168
c	2.0	0.047	0.718	0.376	0.238
c	2.5	0.046	0.678	0.465	0.235
c	3.0	0.037	0.619	0.523	0.224

3.6. Calibration Sensitivity and Phase-I Estimation

A common-random-number OFAT experiment with $R = 1000$ is summarized in Table 9. The clearest drivers of native PFA inflation are serial dependence, failure to preserve dependence when $L_B = 1$, and a longer monitoring horizon. The reference-length results are nonmonotone and their intervals overlap substantially. Increasing N_{boot} from 250 to 2000 has negligible effect. These results point first to dependence reproduction, and only secondarily to finite-Phase-I estimation, as the dominant practical calibration issue in this data-generating process.

A residual-only double-bootstrap diagnostic was also used to examine Phase-I uncertainty. With equal-share and nominal $\alpha = 0.05$, the mean outer PFA is 0.105 and its 90th percentile is 0.247. Lowering the inner target to 0.02 reduces the mean to 0.061 but leaves the

Table 9. OFAT native-PFA sensitivity for equal-share ($R = 1000$); intervals are Wilson 95% CIs

Factor	Level	PFA	95% CI
ϕ	0	0.091	0.075–0.110
	0.20	0.090	0.074–0.109
	0.35	0.105	0.087–0.126
	0.50	0.124	0.105–0.146
	0.70	0.191	0.168–0.217
L_B	1	0.371	0.342–0.401
	5	0.102	0.085–0.122
	10	0.105	0.087–0.126
	20	0.140	0.120–0.163
n_0	600	0.122	0.103–0.144
	1200	0.105	0.087–0.126
	2400	0.110	0.092–0.131
H	150	0.074	0.059–0.092
	300	0.105	0.087–0.126
	600	0.146	0.125–0.169
N_{boot}	250	0.110	0.092–0.131
	500	0.108	0.090–0.129
	1000	0.105	0.087–0.126
	2000	0.104	0.087–0.124

90th percentile at 0.157. Thus a simple residual-only correction improves average calibration but does not provide reliable conditional control across Phase-I references. We therefore treat α as a design target, report the achieved PFA, and make no exact conditional-control claim.

3.7. Reference/Deployment Mismatch Diagnostic

The forward-CF reference is produced by temporary models, whereas deployment residuals come from the final model fitted on all Phase-I observations. To measure this discrepancy directly, we repeated the equal-share analysis for $R = 1000$ replications using either the practical forward-CF reference or an oracle reference generated by the final model on an independent stable stream of the same length.

The distributional mismatch is small on average. The median difference (CF minus oracle) is -0.001 (95% CI $[-0.009, 0.006]$), the robust-scale ratio is 1.012 (1.008–1.015), and the centered 95th-absolute-residual quantile ratio is 1.019 (1.015–1.022). The corresponding ratios of equal-share limits are 1.006 for h_A and 1.008 for h_B. Because the forward-CF reference is slightly more dispersed, its native limits are marginally higher. This gives a lower native PFA (0.105 vs. 0.124) and native P_D values that are 0–3.6 percentage points lower across the six $q = 1$ mechanisms (Table 10). These differences reflect a small shift in operating point; they do not establish a sensitivity advantage for either reference.

Table 10. Reference/deployment mismatch at $\Delta\text{MSE} = 1$ ($R = 1000$), using native limits; Table 5 gives PFA-calibrated results

Mechanism	CF	Oracle	Diff. (95% CI)
Abrupt location	0.844	0.860	-0.016 $[-0.038, 0.006]$
Gradual location	0.717	0.717	0.000 $[-0.026, 0.026]$
Centered interaction	0.589	0.621	-0.032 $[-0.061, -0.003]$
Nonlinear coefficient	0.556	0.588	-0.032 $[-0.060, -0.004]$
Noise-scale inflation	0.351	0.387	-0.036 $[-0.063, -0.009]$
Location+scale	0.634	0.670	-0.036 $[-0.064, -0.008]$

4. Real-Data Studies

4.1. Gas-Sensor Illustration

We use the ethanol subset of the UCI Gas Sensor Array Drift at Different Concentrations data [18–20]. The response is $Y = \log(1 + C)$, where C is ethanol concentration, and the predictors are the 128 supplied sensor-response features. Batches 1–3 (399 observations) form the operational Phase-I reference and Batches 4–10 (1,242 observations) form Phase II. Batch 1 predicts Batch 2 and Batches 1–2 predict Batch 3, producing 316 forward-CF residuals. Exact within-batch timestamps are unavailable, so the empirical bootstrap uses block length 1 and $H = 1242$. XGBoost uses 120 trees, maximum depth 3, learning rate 0.05, subsample and column-subsample fractions 0.90, ℓ_2 regularization 1, and minimum child weight 3 [21].

For equal-share, $\widehat{\mu}_e = -0.199$ and $\widehat{\sigma}_e = 0.161$, with $h_A = 0.951$, $h_B = 1.348$, and $\alpha^*_{\text{ch}} = 0.0275$. The negative reference center is important for interpretation: a deployment residual near zero is standardized to approximately $(0 + 0.199)/0.161 = 1.23$. The early Batch-4 crossing is therefore interpreted as a reference-mismatch warning rather than evidence of prediction deterioration. Batch 4 has a small log-scale mean absolute error (MAE; 0.087), and the early signed excursion is consistent with a location mismatch between the temporary-model reference and the final-model deployment residuals. This real-data example is therefore more severe than the controlled mismatch experiment and illustrates why Phase-I stability and reference location must be checked in practice.

Table 11. Ethanol prediction-error diagnostics. RMSE denotes root mean squared error; the reference row reports the forward-CF residual sequence, and clipping proportions are reported for Phase-II batches

Period	n	MAE	RMSE	$P(\lvert r\rvert > c)$
Reference (forward-CF)	316	0.195	0.239	–
Batch 4	12	0.087	0.113	0.000
Batch 5	20	0.105	0.163	0.050
Batch 6	110	0.222	0.282	0.591
Batch 7	360	0.768	0.916	0.925
Batch 8	40	0.431	0.552	0.925
Batch 9	100	0.125	0.138	0.000
Batch 10	600	0.739	0.839	0.802

Batches 7, 8, and 10 have clearly larger errors (MAE 0.431–0.768), and 80–93% of their standardized residuals exceed c. Batch 6 is intermediate: its MAE (0.222) is close to the forward-CF reference MAE (0.195), but the negative reference center means that a residual of approximately $+0.20$ already gives $r_t \approx c = 2.5$. Its exceedance proportion of 0.591 therefore partly reflects reference-location mismatch and should not be interpreted as large prediction error alone. Batch 9 has no new $|r_t| > c$ observations, but the magnitude EWMA decays slowly because it retains memory from Batches 7–8. Fig. 2 and Table 11 are used for channel attribution and are not labeled estimates of P_D or delay.

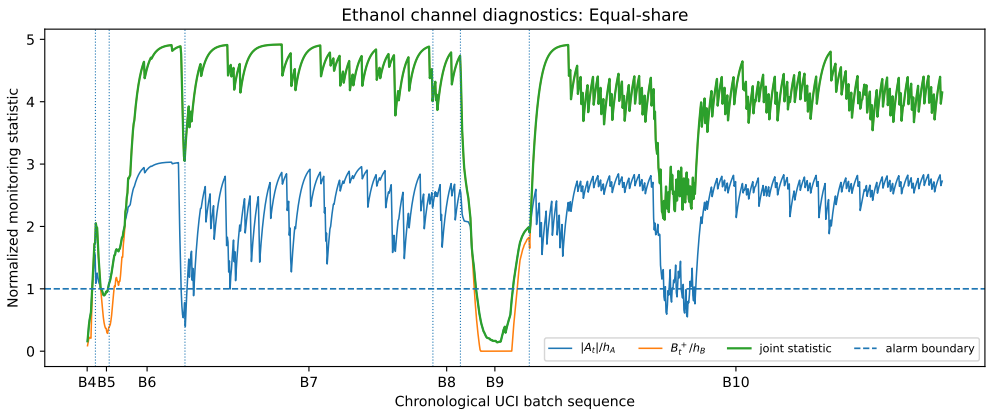

Fig. 2. Equal-share channel diagnostics for the ethanol monitoring sequence. The signed and magnitude EWMAs are divided by their channel-specific limits; the joint statistic is their maximum and the dashed line at one is the alarm boundary. Dotted lines mark batch boundaries

4.2. Metro Traffic Stress Test

The second domain is the UCI Metro Interstate Traffic Volume data [22]. After duplicate timestamps are collapsed, 40,575 hourly records remain. The response is $Y = \log(1 + \texttt{traffic_volume})$. Predictors are temperature, hourly rain and snow, cloud cover, a holiday indicator, and sine/cosine encodings of hour, day of week, and month; the text weather labels are not used. The first 12,000 observations form Phase I. Expanding forward cross-fitting uses 4,000 initial observations followed by four blocks of 2,000, yielding 8,000 reference residuals. Their log-scale MAE is 0.137, with $\widehat{\mu}_e = 0.022$ and $\widehat{\sigma}_e = 0.132$; 0.081 of the standardized reference residuals satisfy $|r| > c = 2.5$. Monitoring uses $L_B = 10$, $H = 300$, $\lambda = 0.10$, and $c = 2.5$. XGBoost uses the same 120-tree real-data specification as in the ethanol study. The final model is fitted on all Phase-I data and evaluated on the remaining 28,575 observations. These observations yield 95 complete non-overlapping windows of length $H = 300$; the final 75 observations are not included in the windowed stress-test analysis.

The unmodified windows are not assumed in the control. Equal-share alarms in 26/95 windows (0.274, 95% CI 0.194–0.371); FCF-R alarms in 26/95 and FCF-M in 24/95. For equal-share, the alarm proportions across five consecutive groups of 19 windows are 0.474, 0.211, 0.158, 0.263, and 0.263, so the elevated deployment alarm rate is not a monotone function of distance from Phase I. Of the 26 equal-share alarm windows, 20 are also alarmed by both single-channel charts, three coincide only with FCF-R, and three only with FCF-M. These values are deployment alarm rates, not false-alarm probabilities, because natural nonstationarity may be present.

For a controlled paired stress test, perturbations are injected at observation 51 of each actual residual window. Across the four methods reported in Table 12, five of the 95 windows contain at least one pre-injection alarm; these windows are removed for every method, leaving a common set of 90 windows for paired comparison. The prespecified moderate perturbations are a location shift of $1.5\widehat{\sigma}_e$, a temperature-linked perturbation $1.25\widehat{\sigma}_e Z_{\text{temp,t}}$, and residual-scale inflation by a factor of 2. The baseline column is the alarm proportion over the same 100-observation post-change window with no injection.

The scale perturbation is applied to the raw residual, $e_t^* = \gamma e_t$, so it scales about zero rather than about the forward-CF center. Under the same reference standardization,

$$r_t^* = \gamma r_t + (\gamma - 1)\widehat{\mu}_e/\widehat{\sigma}_e \tag{18}$$

Here $\widehat{\mu}_e/\widehat{\sigma}_e \approx 0.17$, so for $\gamma = 2$ the deterministic signed displacement is only about 0.17 robust standard deviations. The signed channel nevertheless responds to scale inflation (FCF-R increase: +0.189), because scaling about zero also doubles any signed departure already present in an actual deployment window. The reference also shows substantial tail mass: the MAE-to-robust-scale ratio is $0.137/0.132 \approx 1.04$, compared with $\sqrt{2/\pi} \approx 0.80$ for a centered Gaussian, and 8.1% of the reference residuals already satisfy $|r| > 2.5$. Clipping can therefore limit some additional information available to the magnitude channel. Even so, the magnitude-only chart and the equal-share chart respond at least as strongly as the signed chart in this experiment (FCF-M 0.356, FCF-DS equal-share 0.400, FCF-R 0.311); with only 90 windows, however, the FCF-M–FCF-R difference corresponds to four windows and should not be overinterpreted. We therefore interpret the Metro scale experiment as a residual-amplitude stress test rather than a pure conditional-variance perturbation; clean magnitude specialization is assessed in the controlled noise-scale simulation.

Across the three perturbations, the Metro results reproduce the qualitative pattern of the simulation: the signed channel dominates the location and predictor-

Table 12. Paired Metro stress test on 90 common post-Phase-I windows. Injected alarm proportions are followed by paired increases over baseline in parentheses

Method	Baseline	Abrupt	Relationship	Scale
FCF-R	0.122	0.756 (+0.633)	0.522 (+0.400)	0.311 (+0.189)
FCF-M	0.122	0.244 (+0.122)	0.200 (+0.078)	0.356 (+0.233)
Common-limit DS	0.122	0.344 (+0.222)	0.267 (+0.144)	0.356 (+0.233)
FCF-DS equal-share	0.133	0.744 (+0.611)	0.489 (+0.356)	0.400 (+0.267)

linked perturbations, the common-limit rule loses much of this directional sensitivity (abrupt 0.344 versus 0.744 for equal-share), and equal-share remains close to the better single-channel chart in each case.

For the equal-share chart, exact McNemar tests comparing each injection with its unmodified counterpart give $p = 1.38 \times 10^{-13}$ for the abrupt shift, $p = 1.86 \times 10^{-7}$ for the predictor-linked perturbation, and $p = 1.19 \times 10^{-7}$ for scale inflation. Normal-approximation paired 95% intervals for the increases are [0.492, 0.730], [0.238, 0.473], and [0.175, 0.359], respectively. Thus the Metro study supplies quantitative evidence from a different regression domain without assuming that the unmodified deployment stream is stationary.

5. Discussion

The experiments clarify the contribution of FCF-DS-EWMA. Forward cross-fitting addresses the optimism of an in-sample residual reference, while the signed and magnitude channels are complementary by design. The equal-share rule is not intended to outperform the better specialist for every mechanism; instead, it reduces the penalty associated with not knowing in advance which residual characteristic will be informative. This behavior is supported by the common-limit and Bonferroni ablations.

Calibration remains the main statistical limitation. The unified OFAT study shows that serial dependence and how well it is preserved by the bootstrap are the clearest sources of PFA inflation: stronger ϕ, $L_B = 1$, and longer H all increase PFA materially, whereas reference length has no monotone effect and increasing N_{boot} beyond 250–500 contributes little. Phase-I estimation still matters, as demonstrated by the double-bootstrap diagnostic, but the present evidence does not support attributing the inflation primarily to reference size. Dependence-aware residual modeling or prewhitening is a natural extension, although it may trade sustained-shift sensitivity against better calibration.

The reference/deployment mismatch is modest in the controlled simulation but should not be ignored. Its average effect on robust scale and channel limits is only about 1–2%. The forward-CF reference is slightly more dispersed than the oracle reference, so its native limits are marginally higher; this produces a lower native PFA and correspondingly lower native P_D. The observed 0–3.6 percentage-point differences therefore reflect a small operating-point shift, not evidence that the oracle reference is intrinsically more sensitive. In the ethanol data, the reference-location difference is much more visible: the early B4 crossing occurs while prediction error is small. That crossing is treated as a reference-mismatch warning, not as a validated deterioration event. This example emphasizes that the Phase-I segment used for forward cross-fitting must itself be substantively stable.

Ethanol is an unlabeled diagnostic example that illustrates channel attribution and reference sensitivity. Metro uses actual post-Phase-I residual windows from a different application domain and then applies paired controlled perturbations. Its unmodified alarm rate is already substantial, which is itself evidence that deployment residual behavior differs from the Phase-I reference; the paired injections quantify additional sensitivity without assuming that the unmodified windows are in control.

An alarm from FCF-DS should therefore be interpreted as evidence that the prediction-error process differs from its reference behavior. It is not proof of a specific form of concept drift. The six simulated mechanisms cover location, centered interaction, nonlinear-coefficient, unconditional noise-scale, and mixed location–scale changes, but they do not exhaust conditional heteroskedasticity or alternative interaction structures. In addition, the λ/c sensitivity study is limited to the equal-share chart and does not retune FCF-R and FCF-M over the same grid; it should therefore not be interpreted as a tuning-robust comparison of specialist deficits. Post-alarm diagnosis, delayed labels, stronger conditional calibration guarantees, and model-updating policies are separate problems.

6. Conclusion

In this study, we have presented a forward-cross-fitted two-channel EWMA chart for fixed-model regression monitoring. The equal-share construction uses channel-specific finite-horizon limits and a jointly selected channel-wise exceedance level. A unified Monte Carlo pipeline with independent calibration and verification, matched ΔMSE, six deterioration mechanisms, and explicit combination-rule ablations confirms the complementary behavior of signed and magnitude monitoring and the directional advantage of equal-share calibration over a common-limit maximum. False-alarm calibration remains approximate and is most sensitive to serial dependence, block choice, and monitoring horizon. The oracle diagnostic shows that temporary-model/final-model mismatch produces only a small shift in operating point in the controlled simulation. The ethanol and Metro studies illustrate, respectively, reference sensitivity and paired stress testing on actual post-deployment residual sequences. An FCF-DS alarm indicates that the prediction-error

process differs from its Phase-I reference behavior; it does not by itself identify a specific mechanism of concept drift.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions, which helped improve the quality and clarity of the manuscript.

Reference

[1] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, Learning under concept drift: A review, IEEE Trans. Knowl. Data Eng. vol. 31, iss. 12, pp. 2346–2363, Oct. 2019. https://doi.org/10.1109/TKDE.2018.2876857

[2] J. M. Lucas and M. S. Saccucci, Exponentially weighted moving average control schemes: Properties and enhancements, Technometrics, vol. 32, iss. 1, pp. 1–12, 1990. https://doi.org/10.1080/00401706.1990.10484583

[3] G. Chen, S. W. Cheng, and H. Xie, Monitoring process mean and variability with one EWMA chart, Journal of Quality Technology, vol. 33, iss. 2, pp. 223–233, 2001. https://doi.org/10.1080/00224065.2001.11980069

[4] A. Ostadsharif Memar and S. T. A. Niaki, The Max EWMAMS control chart for joint monitoring of process mean and variance with individual observations, Qual. Reliab. Eng. Int. vol. 27, iss. 4, pp. 499–514, 2011. https://doi.org/10.1002/qre.1146

[5] W. H. Woodall, D. J. Spitzner, D. C. Montgomery, and S. Gupta, Using control charts to monitor process and product quality profiles, Journal of Quality Technology, vol. 36, iss. 3, pp. 309–320, 2004. https://doi.org/10.1080/00224065.2004.11980276

[6] W. A. Jensen, L. A. Jones-Farmer, C. W. Champ, and W. H. Woodall, Effects of parameter estimation on control chart properties: A literature review, Journal of Quality Technology, vol. 38, iss. 4, pp. 349–364, 2006. https://doi.org/10.1080/00224065.2006.11918623

[7] A. Gandy and J. T. Kvaløy, Guaranteed conditional performance of control charts via bootstrap methods, Scand. J. Stat. vol. 40, iss. 4, pp. 647–668, Dec. 2013. https://doi.org/10.1002/sjos.12006

[8] G. J. Ross, N. M. Adams, D. K. Tasoulis, and D. J. Hand, Exponentially weighted moving average charts for detecting concept drift, Pattern Recognit. Lett. vol. 33, iss. 2, pp. 191–198, Jan. 2012. https://doi.org/10.1016/j.patrec.2011.08.019

[9] A. Bifet and R. Gavaldà, Learning from time-changing data with adaptive windowing, in Proc. 7th SIAM Int. Conf. Data Mining, 2007, pp. 443–448.

[10] K. Zhang, A. T. Bui, and D. W. Apley, Concept drift monitoring and diagnostics of supervised learning models via score vectors, Technometrics, vol. 65, iss. 2, pp. 137–149, 2023. https://doi.org/10.1080/00401706.2022.2124310

[11] J. Wu and D. W. Apley, Bootstrapped control limits for score-based concept drift control charts, Technometrics, 2026. https://doi.org/10.1080/00401706.2026.2676855, Published online.

[12] J. Pyeon, D. Cacciarelli, and K. Paynabar, An adaptive sampling framework for detecting localized concept drift under label scarcity, 2025. [Online] Available: https://doi.org/10.48550/arXiv.2511.02452.

[13] B. Wang, S. Sun, J. Wang, and Y. Tang, Robust outlier detection and low-latency concept drift detection for data stream regression: A dual-channel architecture, Concurrency Computat.: Pract. Exper. vol. 38, iss. 9, Apr. 2026, Art. no. e70730. https://doi.org/10.1002/cpe.70730

[14] C. Qian, G. Wang, Z. Wang, and C. Zou, Changepoint detection in complex models: Cross-fitting is needed, 2024. [Online] Available: https://doi.org/10.48550/arXiv.2411.07874.

[15] L. J. Tashman, Out-of-sample tests of forecasting accuracy: An analysis and review, International Journal of Forecasting, vol. 16, iss. 4, pp. 437–450, Dec. 2000. https://doi.org/10.1016/S0169-2070(00)00065-0

[16] P. J. Rousseeuw and C. Croux, Alternatives to the median absolute deviation, J. Amer. Statist. Assoc. vol. 88, iss. 424, pp. 1273–1283, 1993. https://doi.org/10.1080/01621459.1993.10476408

[17] D. V. Hinkley, Inference about the change-point from cumulative sum tests, Biometrika, vol. 58, iss. 3, pp. 509–523, Dec. 1971. https://doi.org/10.1093/biomet/58.3.509

[18] A. Vergara, Gas Sensor Array Drift at Different Concentrations, UCI Machine Learning Repository, 2012. [Online] Available: https://doi.org/10.24432/C5MK6M.

[19] A. Vergara, S. Vembu, T. Ayhan, M. A. Ryan, M. L. Homer, and R. Huerta, Chemical gas sensor drift compensation using classifier ensembles, Sensors and Actuators B: Chemical, vol. 166–167, pp. 320–329, May 2012. https://doi.org/10.1016/j.snb.2012.01.074

[20] I. Rodríguez-Luján, J. Fonollosa, A. Vergara, M. Homer, and R. Huerta, On the calibration of sensor arrays for pattern recognition using the minimal number of experiments, Chemom. Intell. Lab. Syst. vol. 130, pp. 123–134, Jan. 2014. https://doi.org/10.1016/j.chemolab.2013.10.012

[21] T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[22] J. Hogue, Metro Interstate Traffic Volume, UCI Machine Learning Repository, 2019. [Online] Available: https://doi.org/10.24432/C5X60B.