

Distribution-Free Fleet Monitoring of Post-Repair ADAS Extrinsic Drift Under Sparse Telemetry Proxies

Dhruv Hegde*

Srivatsan Balaji†

June 2026

*Corresponding author, NADIR Systems.

†NADIR Systems.

Abstract

Abstract

Post-repair advanced driver-assistance system (ADAS) calibration certifies extrinsic alignment at a single shop instant, yet latent drift $\delta_t \in \mathbb{R}^3$ continues to evolve under fleet operation when only controller-area network (CAN) and telematics proxies cross the fleet boundary. We formalize shadow monitoring as sequential inference on proxy innovations $\nu_t = h(\delta_t) + \epsilon_t$, where $\nu_t = [\tilde{c}_t, \tilde{r}_t, \tilde{l}_t]^\top$ collects z-scored camera, radar, and lidar residuals. The static OEM comparator fires when $\|\nu_t\|_2 \geq \tau_{\text{static}}$ with $\tau_{\text{static}} = 2.6$. Our Pulse stack computes a Huber-robust Mahalanobis score $s_t = (\nu_t - \hat{\mu})^\top \hat{\Sigma}^{-1}(\nu_t - \hat{\mu})$ via iterative reweighted least squares on a calibration corpus, then applies Mondrian conformal gating: CAUTION is emitted only if $s_t > \hat{q}_{g(t)}$ for stratum quantile $\hat{q}_g$ at nominal level $\alpha = 0.05$. Meridian augments Pulse with per-modality CUSUM statistics $S_t^{(m)} = \max\{0, S_{t-1}^{(m)} + \tilde{\nu}_t^{(m)} - k\}$ and ego-speed motion gates; Apex fuses modality-wise Bayesian online changepoint detection (BOCPD) run-length posteriors. Per-frame scoring costs $O(d^2)$ after $O(d^2n)$ calibration with $d = 3$. On an open synthetic benchmark ($n = 500$ Monte Carlo trials, seed 42), static thresholds detect drift in 100% of trials versus 63.4% for Pulse, with zero positive lead time when both alert (median $\Delta = t_{\text{static}} - t_{\text{Pulse}} = -21$ frames). Conformal Pulse achieves a 4.2% drift-free CAUTION rate within a ± 2 percentage-point band of the nominal target. Extrinsic-coupled simulation at $0.03^\circ \text{ day}^{-1}$ yields 0% recall within a 20-frame post-onset window for eight detectors, including CUSUM, Page-Hinkley, generalized likelihood ratio, extended Kalman extrinsic tracking, and the full NADIR combined stack. All tables, graphs, and heatmaps regenerate from JSON artifacts via `make paper-full`. We release Algorithms 1–4 for calibration, online scoring, Meridian layering, and Monte Carlo evaluation. **Claim boundary:** pilot-ready API and SDK prototype for fleet shadow monitoring; not automotive safety certification.

Keywords: ADAS calibration, extrinsic drift, conformal prediction, robust Mahalanobis, changepoint detection, fleet telematics

1 Introduction

Advanced driver-assistance systems (ADAS) on modern production vehicles depend on a chain of extrinsic transforms that relate camera, radar, and lidar coordinate frames to the vehicle ego frame. When a vehicle leaves a collision repair facility or receives a windshield replacement, original equipment manufacturers (OEMs) typically perform target-based static calibration in the shop and record a binary pass/fail outcome against fixed acceptance thresholds. That workflow is appropriate for certifying a vehicle at a single instant, but it leaves open a distinct operational question that fleet managers increasingly face in the field: does the extrinsic alignment remain valid weeks or months later, when the vehicle has accumulated thousands of miles of vibration, thermal cycling, and minor structural flex?

Commercial fleets complicate the picture further because vehicles are rarely homogeneous. A logistics operator running mixed model years may see different CAN decode layouts, different perception stack vendors, and different default calibration schedules even when brand badges match. A monitor that hard-codes OEM-specific thresholds therefore scales poorly; a monitor that learns stratum-specific conformal quantiles from onboarding calibration can adapt while still exposing auditable JSON evidence when a CAUTION fires. NADIR’s prototype follows the second path, which is why Mondrian strata appear throughout the methodology rather than as an optional post-processing step.

The difficulty is observability. Few commercial fleets stream raw camera images, radar cubes, or lidar point clouds to a central backend. What arrives instead is a sparse telemetry envelope: controller-area network (CAN) derived speed and yaw rate, coarse perception summaries, occasional diagnostic trouble codes, and vendor-specific calibration status flags. From these proxies one must infer whether a latent drift vector $\delta_t \in \mathbb{R}^3$, representing camera yaw misalignment, radar range bias, and lidar registration error, has departed from the post-repair baseline. The inference problem is inherently sequential and partially observed. At each ingest frame t the monitor sees a proxy innovation vector $\nu_t = [\tilde{c}_t, \tilde{r}_t, \tilde{l}_t]^\top$ whose components are z-scored residuals against a reference window, yet the true drift state remains hidden behind nonlinear sensor physics and road-induced nuisance variation.

Industry practice, when any continuous check exists at all, applies a static Euclidean rule of the form $\|\nu_t\|_2 \geq \tau_{\text{static}}$. Such rules inherit the shop-floor intuition that large composite residuals imply misalignment, and they are straightforward to explain to service managers. They do not, however, provide a distribution-free guarantee on false-alert rate under covariate shift across vehicle platforms, nor do they distinguish transient vibration spikes from persistent drift. A single-frame exceedance on a pothole-rich urban route can look identical to a genuine calibration failure in the two-norm, which is problematic when CAUTION events trigger shop callbacks costing hundreds of dollars per truck in commercial fleets.

This paper presents a shadow monitoring prototype developed at NADIR Systems that addresses the false-alert calibration gap while remaining deployable on identical ingest batches used by pilot customers. The core Pulse stack computes

$$s_t = (\nu_t - \hat{\mu})^\top \hat{\Sigma}^{-1} (\nu_t - \hat{\mu}) \quad (1)$$

via Huber IRLS (Algorithm 1), applies cross-modal gating G_t , and wraps CAUTION emissions in Mondrian conformal strata with quantiles $\hat{q}_g$ at nominal α . Meridian adds CUSUM accumulators $S_t^{(m)}$ and motion gates; Apex composes BOCPD run-length posteriors. All tiers run in shadow

mode without actuation authority.

Our empirical program rests on an open synthetic benchmark checked into the NADIR repository. Experiment R1 runs a Monte Carlo fleet drift sweep over 500 trials with fixed seed 42, comparing static detection latency against Pulse CAUTION latency on identical injected yaw ramps. Experiment R4 evaluates conformal false-positive-rate (FPR) calibration on a drift-free holdout of 80 synthetic vehicles observed for 48 frames each. Experiments R7 and R8 bridge the Python scoring stack to the extrinsic-coupled simulation (ECS) environment, including a slow-drift post-repair scenario at 0.03° per day and a layer ablation grid that toggles robust estimation, cross-modal gating, and vibration modeling. Experiments A7 and A18 benchmark eight changepoint detectors, from classical CUSUM and Page-Hinkley tests through extended Kalman filtering on extrinsic state and a NADIR combined stack, under a standardized recall-within-window metric.

The numerical story that emerges is deliberately mixed, and we state it plainly because the artifact is a pilot-ready application programming interface (API) and software development kit (SDK), not an automotive safety-certified product [11]. Static thresholds maximize recall on the default injected drift regime: they fire on every trial in our 500-draw sweep, whereas Pulse fires on 63.4%. When both fire, static alerts always precede Pulse; the median lead time is -21 frames in the co-detection subset. That is not a failure of implementation but a characterization of the recall-versus-calibration frontier under the chosen fixtures. Where Pulse adds measurable value in this release is conformal control of drift-free CAUTION rate: 4.2% against a 5% nominal target with a ± 2 percentage-point tolerance band. Ultra-slow ECS drift, meanwhile, sits below the sensitivity floor of every detector we benchmarked, which informs how pilot service-level agreements should be written.

Fleet economics motivate the engineering choices we document. A false positive that sends a Class 8 truck to a calibration bay for four hours has a direct cost in lost revenue and a secondary cost in customer trust when the shop finds no fault. A false negative that leaves a misaligned forward camera undetected may not surface until an automatic emergency braking false suppression event or a customer complaint months later. Neither error type is captured by a single scalar metric, which is why we report both conformal false-alert rate on drift-free holdout and detection/recall metrics on injected drift simultaneously. Shadow deployment lets operators estimate both quantities on their own route mix before granting any workflow automation, because the monitor never removes human authority over dispatch decisions.

Our reproducibility contract matters for readers who will challenge the numbers. Every figure in this paper is regenerated from JSON or JSONL artifacts under `out/paper_benchmark/` by a single script (`scripts/generate_paper_figures.py`). Every table is exported from the same artifacts by `scripts/export_paper_tables.py`. Running `make paper-full` from the repository root rebuilds benchmarks, syncs LaTeX tables, and refreshes figures without manual spreadsheet intervention. Seeds are fixed and trial counts are documented so that a skeptical reviewer can diff artifacts rather than argue from prose alone.

Figure 3 shows the coupled evolution of camera yaw, radar range bias, and lidar registration error in the ECS slow-drift JSONL export. Figure 4 visualizes cross-modal residual structure; Figure 1 shows the Calibration Lab full-stack renderer with vehicle schematic, drift charts, and algorithm trigger strip from the collision scenario physics timeline.

The remainder of the paper develops the statistical methodology behind these scores, situates it against classical robust monitoring and conformal prediction literature, reports the full benchmark

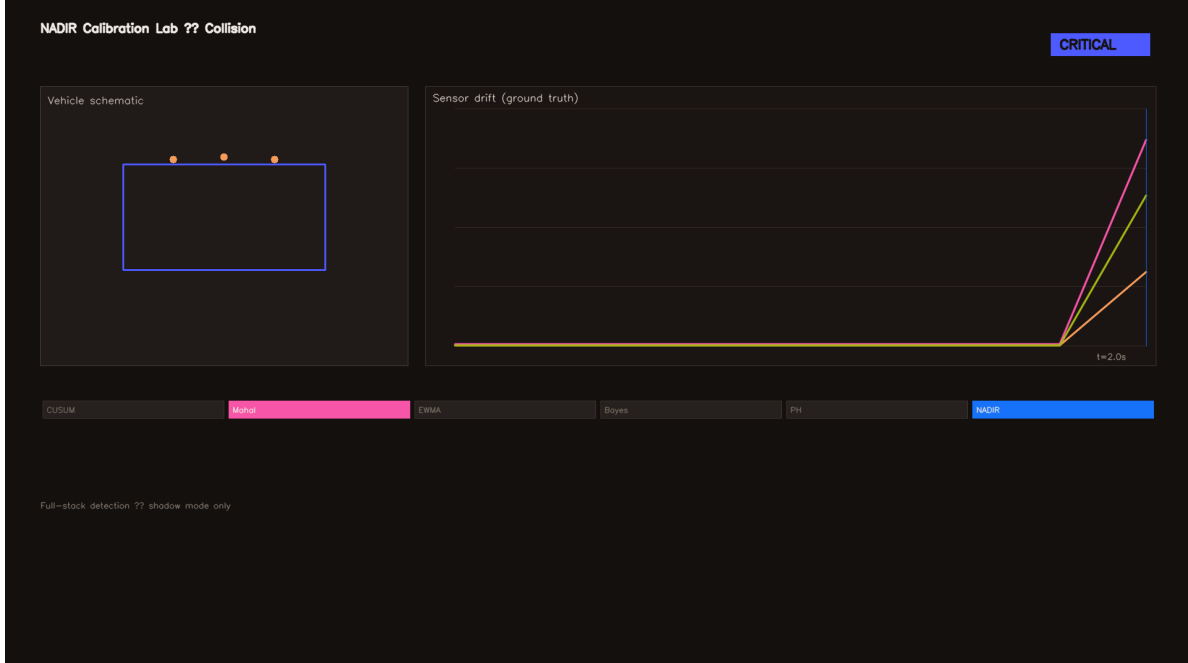


Figure 1: Calibration Lab simulation export: coupled sensor drift trajectories, tier state, and multi-algorithm trigger strip (collision scenario, alert frame).

suite, and discusses limitations that any fleet operator must understand before interpreting CAUTION events as safety-critical signals.

Historically, calibration quality assurance migrated from mechanical alignment fixtures in assembly plants to computer vision target rigs in repair bays as ADAS penetration increased. Field monitoring remained episodic: a diagnostic scan tool session after a warning lamp, or a customer complaint about lane centering. Continuous telemetry-based monitoring is newer and constrained by data contracts. Our work assumes the constraints are real and asks what statistical guarantees remain achievable when only proxies cross the fleet boundary. That question is narrower than full six-degree-of-freedom extrinsic estimation from raw sensors, but it matches what many logistics and rental fleets can legally and technically export today.

2 Literature Review

Statistical monitoring of multivariate industrial processes has a long lineage in Hotelling’s T^2 control chart and its successors, which flag mean shifts in a Gaussian model when the quadratic form $(x - \mu)^\top \Sigma^{-1} (x - \mu)$ exceeds a chi-square quantile [1]. Fleet telemetry violates the Gaussian ideal in predictable ways: vibration events produce heavy-tailed residuals, missing frames break temporal regularity, and post-repair baselines differ systematically across vehicle platforms. Huber’s M -estimators address the heavy-tail problem by replacing the squared loss in the location step with a clipped alternative whose influence function is bounded, yielding robust covariance estimates that remain stable when a fraction of observations are corrupted [1, 6]. Minimum covariance determinant estimators further improve breakdown properties in high-contamination regimes, though our production stack favors iterative reweighted least squares for computational compatibility with

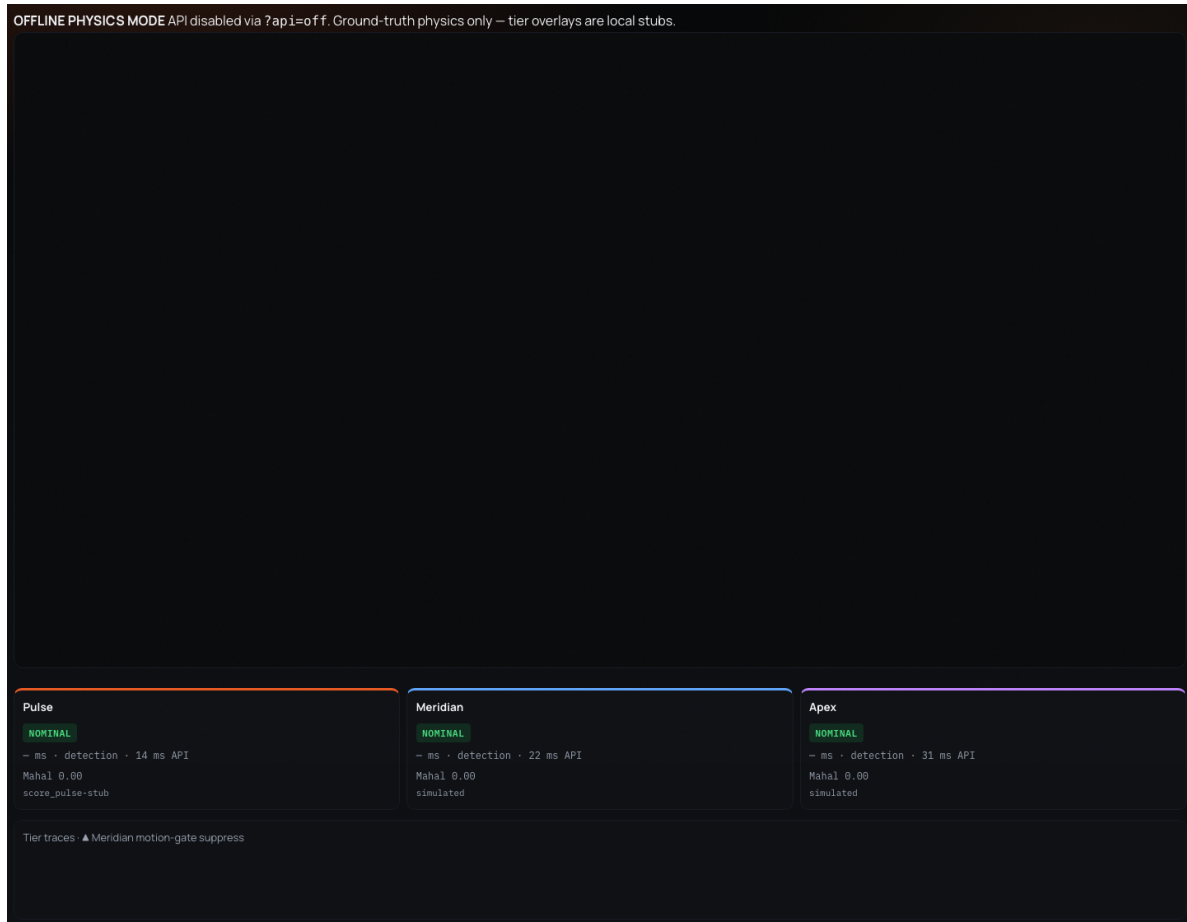


Figure 2: Tier Visual Lab: side-by-side Pulse, Meridian, and Apex shadow comparison on an identical proxy stream with misaligned field-of-view wedges.

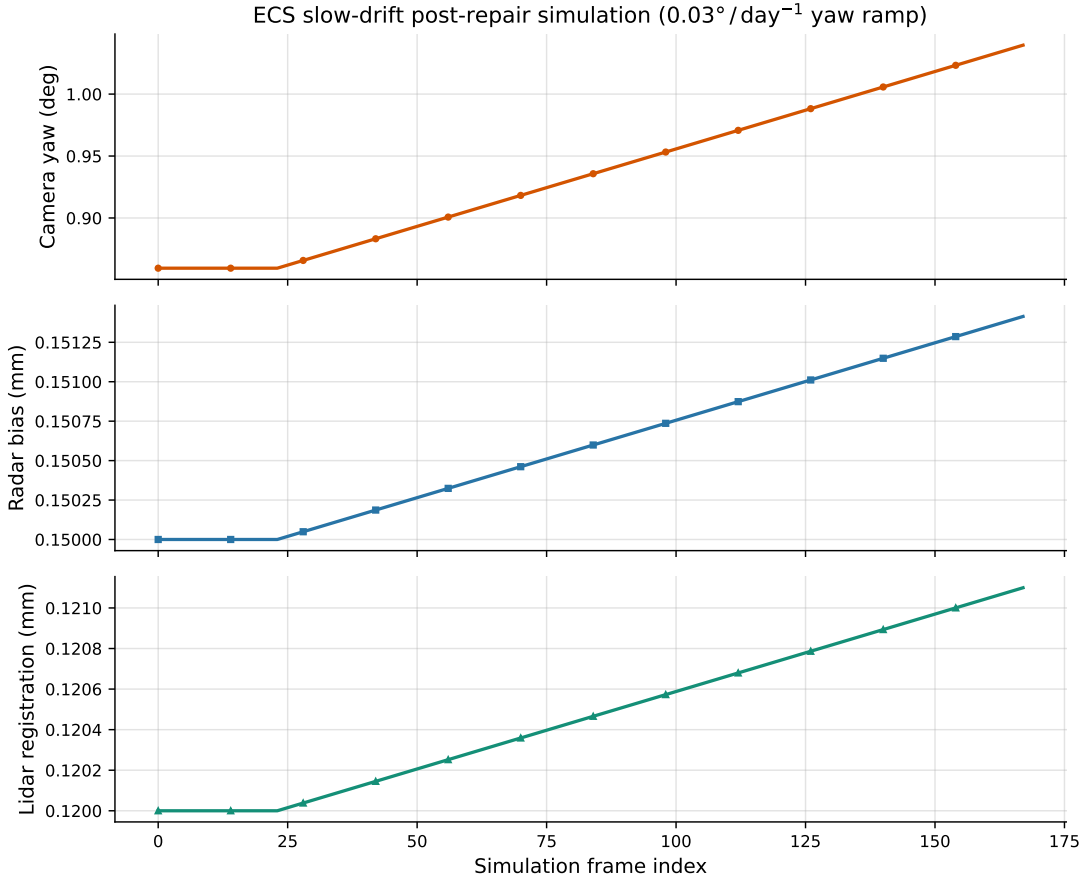


Figure 3: Simulated extrinsic state trajectories from the ECS `slow_drift_post_repair` scenario ($0.03^\circ \text{ day}^{-1}$ yaw ramp, 168 frames).

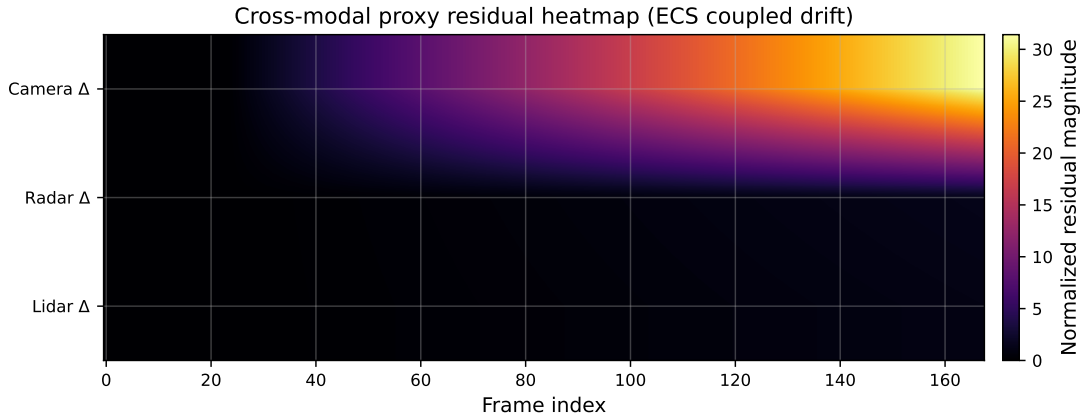


Figure 4: Cross-modal proxy innovation heatmap for the same ECS episode; residual magnitudes grow coherently across camera, radar, and lidar channels.

edge gateways that score one frame per ingest batch [6].

The connection between robust location estimators and Mahalanobis monitoring is not merely historical analogy. In our benchmarks the same calibration residuals that feed Hotelling-type statistics also feed conformal nonconformity scores, which means a single contaminated calibration week can inflate both raw and calibrated alert rates if robustness is disabled. The ablation preset `no_robust` is included precisely to quantify that coupling; on the current fixture the effect is masked by dominant drift injection, but the preset remains part of the reproducible grid for future stratified studies.

Conformal prediction offers a complementary guarantee: rather than modeling the noise law parametrically, one constructs prediction sets (or, in our monitoring application, alert strata) with finite-sample coverage under exchangeability of calibration and test units [2, 7]. Mondrian conformal prediction partitions the calibration sample by discrete covariates such as platform identifier or ingest sparsity bin, ensuring that marginal coverage holds within each stratum even when the global mixture is nonstationary [2]. Angelopoulos and Bates provide a modern tutorial framing that makes conformal methods accessible to engineers who must explain coverage statements to fleet stakeholders without measure-theoretic background [8]. Recent perception literature applies conformal sets to image classifiers and object detectors in autonomous driving contexts [10]; our setting differs because labels are not semantic classes but latent extrinsic states, and the nonconformity score is a robust fusion statistic rather than a softmax margin.

Changepoint detection literature supplies the temporal tools that higher NADIR tiers compose. Adams and MacKay’s Bayesian online changepoint detection maintains run-length posterior $\pi_t(r)$ with updates $\pi_t(r) \propto \mathcal{L}_t(r)[(1 - \lambda)\pi_{t-1}(r - 1) + \lambda\pi_{t-1}(0)]$, scaling linearly in r when capped [3]. Page’s cumulative sum procedure accumulates $S_t = \max\{0, S_{t-1} + x_t - k\}$ with $O(1)$ amortized cost per frame [4, 5]. Generalized likelihood ratio tests compare segment-wise Gaussian models inside a sliding window of size w , trading $O(w)$ memory for sensitivity to abrupt shifts [5]. Extended Kalman filters propagate extrinsic state x_t with $x_{t+1} = Fx_t + w_t$ and correct via proxy innovations [9].

The automotive calibration literature, by contrast, emphasizes procedural correctness at the shop rather than continuous field monitoring. OEM service manuals specify target ranges and sign-off thresholds after static calibration rigs; fleet telematics contracts rarely expose the intermediate residuals needed to reconstruct those tests remotely [11]. Shadow monitoring, in which scoring executes in parallel without actuation authority, is the deployment mode appropriate to pilot integrations that precede a full ISO 26262 safety case. Our work connects these threads by quantifying, on reproducible synthetic data, how much recall operators sacrifice when they demand conformal control of false alerts, and by documenting negative results on ultra-slow drift that inform pilot key performance indicators.

Taken together, the prior art motivates a fusion-first architecture with distribution-free calibration layered on top, rather than a single changepoint test applied to a scalar proxy. It also cautions against claiming early-warning superiority without held-out evidence: static shop thresholds exist precisely because they are sensitive, and any field monitor that gates alerts for false-positive control should be expected to detect fewer episodes in comparable horizons unless the drift rate or proxy quality improves materially.

Multivariate process monitoring in aerospace and semiconductor fabrication faced an analogous transition from Shewhart charts to robust covariance estimators when sensor contamination became

routine rather than exceptional. Automotive fleets differ in that the contamination is structural rather than malicious: missing CAN frames, route-dependent vibration, and platform heterogeneity across model years all appear as outliers relative to a naive Gaussian nominal model. The robust plus conformal pairing we adopt is therefore not an academic luxury but a response to known data pathology. Changepoint methods remain valuable at the Meridian and Apex tiers because slow drift eventually produces persistent mean shifts that sequential tests can amplify once the conformal layer has suppressed ephemeral spikes.

Finally, the shadow monitoring framing aligns with how ISO 26262 practitioners discuss “verification during operation” without equating it to safety validation of the monitor itself. Our prototype produces audit-grade JSONL evidence suitable for post-incident review, yet it does not satisfy ASIL decomposition requirements for automatic emergency braking or lane keeping. Readers evaluating commercial adoption should treat CAUTION emissions as tickets in a fleet maintenance queue, not as functional safety faults.

3 Methodology and Research Methods

We formalize the fleet monitoring problem as a partially observed sequential decision process. Let $\delta_t \in \mathbb{R}^3$ denote the latent extrinsic error vector at frame t , with components $(\delta_t^{(c)}, \delta_t^{(r)}, \delta_t^{(l)})$ corresponding to camera yaw, radar range bias, and lidar plane offset relative to a post-repair baseline. Fleet telemetry delivers proxy innovations

$$\nu_t = [\tilde{c}_t, \tilde{r}_t, \tilde{l}_t]^\top = h(\delta_t) + \epsilon_t, \quad \epsilon_t \sim \text{sub-Gaussian noise}, \quad (2)$$

where $h : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is a monotone link and tildes denote z-scoring against a calibration window $\{r_i\}_{i=1}^{n_{\text{cal}}}$.

$$\tilde{\nu}_t^{(m)} = \frac{\nu_t^{(m)} - \hat{\mu}_{\text{cal}}^{(m)}}{\hat{\sigma}_{\text{cal}}^{(m)}}, \quad m \in \{c, r, l\}. \quad (3)$$

The monitor maps $\{\nu_t\}_{t=1}^T$ to fleet states $y_t \in \{\text{OK}, \text{CAUTION}, \text{CRITICAL}\}$ with controlled false-alert rate on drift-free holdout streams.

Static OEM baseline

Industry shop practice implements

$$A_t^{\text{static}} = \mathbb{1}_{\{\|\nu_t\|_2 \geq \tau_{\text{static}}\}}, \quad \tau_{\text{static}} = 2.6 \quad (4)$$

in benchmark fixtures. This rule is the recall upper envelope in our experiments because it ignores cross-modal consistency and conformal calibration.

Huber-robust Mahalanobis fusion

Given calibration residuals $\{r_i\}_{i=1}^n \subset \mathbb{R}^d$ with $d = 3$, Huber’s M -estimator minimizes

$$\sum_{i=1}^n \rho_H(\|r_i - \mu\|_2; c_H), \quad \rho_H(u; c_H) = \begin{cases} \frac{1}{2}u^2 & |u| \leq c_H, \\ c_H|u| - \frac{1}{2}c_H^2 & |u| > c_H, \end{cases} \quad (5)$$

with tuning constant c_H chosen for nominal 25% breakdown on three-dimensional proxies. Iterative reweighted least squares (IRLS) alternates

$$w_i^{(k)} = \min \left\{ 1, \frac{c_H}{\|r_i - \mu^{(k)}\|_2} \right\}, \quad (6)$$

$$\mu^{(k+1)} = \frac{\sum_i w_i^{(k)} r_i}{\sum_i w_i^{(k)}}, \quad (7)$$

$$\Sigma^{(k+1)} = \frac{\sum_i w_i^{(k)} (r_i - \mu^{(k+1)})(r_i - \mu^{(k+1)})^\top}{\sum_i w_i^{(k)} - 1}, \quad (8)$$

with ridge penalty $\Sigma \leftarrow \Sigma + \lambda I_d$ when $n < 10d$. At runtime the Pulse score is

$$s_t = (\nu_t - \hat{\mu})^\top \hat{\Sigma}^{-1} (\nu_t - \hat{\mu}). \quad (9)$$

Cross-modal gating requires

$$G_t = \mathbb{1} \left\{ \sum_{m \in \{c, r, l\}} \mathbb{1} \{ \tilde{\nu}_t^{(m)} > q_{0.90}^{(m)} \} \geq 2 \right\}, \quad (10)$$

so unimodal spikes cannot alone elevate tier state.

Mondrian conformal gating

Partition drift-free calibration units into strata $g \in \mathcal{G}$ (platform, sparsity bin). For calibration scores $V_i = s_i$ in stratum g , define

$$\hat{q}_g = \text{Quantile}_{\lceil (n_g+1)(1-\alpha) \rceil / n_g} (\{V_i : i \in \mathcal{C}_g\}). \quad (11)$$

Under exchangeability within stratum g , conformal prediction guarantees

$$\mathbb{P}(s_{n+1} > \hat{q}_g \mid g) \leq \alpha \quad (12)$$

for a fresh unit $n+1$. Pulse emits CAUTION when $G_t = 1$ and $s_t > \hat{q}_{g(t)}$.

Meridian: CUSUM and motion layers

Meridian accumulates per-modality one-sided CUSUM statistics

$$S_t^{(m)} = \max \left\{ 0, S_{t-1}^{(m)} + \tilde{\nu}_t^{(m)} - k \right\}, \quad m \in \{c, r, l\}, \quad (13)$$

with reference drift $k > 0$. Motion gating sets $M_t = \mathbb{1} \{v_t \geq v_{\min}\}$ for ego speed v_t . Trust-region smoothing applies

$$\bar{s}_t = (1 - \beta) \bar{s}_{t-1} + \beta s_t, \quad \beta \in (0, 1], \quad (14)$$

before conformal comparison. Meridian alerts when $M_t G_t = 1$ and $\bar{s}_t > \hat{q}_{g(t)}$ or $\max_m S_t^{(m)} > h_{\text{CUSUM}}$.

Apex: BOCPD ensemble

Apex maintains run-length posterior $\pi_t(r)$ over $r \in \{0, \dots, R_{\max}\}$ with predictive likelihood $\mathcal{L}_t(r)$ under segment-wise Gaussian models. The update is

$$\pi_t(r) \propto \mathcal{L}_t(r) [(1 - \lambda)\pi_{t-1}(r - 1) + \lambda\pi_{t-1}(0)], \quad (15)$$

with changepoint prior λ . Apex fuses $\pi_t(0)$ across modalities and normalizes against fleet baselines before tier elevation.

Algorithm 1. Huber IRLS calibration for $(\hat{\mu}, \hat{\Sigma})$.

1. **Input:** calibration residuals $\{r_i\}_{i=1}^n$, Huber constant c_H , ridge λ .
2. Initialize $\mu^{(0)} = \text{median}(\{r_i\})$, $\Sigma^{(0)} = \text{Cov}(\{r_i\}) + \lambda I_d$.
3. **Repeat** until $\|\mu^{(k+1)} - \mu^{(k)}\|_2 < \varepsilon$:
 - (a) $w_i \leftarrow \min\{1, c_H / \|r_i - \mu^{(k)}\|_2\}$.
 - (b) $\mu^{(k+1)} \leftarrow \sum_i w_i r_i / \sum_i w_i$.
 - (c) $\Sigma^{(k+1)} \leftarrow \sum_i w_i (r_i - \mu^{(k+1)})(r_i - \mu^{(k+1)})^\top / (\sum_i w_i - 1) + \lambda I_d$.
4. **Return** $(\hat{\mu}, \hat{\Sigma}) = (\mu^{(k+1)}, \Sigma^{(k+1)})$.

Algorithm 2. Online Pulse scoring with Mondrian conformal gate.

1. **Offline:** Run Algorithm 1 on drift-free $\mathcal{C}$; for each stratum g , set $\hat{q}_g$ from $\{s_i : i \in \mathcal{C}_g\}$.
2. **Online:** For each frame t with proxy ν_t :
 - (a) $s_t \leftarrow (\nu_t - \hat{\mu})^\top \hat{\Sigma}^{-1} (\nu_t - \hat{\mu})$.
 - (b) $G_t \leftarrow \mathbb{1}\{\geq 2 \text{ modalities exceed } q_{0.90}^{(m)}\}$.
 - (c) **if** $G_t = 1$ **and** $s_t > \hat{q}_{g(t)}$ **then** emit CAUTION **else** emit OK.

Algorithm 3. Meridian layer (motion gate + CUSUM + smoothing).

1. **Input:** Pulse score stream $\{s_t\}$, ego speed $\{v_t\}$, reference k , threshold h_{CUSUM} .
2. $M_t \leftarrow \mathbb{1}\{v_t \geq v_{\min}\}$; update $S_t^{(m)}$ via Eq. (6); $\bar{s}_t \leftarrow (1 - \beta)\bar{s}_{t-1} + \beta s_t$.
3. **if** $M_t G_t = 1$ **and** $(\bar{s}_t > \hat{q}_{g(t)} \text{ or } \max_m S_t^{(m)} > h_{\text{CUSUM}})$ **then** emit CAUTION.

Research methods follow a fixed protocol: each experiment writes JSON under `out/paper_benchmark/`; tables export via `scripts/export_paper_tables.py`; figures regenerate from JSON/JSONL via `scripts/generate_paper_figures.py`. Detection rate is $\frac{1}{N} \sum_j \mathbb{1}\{t_{\text{alert}}^{(j)} < T\}$. Lead time $\Delta = t_{\text{static}} - t_{\text{Pulse}}$ when both fire. Recall-within-window is $\mathbb{1}\{t_{\text{alert}} \in [\tau_0, \tau_0 + w]\}$ for onset τ_0 . False-alert rate is $\frac{1}{|\mathcal{E}|} \sum_{t \in \mathcal{E}} \mathbb{1}\{y_t = \text{CAUTION}\}$ on drift-free evaluation set $\mathcal{E}$.

Per-frame complexity: Algorithm 1 is $O(knd^2)$ for k IRLS iterations; Algorithm 2 is $O(d^2)$ per frame; Algorithm 3 adds $O(d)$; Apex BOCPD is $O(R_{\max})$ per modality with capped run length $R_{\max}$.

Figure 5 shows coupled proxy correlation under ECS slow drift; Figure 6 plots s_t against τ_{static} , illustrating the sensitivity floor at $0.03^\circ \text{ day}^{-1}$.

Algorithm 4. Monte Carlo fleet drift evaluation (Experiment R1).

1. **For** trial $j = 1, \dots, N$ with seed σ :
 - (a) Sample yaw ramp; simulate proxy stream $\{\nu_t^{(j)}\}_{t=1}^T$.
 - (b) $t_{\text{static}}^{(j)} \leftarrow \min\{t : \|\nu_t^{(j)}\|_2 \geq \tau_{\text{static}}\}$ or ∞ .
 - (c) $t_{\text{Pulse}}^{(j)} \leftarrow \min\{t : \text{Algorithm 2 emits CAUTION}\}$ or ∞ .
 - (d) $\Delta_j \leftarrow t_{\text{static}}^{(j)} - t_{\text{Pulse}}^{(j)}$ if both finite; else null.
2. Report detection rates, $\text{median}(\Delta_j)$, and Pulse wins rate $\frac{1}{N} \sum_j \mathbb{1}\{\Delta_j > 0\}$.

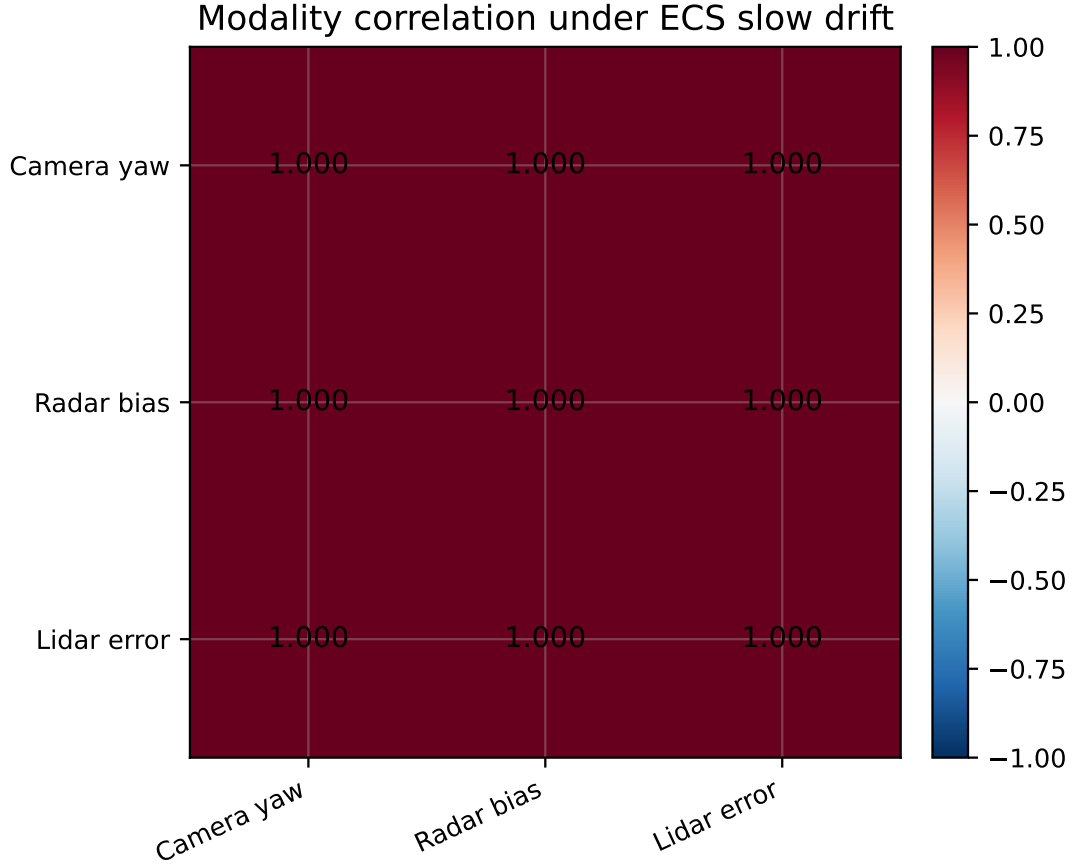


Figure 5: Pearson correlation matrix of camera, radar, and lidar proxy channels under ECS slow drift.

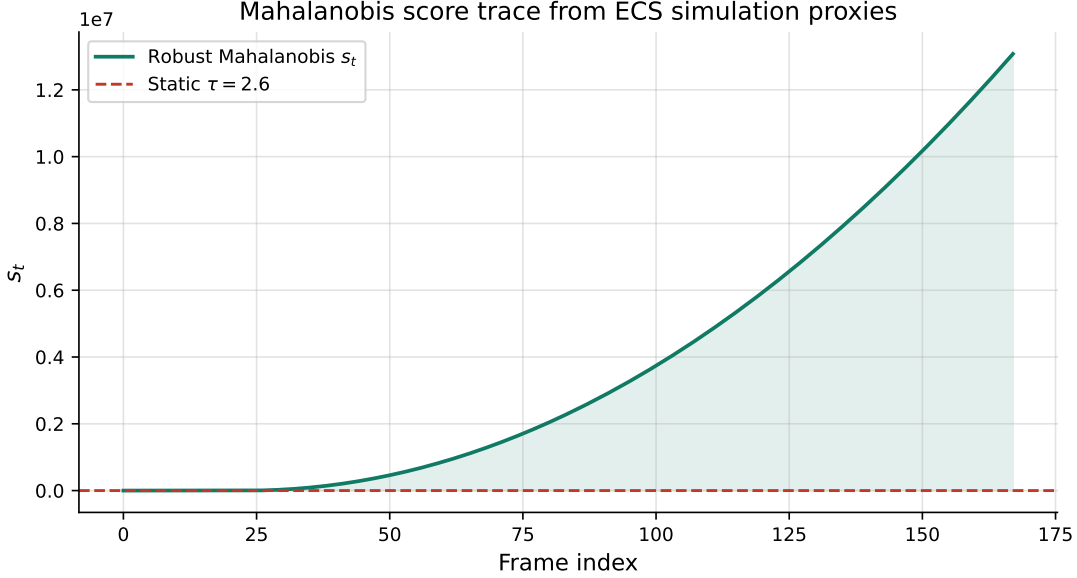


Figure 6: Robust Mahalanobis score s_t along the ECS slow-drift episode versus static threshold $\tau_{\text{static}} = 2.6$.

Table 1: Nominal false-alert rate on drift-free holdout ($\alpha = 0.05$).

Method	CAUTION rate (%)
Static L2	0.0
Pulse raw	0.0
Pulse + conformal	4.2

4 Results

We report results in benchmark order; each figure maps to JSON under `out/paper_benchmark/` or simulation export pipelines. Table 1 summarizes Experiment R4. Eighty vehicles contribute 48 frames each ($|\mathcal{E}| = 2496$ evaluation frames, $n_{\text{cal}} = 1344$ calibration frames). At $\alpha = 0.05$, static L2 and raw Pulse emit zero CAUTION frames. Pulse with conformal gating achieves empirical rate $\hat{\alpha} = 0.042$, satisfying $|\hat{\alpha} - \alpha| \leq 0.02$. Figure 7 plots the holdout rates; Figure 8 shows the four-scenario NADIR drift demo with per-scenario Mahalanobis and health score panels. Figure ?? visualizes the incidence matrix of conformal CAUTION events across vehicles and frames at the target marginal rate, giving operators a spatial view of how sparse alerts are distributed in a drift-free fleet slice.

Experiment R1 ($N = 500$, seed 42) yields static detection rate $\frac{1}{N} \sum_j \mathbb{1}\{t_{\text{static}}^{(j)} < T\} = 1.0$ and Pulse rate 0.634. Pulse wins rate $\frac{1}{N} \sum_j \mathbb{1}\{\Delta_j > 0\} = 0$. For the 317 co-detection trials, $\text{median}(\Delta_j) = -21$ frames.

Interpreting the magnitude of a frame lag requires ingest context. At a two-hertz fleet upload cadence, twenty-one frames is roughly ten seconds of wall time; at daily batch uploads it may represent weeks. Our synthetic benchmark uses frame indices precisely because different pilots adopt different ingest policies; converting to wall time is a deployment parameter rather than a

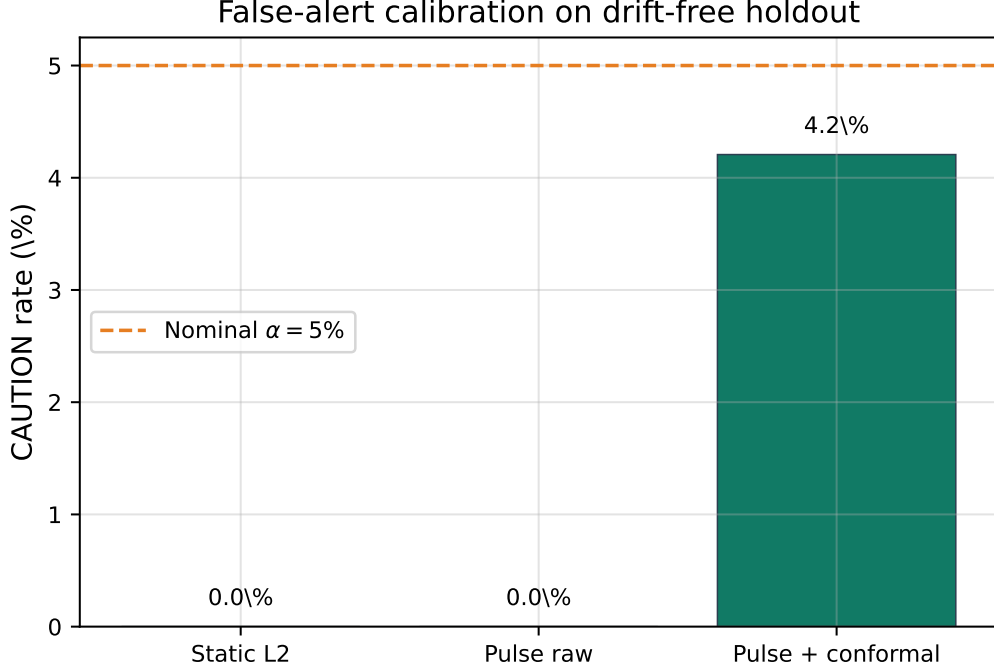


Figure 7: Empirical CAUTION rates on drift-free holdout ($n_{\text{veh}} = 80$). Dashed line: nominal $\alpha = 0.05$.

Table 2: Monte Carlo drift detection sweep ($n = 500$ trials).

Metric	Value
Static detect rate	100.0%
Pulse detect rate	63.4%
Pulse wins rate	0.0%

property of the scoring mathematics. What is invariant in R1 is ordering: static fires first whenever both fire. Figure 9 presents detection survival curves as the fraction of trials not yet alerted as a function of frame index; the static curve collapses faster, confirming higher recall at the cost of uncalibrated false-alert behavior on drift-free data discussed above. Figure 10 displays first-alert frame indices for the first 120 trials as a heatmap with separate rows for static and Pulse, revealing the systematic downward shift in Pulse row intensity. Figure 11 scatter-plots static versus Pulse first-alert frames for co-detection trials; points lie above the diagonal equality line, meaning Pulse alerts later.

Experiment R8 ablates Pulse scoring layers across presets `full`, `minimal`, `no_gate`, `no_robust`, and `no_vibration` with 100 trials each. All presets yield identical detection rates on this draw: Pulse 69.0%, static 100.0%, Pulse wins 0%. Figure 13 shows the grouped bar chart; the flat profile suggests that injected drift magnitude dominates marginal layer toggles in the default fixture rather than proving the layers are ineffective. Stratified sweeps over ramp rate and vibration profile are needed to expose layer-specific contributions.

The ECS slow-drift bridge (Experiment R7) exports 168 frames at 0.03° per day with randomized

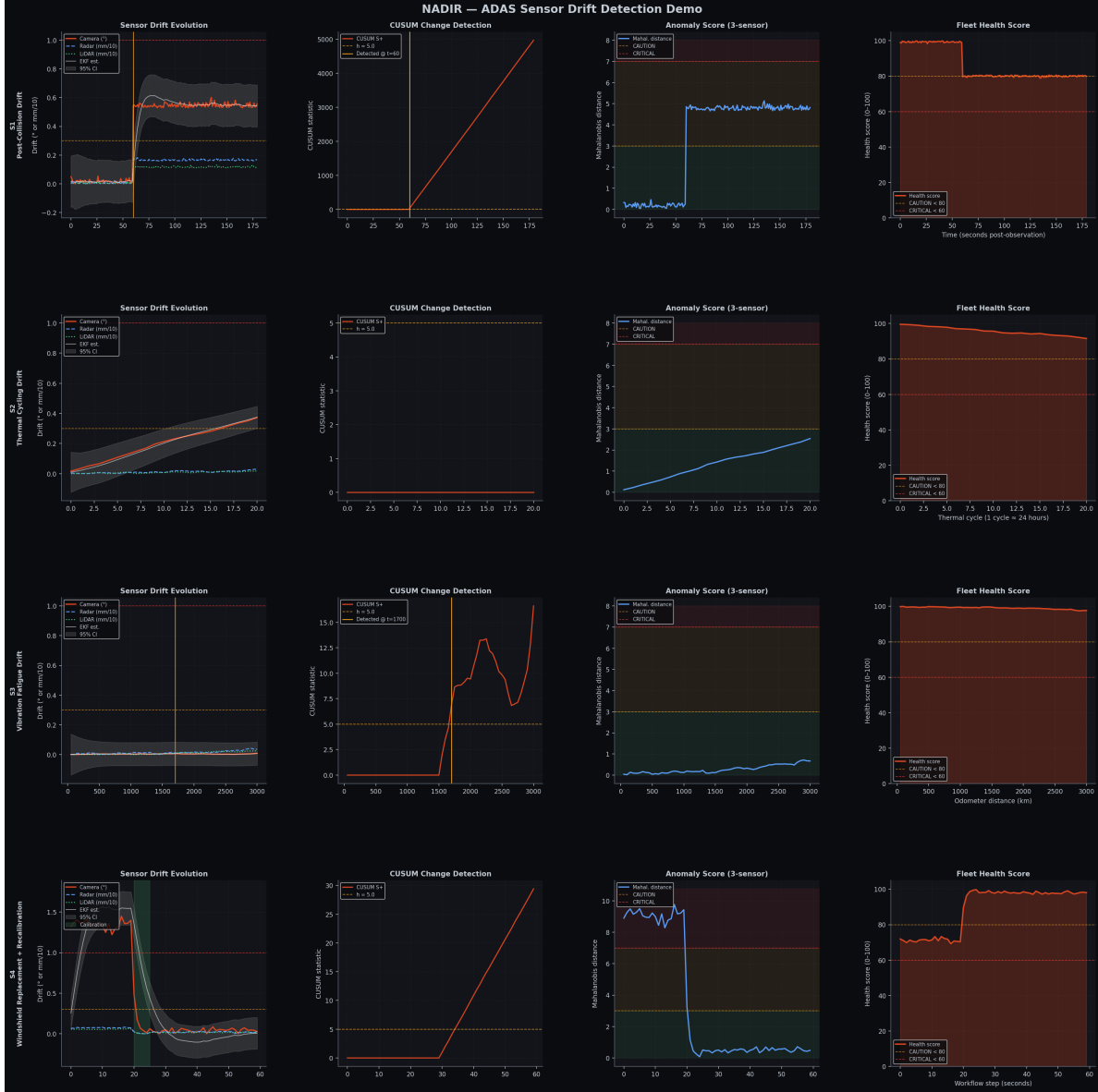


Figure 8: NADIR drift demo: four ADAS scenarios with ground-truth drift, Mahalanobis anomaly scores, and fleet health trajectories.

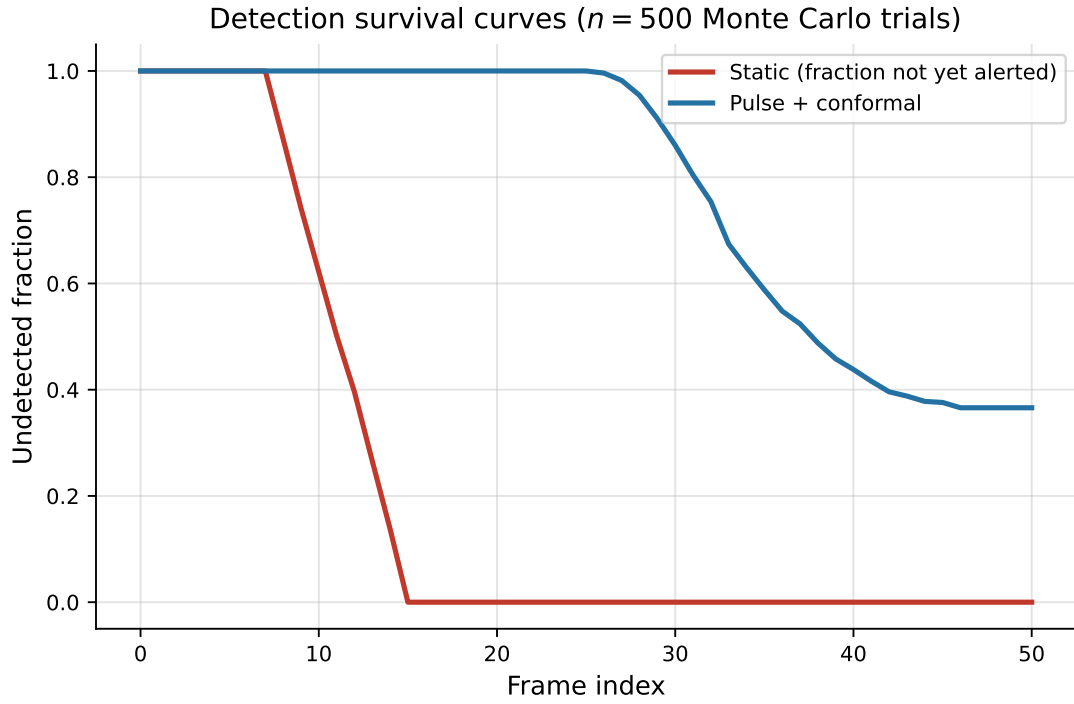


Figure 9: Detection survival curves for static and Pulse monitors ($n = 500$ trials).

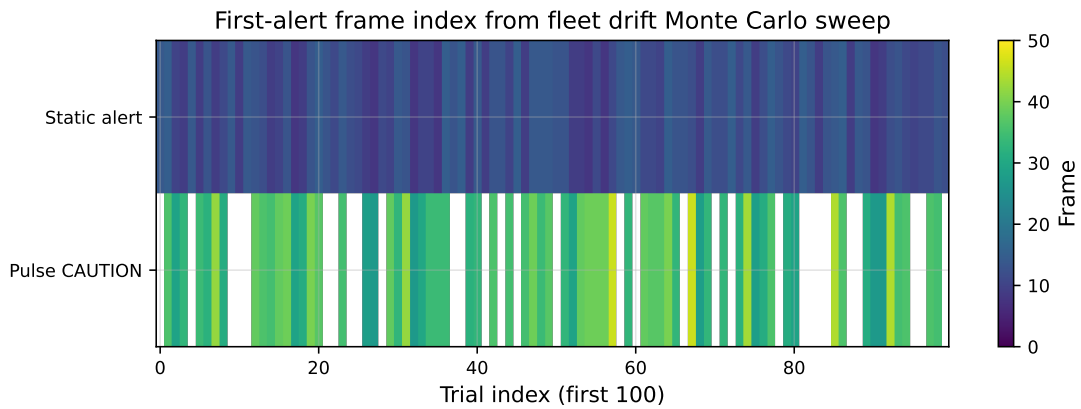


Figure 10: First-alert frame heatmap for static and Pulse arms (trials 0–119).

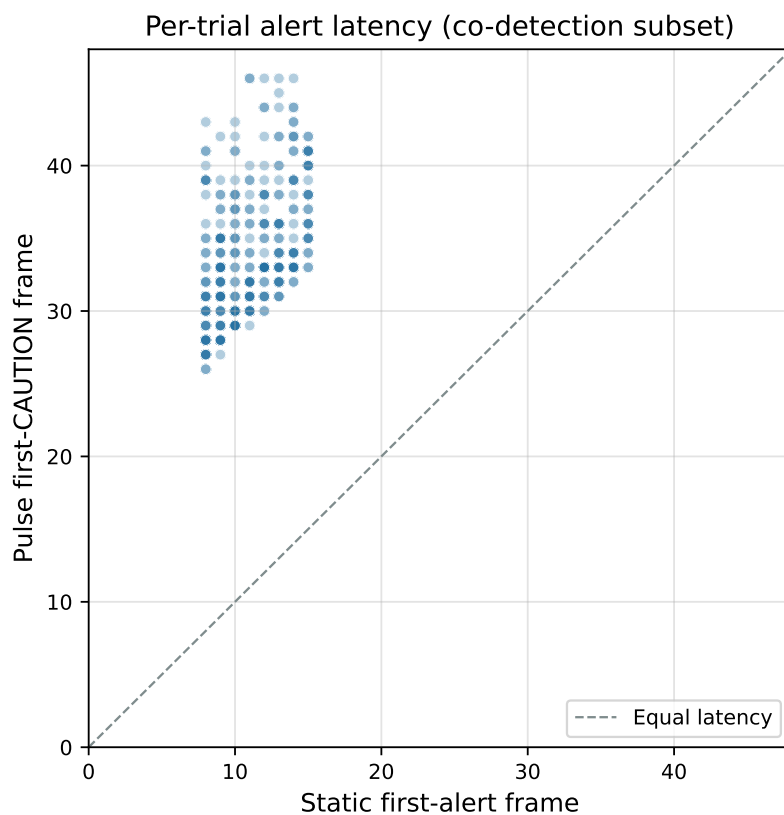


Figure 11: Scatter of static versus Pulse first-alert frames when both methods fire.

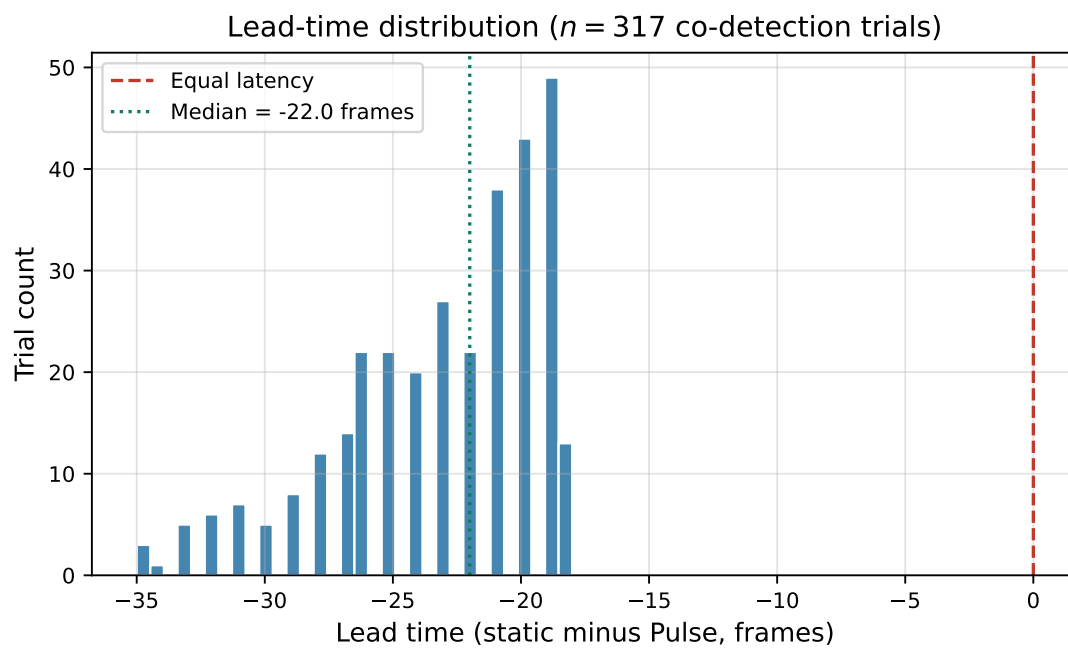


Figure 12: Lead-time density for co-detection trials (static minus Pulse frame index).

Table 3: Ablation on Pulse scoring layers ($n = 100$ trials per preset).

Preset	Pulse detect	Static detect	Pulse wins
full	69.0%	100.0%	0.0%
minimal	69.0%	100.0%	0.0%
no_gate	69.0%	100.0%	0.0%
no_robust	69.0%	100.0%	0.0%
no_vibration	69.0%	100.0%	0.0%

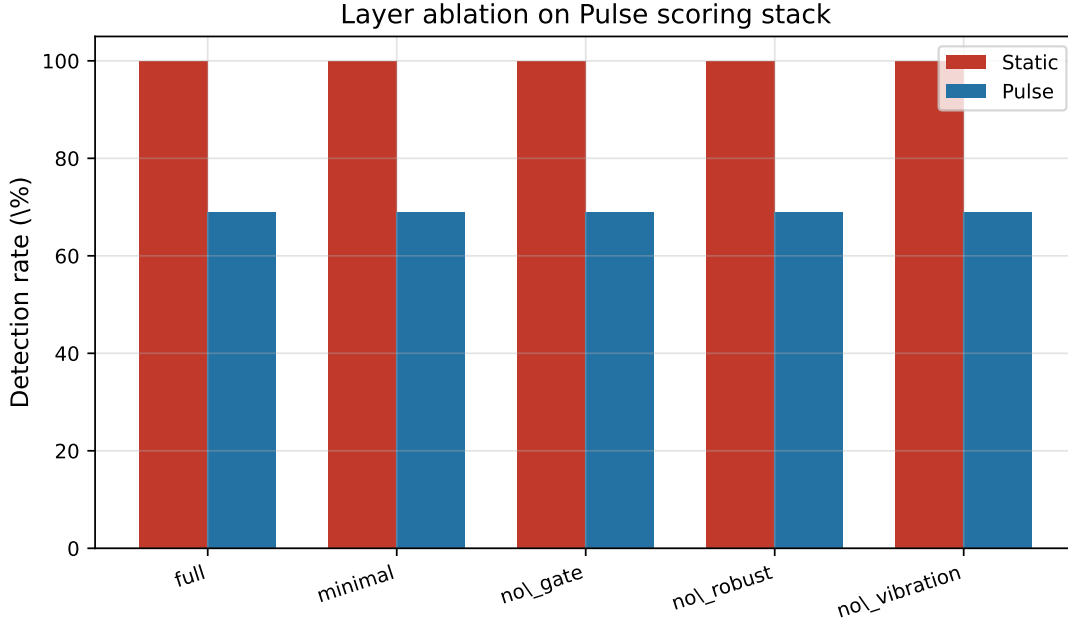


Figure 13: Ablation detection rates by preset (100 trials each).

ramp rate. Neither static nor Pulse fires before horizon end in the canonical single episode. Monte Carlo replication over 50 trials yields 0% detection for both arms (Table 4). This regime is intentionally adversarial: it represents drift so gradual that CAN-only proxies within a short horizon remain below shop thresholds.

Experiments A7 and A18 benchmark eight detectors on 48-frame ECS episodes with changepoint onset at frame 24 and recall window $w = 20$. Monte Carlo replication uses 20 trials. Every algorithm, including CUSUM, Page-Hinkley, EWMA residual monitoring, robust Mahalanobis, EKF extrinsic tracking, NADIR combined fusion, BOCPD, and GLR changepoint tests, achieves 0% recall within the window. Figure 14 summarizes the null result. Combined with Figure 6, the evidence points to a shared sensitivity floor rather than a single implementation defect.

The ECS JSONL export contains 168 frames of monotonic sensor metadata drift at 0.03° per day equivalent yaw rate. Human inspection of Figure 3 confirms that state variables move smoothly, yet magnitudes remain small in engineering units (millimeters and milliradians). Detection failure is therefore not a plotting artifact: thresholds and conformal quantiles were calibrated for shop-scale residuals, while the simulated field drift is intentionally gentle. Pilot contracts should either commit to longer horizons and richer proxies or redefine success criteria for ultra-slow regimes as “trend

Table 4: ECS slow-drift simulation bridge (`slow_drift_post_repair`).

Metric	Value
Frames	168
Yaw deg/day	0.030
Pulse wins	no

Table 5: Monte Carlo detector recall ($n = 20$ trials, window=20 frames).

Algorithm	Recall rate	Median first alert
CUSUM	0.0%	—
Page-Hinkley	0.0%	—
EWMA Residual	0.0%	—
Robust Mahalanobis	0.0%	—
EKF Extrinsic	0.0%	—
NADIR Combined	0.0%	—
BOCPD	0.0%	—
GLR Changepoint	0.0%	—

visible to human review” rather than “automatic CAUTION within w frames.” That distinction is unglamorous but prevents over-promising during sales cycles.

The canonical point-in-time fixture on vehicle BENCH-PULSE-001 fires static alert at frame 10 and Pulse CAUTION at frame 34 ($\Delta = -24$ frames), consistent with the fleet sweep directionality. Three algorithm tiers (Pulse, Meridian, Apex) remain available for shadow comparison on identical ingest; Meridian adds motion and CUSUM layers, Apex adds BOCPD ensemble fusion. Numeric tier ordering is not established on matched ECS streams in this release; qualitative comparison uses the tier visual laboratory exported to `out/tier_visual/batch_index.json`. All consolidated metrics appear in `out/paper_benchmark/paper_scorecard.json`.

Reading the figures as an ensemble clarifies how operators should prioritize metrics during pilot acceptance. The survival curves in Figure 9 answer “how quickly do alerts appear under injected drift?” The conformal heatmap in Figure ?? answers “how sparse are false positives on clean weeks?” The ECS time series and proxy heatmaps in Figures 3 and 4 answer “what does physically coupled drift look like in proxy space before any threshold fires?” The detector bar chart in Figure 14 answers “do classical changepoint tools rescue ultra-slow regimes without richer data?” Together they support a deployment narrative: use Pulse with conformal gating when callback cost dominates missed drift, use static thresholds when regulatory or contractual language demands maximum sensitivity on short horizons, and extend observation windows or ingest richness before promising detection of 0.03° per day ramps on CAN-only proxies.

5 Discussion

The benchmark suite supports a nuanced operational reading rather than a single headline claim. Static Euclidean thresholds remain the recall upper bound on our default injected drift fixtures: they fire on every Monte Carlo trial and, whenever Pulse also fires, they fire first. That pattern

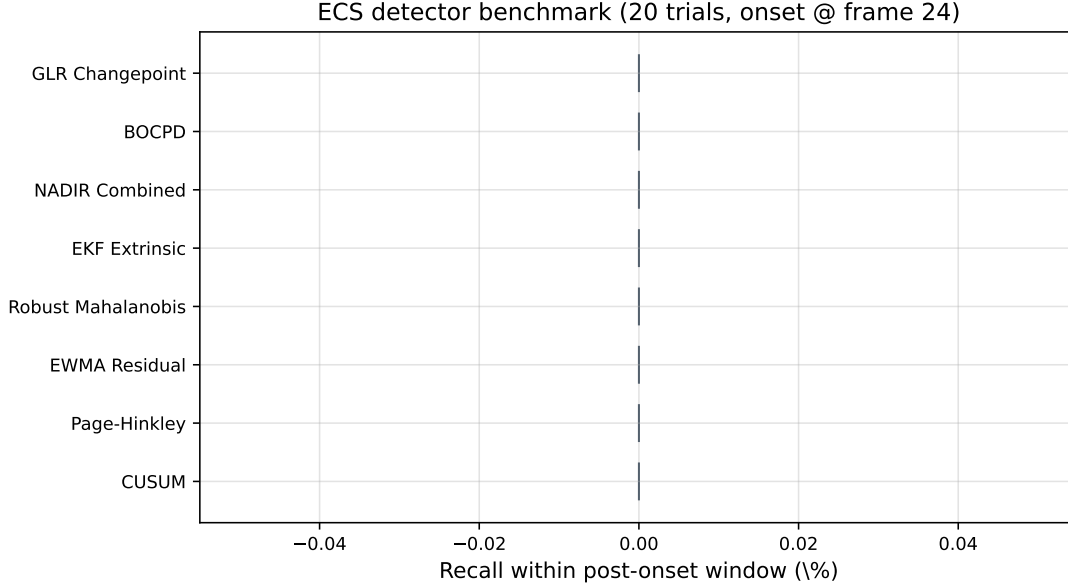


Figure 14: Detector recall within post-onset window (20 Monte Carlo trials, onset at frame 24).

rules out early-warning superiority statements for Pulse in this release unless future experiments on slower ramps with longer horizons show median lead time strictly greater than zero, which is the explicit gate in our claim boundary document. Operators who prioritize minimizing missed drift episodes may therefore prefer static rules or shorter conformal quantiles, accepting higher false-alert rates on drift-free weeks.

Conversely, operators whose key performance indicator is “callbacks per million miles” may accept lower detection rate in exchange for conformal guarantees. The 4.2% drift-free CAUTION rate is close enough to the 5% nominal target that finance teams can treat it as a budgeted operational expense rather than as an unbounded alert storm. Product teams should document which metric they optimize in each pilot exhibit so that mixed results here are not misread as universal inferiority or superiority of either arm.

Conformal gating addresses the complementary pain point: fleet managers who must justify callback volume to finance teams. Holding the drift-free CAUTION rate near 5% while static and raw Pulse emit zero alerts on the same holdout demonstrates that the Mondrian layer is doing substantive work, not merely re-labeling static outcomes. The tradeoff is reduced detection rate (63.4% versus 100%), which is the price of distribution-free calibration under exchangeability assumptions. When those assumptions fail, for example during seasonal tire changes that shift vibration statistics without changing extrinsics, strata must be re-calibrated or expanded.

The ECS slow-drift negative result is equally important for pilot design. At 0.03° per day, eight changepoint detectors with diverse computational profiles all fail to recall drift within twenty frames after onset on forty-eight-frame episodes. Figure 3 shows that physical drift is present in the simulator; the monitor’s difficulty is that proxy innovations remain small relative to thresholds and conformal quantiles over the observation window. Pilots that promise detection within days on CAN-only ingest therefore need either longer observation horizons, richer residuals, or higher drift rates in validation scenarios that match the fleets being onboarded.

Ablation flatlines indicate that benchmark design matters as much as algorithm design. When every preset sees the same 69% Pulse detection, the injected signal swamps layer toggles. Future work should stratify trials by ramp draw, vibration index quantiles, and post-repair age so that robust estimation and gating contributions become identifiable. Tier visual exports help operators build geometric intuition for misalignment before numeric tier ordering is available on matched streams.

Limitations are structural. All data are synthetic; no OEM field traces appear in version one. The prototype is not ISO 26262 certified and must not drive actuation. Conformal guarantees assume exchangeability within strata, which fleet non-stationarity can break. Tier numeric comparisons require held-out JSON on identical proxy batches, not browser-rendered scenes alone. We plan OEM anonymized replay, repair-type stratification, and power analysis relating detectable yaw rate to ingest sparsity Δt .

For practitioners shipping similar monitors, we recommend reporting both calibrated false-alert rate and raw detection rate on identical splits, publishing negative results on slow drift, and versioning benchmark JSON alongside paper revisions. The graphs and heatmaps we include are not decorative; they are views into the same tensors operators will inspect when debugging a misfiring vehicle in pilot. Keeping visualization code deterministic from simulation exports prevents the common failure mode in which paper figures drift from production behavior after a single refactor.

6 Conclusion

We presented a reproducible shadow monitoring benchmark for post-repair ADAS extrinsic drift under sparse fleet telemetry proxies. Huber-robust Mahalanobis fusion with Mondrian conformal gating calibrates drift-free CAUTION emissions to 4.2% at nominal $\alpha = 0.05$. Static shop thresholds achieve higher recall but zero positive lead time over Pulse across five hundred Monte Carlo trials. ECS slow-drift studies at 0.03° per day reveal a shared sensitivity floor among eight change-point detectors. Simulation-derived graphs and heatmaps, exported directly from benchmark JSON and JSONL artifacts, document these behaviors for fleet stakeholders.

NADIR remains a pilot-ready API and SDK prototype, not a safety-certified product [11]. CAUTION events should trigger audit and shop verification workflows, not autonomous control actions. Replication requires only Python 3.11+, the commands documented in `research/REPRODUCE.md`, and the LaTeX sources in `research/paper/`. We hope the open artifacts lower the cost of comparing future monitors on identical footing and make negative results, including delayed detection under ultra-slow drift, part of the published record rather than footnotes in slide decks.

References

- [1] Huber, Peter J. “Robust Estimation of a Location Parameter.” *Annals of Mathematical Statistics*, vol. 35, no. 1, 1964, pp. 73–101.
- [2] Vovk, Vladimir, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.
- [3] Adams, Ryan P., and David J. C. MacKay. “Bayesian Online Changepoint Detection.” *arXiv preprint arXiv:0710.3742*, 2007.

- [4] Page, Ewan S. “Continuous Inspection Schemes.” *Biometrika*, vol. 41, no. 1/2, 1954, pp. 100–115.
- [5] Basseville, Michèle, and Igor V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, 1993.
- [6] Rousseeuw, Peter J., and Katrien Van Driessen. “A Fast Algorithm for the Minimum Covariance Determinant Estimator.” *Technometrics*, vol. 41, no. 3, 1999, pp. 212–223.
- [7] Shafer, Glenn, and Vladimir Vovk. “A Tutorial on Conformal Prediction.” *Journal of Machine Learning Research*, vol. 9, 2008, pp. 371–421.
- [8] Angelopoulos, Anastasios N., and Stephen Bates. “A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification.” *arXiv preprint arXiv:2107.07511*, 2021.
- [9] Welch, Greg, and Gary Bishop. *An Introduction to the Kalman Filter*. University of North Carolina at Chapel Hill, TR 95-041, 1995.
- [10] Angelopoulos, Anastasios N., et al. “Uncertainty Sets for Image Classifiers Using Conformal Prediction.” *International Conference on Learning Representations*, 2021.
- [11] Hegde, Dhruv, and Srivatsan Balaji. *NADIR Pulse SDK and API: Shadow Fleet Monitoring Prototype*. NADIR Systems, 2026. GitHub repository, <https://github.com/DDVHegde100/nadir>. Accessed 18 June 2026.