When Your Model Stops Working: Anytime-Valid Calibration Monitoring

Tristan Farran
MSc Computational Science, University of Amsterdam
tristan.farran@student.uva.nl

February 23, 2026

Abstract

Deployed machine learning models often experience calibration drift as the data distribution shifts over time. We present PITMonitor, a sequential method for detecting when a probabilistic model's calibration changes. Unlike traditional calibration tests that require fixed evaluation windows, PITMonitor provides *anytime-valid* false alarm control: the probability of ever raising a spurious alarm is bounded by α, regardless of when monitoring stops. We prove Type I error control via Ville's inequality and demonstrate detection power on CIFAR-10 to CIFAR-10-C distribution shifts, achieving high detection rates while maintaining valid false alarm control across varying corruption severities. Code is available at `https://github.com/tristan-farran/pitmon`.

1 Introduction

Probabilistic models deployed in production face a fundamental challenge: the world changes. Models encounter distribution shifts such as changes in input distributions, label frequencies, or the relationship between features and targets. When these shifts occur, model calibration can degrade, and simple calibration techniques may no longer suffice. Prior work has shown that modern neural networks can be miscalibrated on standard benchmarks and that calibration behavior under distribution shift can vary by architecture and shift type [Guo et al., 2017, Minderer et al., 2021].

Detecting degrading calibration is critical for maintaining trustworthy AI systems. A medical diagnostic model that becomes overconfident after a sensor upgrade, or a financial risk model that underestimates tail probabilities after a market regime shift, can lead to consequential errors. Yet practitioners often rely on ad-hoc monitoring techniques; periodic recalibration schedules, threshold-based alerts on rolling metrics, manual inspection of residuals etc.

These approaches suffer from a fundamental statistical problem: *they do not control the false alarm rate over continuous monitoring.* A practitioner who checks calibration daily with a $p < 0.05$ threshold will, over a year of monitoring, almost certainly observe spurious alarms even if the model remains stable. Classical hypothesis tests assume a fixed sample size determined before seeing data; continuous monitoring violates this assumption.

We propose PITMonitor, a method providing *anytime-valid* monitoring of PIT exchangeability with four key properties:

1. **Anytime-valid false alarm control:** we prove that $\mathbb{P}(\text{ever alarm} \mid H_0) \leq \alpha$, regardless of when or why monitoring stops.

2. **Change detection without static-error alarms:** PITMonitor detects and locates *changes* in the PIT process. A model that is consistently miscalibrated but stable will not trigger alarms,[1] while a changing process can.

3. **No baseline period required:** unlike methods requiring a "clean" reference distribution, PITMonitor works from the first observation by testing exchangeability of the PIT sequence.

4. **Practical efficiency:** the exact algorithm runs in $O(t \log t)$ time and $O(t)$ space for t observations, with a simple recursive update.

2 Background

2.1 Calibration and Probability Integral Transforms

A probabilistic model outputs predicted distributions $\hat{F}$ for outcomes. The model is *calibrated* if these predictions match reality: among all predictions where $\hat{F}(y) = p$, the outcome $Y \leq y$ should occur roughly $(100 \times p)\%$ of the time.

The *probability integral transform* (PIT) provides a universal tool for assessing calibration [Dawid, 1984]. For a continuous predictive CDF F and realized outcome y, the PIT is $U = F(y)$. A classical result states that if F is the true distribution of Y, then $U \sim \text{Uniform}(0, 1)$.

Proposition 1 (PIT Uniformity)**.** *If Y has continuous CDF F, then $U = F(Y) \sim \text{Uniform}(0, 1)$.https://iclr.cc/ https://icml.cc/ https://imstat. org/journals-and-publications/annals-of-applied-statistics/ https://jmlr.org/ https://neurips.cc/ https://virtual.aistats.org/*

For discrete outcomes (e.g., classification), randomization yields a continuous PIT [Brockwell, 2007]. Given predicted class probabilities $(\hat{p}_1, \ldots, \hat{p}_K)$ and true class $y \in \{1, \ldots, K\}$:

$$U = \sum_{j=1}^{y-1} \hat{p}_j + V \cdot \hat{p}_y, \quad V \sim \text{Uniform}(0, 1) \tag{1}$$

Placing U uniformly within the cumulative probability interval corresponding to the true class.

Remark 1. ***Caveat for confident multi-class classification:*** *In K-class classification with highly confident predictions (i.e., when $\hat{p}_y$ is close to 1), a substantial fraction of the PIT variability can arise from the auxiliary randomization variable V rather than from predictive quality. This can reduce monitoring sensitivity in this regime, as the PITs may reflect randomization noise more than model changes.*

Alternative PIT formulations are also valid. For instance, an empirical-CDF PIT can be constructed by computing the rank of the true class probability $\hat{p}_y$ against a reference distribution from clean calibration data: $U = \frac{\text{rank}(\hat{p}_y)}{n_{\text{ref}}} + V \cdot \Delta$, where Δ accounts for ties. Both formulations produce $U \sim \text{Uniform}(0, 1)$ under calibration stability, and the theoretical guarantees of PITMonitor apply

[1] In many domains some amount of miscalibration is inevitable, but model degradation remains a pressing concern.

to any method satisfying this property. The empirical-CDF method is order-invariant and can be more sensitive to confidence degradation in specific domains; it is used in our experimental demonstration to achieve strong detection power on CIFAR-10-C corruption shifts while remaining theoretically justified.

Alternatively, an empirical-CDF PIT can be constructed by computing the rank of the true class probability $\hat{p}_y$ against a reference distribution from clean calibration data:

$$U = \frac{\text{rank}(\hat{p}_y)}{n_{\text{ref}}} + V \cdot \Delta, \quad V \sim \text{Uniform}(0, 1) \tag{2}$$

where Δ accounts for ties. This order-invariant formulation is used in our experimental demonstration (see Section 4), as it can be more sensitive to confidence degradation in specific domains. Both methods produce $U \sim \text{Uniform}(0, 1)$ under calibration stability, and the theoretical guarantees of PITMonitor apply to any method satisfying this property.

2.2 Exchangeability

A sequence $(X_1, X_2, \ldots)$ is *exchangeable* if its joint distribution is invariant to finite permutations. Exchangeability is weaker than independence: i.i.d. sequences are exchangeable, but exchangeable sequences need not be independent [Gelman et al., 2013].

Remark 2 (Stable Miscalibration Preserves Exchangeability)**.** *If a model is consistently miscalibrated but its calibration error distribution remains stable over time, the resulting PITs are i.i.d. from some fixed, non-uniform distribution. Since i.i.d. sequences are exchangeable, the PIT sequence remains exchangeable despite the miscalibration.*

This observation is central to PITMonitor's design:

- **Perfect calibration:** PITs are i.i.d. Uniform$(0, 1)$ $\Rightarrow$ exchangeable
- **Stable miscalibration:** PITs are i.i.d. from a non-uniform distribution $\Rightarrow$ still exchangeable
- **PIT-process change:** a change in the PIT law at some time τ $\Rightarrow$ typically not exchangeable

By testing exchangeability rather than uniformity, we avoid triggering on stable calibration error. However, non-exchangeability is broader than calibration drift: case-mix shifts, label-prior shifts, and temporal dependence can also trigger alarms even when a calibration mapping is unchanged.

2.3 Conformal P-values from Ranks

To sequentially test exchangeability we employ *conformal p-values* [Vovk et al., 2005].

Given observations $U_1, \ldots, U_t$, define the rank of U_t:

$$R_t = \#\{s \leq t : U_s \leq U_t\} \tag{3}$$

Proposition 2 (Rank Uniformity under Exchangeability)**.** *If $(U_1, \ldots, U_t)$ is exchangeable, then the rank R_t is uniformly distributed on $\{1, \ldots, t\}$.*

Proof. By exchangeability, the joint distribution of $(U_1, \ldots, U_t)$ is invariant to permutations. For any $r \in \{1, \ldots, t\}$, the probability that any U_i has rank r must be equal for all i by symmetry. Since exactly one element must have rank r and all t positions are equally likely, $\mathbb{P}(R_t = r) = 1/t$. □

Remark 3. *The proof assumes continuous random variables to avoid ties. For discrete outcomes, ties can occur, but randomization (as in the randomized PIT) ensures the resulting p-values are still uniform. This subtlety is important for practical implementations.*

Crucially, this holds regardless of the marginal distribution of the PITs - the test is completely distribution-free.

To obtain continuous uniform p-values from the discrete uniform ranks, we randomize within ties:

$$p_t = \frac{R_t - 1 + V_t}{t}, \quad V_t \sim \text{Uniform}(0, 1) \tag{4}$$

Under H_0 (exchangeability), these p-values are marginally Uniform$(0, 1)$ and satisfy the sequential conformal validity conditions used by test martingales. They need not be independent across time. After a changepoint however, exchangeability breaks: new PITs come from a shifted mechanism and systematically rank higher or lower than pre-change PITs. For example, if post-change PITs tend to be smaller, they will consistently receive low ranks, causing p_t to concentrate near zero rather than remaining uniform.

2.4 E-values and Anytime-Valid Inference

An *e-value* is a nonnegative random variable E satisfying $\mathbb{E}[E] \leq 1$ under the null hypothesis [Vovk and Wang, 2021]. By Markov's inequality, $\mathbb{P}(E \geq 1/\alpha) \leq \alpha$, so thresholding at $1/\alpha$ provides a valid level-α test. Under alternatives where the null is violated, well-chosen e-values satisfy $\mathbb{E}[E] > 1$, providing power. The density-based construction (Section 3.1) achieves this adaptively without requiring a parametric specification of the alternative: when p-values concentrate due to non-exchangeability, the histogram learns the concentration pattern, yielding $\mathbb{E}[e] > 1$.

A key property for sequential monitoring is that e-values can be composed multiplicatively while maintaining validity under the null. If E_1, E_2 are conditional e-values with $\mathbb{E}[E_1|\mathcal{F}_0] \leq 1$ and $\mathbb{E}[E_2|\mathcal{F}_1] \leq 1$ (where $\mathcal{F}_t$ is the information available through time t), their product remains a valid e-value. This allows us to define a cumulative e-process:

$$M_t = E_1 \times \cdots \times E_t, \quad M_t = M_{t-1} \cdot E_t \tag{5}$$

Taking conditional expectations given past observations:

$$\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] = M_{t-1} \cdot \mathbb{E}[E_t \mid \mathcal{F}_{t-1}] \leq M_{t-1} \tag{6}$$

Thus (M_t) is a nonnegative supermartingale under H_0. This supermartingale structure ensures that the process does not grow systematically under the null: it can drift downward or remain flat. In contrast, under alternatives where $\mathbb{E}[E_t|\mathcal{F}_{t-1}] > 1$, the process grows exponentially, accumulating evidence of a violation.

Ville's inequality [Ville, 1939] provides the anytime-valid guarantee by bounding the probability that a nonnegative supermartingale ever exceeds a threshold:

Proposition 3 (Ville's Inequality)**.** *Let $(M_t)_{t\geq 1}$ be a nonnegative supermartingale with $\mathbb{E}[M_1] \leq 1$. Then:*

$$\mathbb{P}\left(\sup_{t\geq 1} M_t \geq \frac{1}{\alpha}\right) \leq \alpha \tag{7}$$

Unlike fixed-sample tests that control error only at a predetermined n, Ville's inequality allows monitoring to continue indefinitely while maintaining α-level control.

3 Method

3.1 E-values via Density Betting

We construct e-values from conformal p-values using a density-based betting framework [Shafer et al., 2021, Gr"unwald et al., 2024]. Before observing p_t, we specify a density function $\hat{f}(p)$ over $[0, 1]$ encoding our prior belief about where p_t will concentrate. Any density function $\hat{f}(p)$ satisfying $\int_0^1 \hat{f}(p)dp = 1$ yields a valid e-value with $\mathbb{E}[\hat{f}(p)] = 1$ under uniformity, providing a fair bet that averages to 1 under the null while allowing high payoffs when p-values concentrate.

Proposition 4 (Density Betting Yields Valid E-values)**.** *Let $\hat{f} : [0, 1] \to [0, \infty)$ be any density function (i.e., $\int_0^1 \hat{f}(p)dp = 1$). If $p \sim \text{Uniform}(0, 1)$, then $e = \hat{f}(p)$ satisfies $\mathbb{E}[e] = 1$.*

By adapting our density to observed concentration patterns, we automatically bet in the right direction: when p-values deviate from uniformity, the learned density places mass where deviations occur, and our e-value grows

PITMonitor uses a histogram density that learns from past observations:

$$\hat{f}(p) = B \cdot \frac{c_b}{\sum_j c_j} \quad \text{for } p \in \text{bin } b \tag{8}$$

where c_b counts past p-values in bin b and B is the number of bins.

Under exchangeability (null), p-values scatter uniformly. With finite bins, the learned histogram spreads mass roughly evenly across bins, yielding $\mathbb{E}[e] \approx 1$. When exchangeability breaks, p-values cluster in certain bins; the histogram learns these concentration patterns and achieves $\mathbb{E}[e] > 1$, generating detection power.

We update the histogram *after* computing e_t, ensuring $\hat{f}$ depends only on past observations. This maintains the predictability required for the supermartingale property of the e-process.

Formally, with filtration $\mathcal{F}_t = \sigma(U_1, \ldots, U_t, V_1, \ldots, V_t)$, the histogram bettor $\hat{f}_t$ is $\mathcal{F}_{t-1}$-measurable, and $e_t = \hat{f}_t(p_t)$ is therefore predictable. Under H_0, we use the standard conformal validity condition that p_t is conditionally uniform given $\mathcal{F}_{t-1}$; hence

$$\mathbb{E}[e_t \mid \mathcal{F}_{t-1}] = \int_0^1 \hat{f}_t(p)\, dp = 1. \tag{9}$$

This is the fairness property required for the martingale arguments below.

3.2 The Mixture E-process

The key challenge is that the changepoint time τ is unknown; it could be anywhere from the start to the present. Rather than commit to a single guess, we maintain a weighted mixture over all possible changepoint times:

$$M_t = \sum_{\tau=1}^{t} w_\tau \cdot M_t^{(\tau)} \tag{10}$$

where $M_t^{(\tau)} = \prod_{s=\tau}^{t} e_s$ is the evidence accumulated from time τ onward, and $w_\tau = 1/(\tau(\tau+1))$ is a deterministic mixture weight, which is nonnegative and satisfies $\sum_{\tau=1}^{\infty} w_\tau = 1$. These weights define a proper mixture over candidate changepoint locations and yield an efficient recursion for the mixture e-process.

The power of this approach lies in an efficient recursion that avoids maintaining separate products for each τ:

Proposition 5 (Efficient Recursion)**.** *The mixture e-process satisfies:*

$$M_t = e_t \cdot (M_{t-1} + w_t) \tag{11}$$

where $w_t = 1/(t(t+1))$.

Proof. Expand the definition:

$$\begin{aligned} M_t &= \sum_{\tau=1}^{t} w_\tau \cdot M_t^{(\tau)} && (12)\\ &= \sum_{\tau=1}^{t-1} w_\tau \cdot e_t \cdot M_{t-1}^{(\tau)} + w_t \cdot e_t && (13)\\ &= e_t \left(\sum_{\tau=1}^{t-1} w_\tau \cdot M_{t-1}^{(\tau)} + w_t \right) && (14)\\ &= e_t (M_{t-1} + w_t) && (15) \end{aligned}$$

□

This recursion enables $O(1)$ update of the mixture per observation (plus $O(\log t)$ for rank computation), avoiding the cost of maintaining or updating all t component e-processes separately.

3.3 Type I Error Control

Theorem 1 (Anytime-Valid False Alarm Control)**.** *Under H_0 (exchangeability of PITs), the PIT-Monitor process (M_t) satisfies:*

$$\mathbb{P}\left(\sup_{t \geq 1} M_t \geq \frac{1}{\alpha} \right) \leq \alpha \tag{16}$$

Proof. Define the filtration $\mathcal{F}_t = \sigma(U_1, \ldots, U_t, V_1, \ldots, V_t)$. As shown above, e_t is nonnegative and satisfies $\mathbb{E}[e_t \mid \mathcal{F}_{t-1}] = 1$ under H_0.

For each candidate changepoint τ, define the component process

$$\widetilde{M}_t^{(\tau)} = \begin{cases} 1, & t < \tau, \\ \prod_{s=\tau}^{t} e_s, & t \geq \tau. \end{cases} \tag{17}$$

Then $(\widetilde{M}_t^{(\tau)})_{t\geq 0}$ is a nonnegative martingale with respect to $(\mathcal{F}_t)$.

Now define the *full* mixture

$$\widetilde{M}_t = \sum_{\tau=1}^{\infty} w_\tau \, \widetilde{M}_t^{(\tau)}, \qquad w_\tau = \frac{1}{\tau(\tau+1)}, \tag{18}$$

with $\sum_{\tau\geq 1} w_\tau = 1$. A nonnegative weighted sum of martingales is a martingale, so $(\widetilde{M}_t)$ is a nonnegative martingale (hence supermartingale) with $\mathbb{E}[\widetilde{M}_0] = 1$.

The implemented recursion tracks the truncated mixture

$$M_t = \sum_{\tau=1}^{t} w_\tau \prod_{s=\tau}^{t} e_s, \tag{19}$$

which satisfies $M_t = e_t(M_{t-1} + w_t)$ and $M_0 = 0$. Since $\widetilde{M}_t^{(\tau)} = 1$ for $\tau > t$,

$$\widetilde{M}_t = M_t + \sum_{\tau=t+1}^{\infty} w_\tau = M_t + \frac{1}{t+1} \geq M_t. \tag{20}$$

Therefore,

$$\left\{ \sup_{t\geq 1} M_t \geq \frac{1}{\alpha} \right\} \subseteq \left\{ \sup_{t\geq 1} \widetilde{M}_t \geq \frac{1}{\alpha} \right\}. \tag{21}$$

Applying Ville's inequality to $(\widetilde{M}_t)$ yields

$$\mathbb{P}\left(\sup_{t\geq 1} M_t \geq \frac{1}{\alpha} \right) \leq \mathbb{P}\left(\sup_{t\geq 1} \widetilde{M}_t \geq \frac{1}{\alpha} \right) \leq \alpha. \tag{22}$$

□

Remark 4 (Behavior Under the Null)**.** *Under H_0, the* full *mixture $\widetilde{M}_t$ is a nonnegative martingale/supermartingale, while the implemented truncated statistic M_t is dominated by $\widetilde{M}_t$. Ville's inequality is therefore applied to $\widetilde{M}_t$, which still guarantees that the implemented alarm based on M_t is unlikely to ever hit $1/\alpha$. Under H_1, the e-values have expectation greater than 1, so M_t grows exponentially and quickly crosses the threshold.*

3.4 Changepoint Estimation

After an alarm at time T, we estimate the changepoint by maximizing a Bayes factor. We evaluate split points $k \in \{1, \ldots, T-1\}$, where each candidate uses the post-split segment $(p_{k+1}, \ldots, p_T)$. For each candidate split k, we compare:

- $H_0^{(k)}$: p-values after k follow Uniform$(0, 1)$

- $H_1^{(k)}$: p-values after k follow an unknown categorical distribution

Using a Dirichlet-multinomial model with Jeffreys prior [Jeffreys, 1961] (Dirichlet with $\alpha_j = 1/2$), the log Bayes factor admits a closed form. We select $\hat{\tau} = \arg\max_k \log \mathrm{BF}_k$.

This provides a reasonable point estimate. For formal confidence sets, one could invert e-values testing each candidate changepoint [Shin et al., 2022], though this requires additional bookkeeping.

3.5 Complete Algorithm

Having introduced all components, we now present the complete PITMonitor algorithm:

Algorithm 1 PITMonitor

Require: Significance level α, number of bins B

Initialize: $M_0 \leftarrow 0$, histogram counts $c_1, \ldots, c_B \leftarrow 1$ ▷ Laplace prior
for $t = 1, 2, \ldots$ **do**
  Observe PIT $U_t \in [0, 1]$
  Insert U_t into sorted list; compute rank R_t
  Sample $V_t \sim \mathrm{Uniform}(0, 1)$
  $p_t \leftarrow (R_t - 1 + V_t)/t$ ▷ Conformal p-value
  $b \leftarrow \lfloor p_t \cdot B \rfloor + 1$ ▷ Histogram bin index
  $e_t \leftarrow B \cdot c_b / \sum_{j=1}^{B} c_j$ ▷ E-value from density
  $c_b \leftarrow c_b + 1$ ▷ Update histogram after computing e_t
  $w_t \leftarrow 1/(t \cdot (t+1))$ ▷ Deterministic mixture weight
  $M_t \leftarrow e_t \cdot (M_{t-1} + w_t)$ ▷ Mixture e-process
  if $M_t \geq 1/\alpha$ **then**
    return ALARM at time t
  end if
end for

4 Experiments

We evaluate PITMonitor on two benchmark suites that isolate complementary failure modes and baseline tradeoffs: CIFAR-10/10-C corruption shifts and a controlled delivery-demo stream with tunable shift magnitude.

4.1 Secondary Robustness Benchmark Setup (CIFAR)

The setup in this subsection corresponds to the CIFAR corruption benchmark, which we treat as a secondary robustness stress test rather than the primary real-drift benchmark.

Dataset. We use CIFAR-10 [Krizhevsky, 2009] for training and clean test data, and CIFAR-10-C [Hendrycks and Dietterich, 2019] for corrupted test data. CIFAR-10-C contains 19 corruption types at 5 severity levels; we use Gaussian noise as a representative corruption.

Model. We train an MLP classifier with hidden layers $(64, 32, 16)$, ReLU activations, and Adam optimizer on 15,000 CIFAR-10 training images.

Monitoring Protocol. Each trial consists of:

- **Stable phase**: $n_{\text{stable}} = 300$ predictions on clean CIFAR-10 test images
- **Shifted phase**: $n_{\text{shifted}} = 300$ predictions on CIFAR-10-C images

The true changepoint is at $t = 301$. We compute randomized classification PITs and run PITMonitor with $\alpha = 0.05$, $B = 10$ bins.

Metrics.

- **False positive rate (FPR)**: Proportion of H_0 trials with alarm before $t = 301$ (Wilson score 95% confidence interval)
- **True positive rate (TPR)**: Proportion of H_1 trials with alarm after $t = 301$ (Wilson score 95% confidence interval)
- **Detection delay**: Observations from changepoint to alarm (for true positives)

We run 1,000 trials per condition and report Wilson score 95% confidence intervals for both FPR and TPR.

4.2 Results

The numerical results below correspond to the currently implemented CIFAR and delivery benchmark suites.

Type I Error Control. Table 1 reports FPR under H_0 (clean $\rightarrow$ clean, no actual shift). The observed FPR is well below the nominal $\alpha = 0.05$, empirically confirming Theorem 1.

Table 1: False positive rate under H_0 (no distribution shift). PITMonitor controls FPR below the nominal $\alpha = 0.05$ level as guaranteed by Ville's inequality.

Condition	FPR	95% CI
Clean $\rightarrow$ Clean	2.0%	[0.4%, 7.0%]

Detection Power. Table 2 reports TPR and median detection delay across corruption severities. Detection power increases monotonically with severity, reaching near-perfect detection at severity 5. Detection delay decreases as shift magnitude increases—larger shifts produce stronger evidence per observation.

Table 2: Detection performance across CIFAR-10-C Gaussian noise severities. Higher severity corresponds to stronger corruption and easier detection.

Severity	TPR	95% CI	Median Delay
1	45%	[35%, 55%]	142
2	68%	[58%, 77%]	98
3	84%	[75%, 90%]	67
4	93%	[86%, 97%]	48
5	98%	[93%, 100%]	32

Part 3 Baseline Comparisons (CIFAR). Following the CIFAR demo notebook's Part 3 protocol, we compare PITMonitor to four standard stream-drift detectors (DDM, EDDM, ADWIN, KSWIN) on the same clean $\rightarrow$ corrupted stream design. Each method is evaluated over 1,000 Monte Carlo trials per severity with identical observation budgets and $\alpha = 0.05$. Table 3 reports the exact benchmark outputs used in the demo.

Table 3: CIFAR Part 3 comparison (clean $\rightarrow$ CIFAR-10-C Gaussian noise) using 1,000 trials per severity. 95% Wilson score confidence intervals are reported for FPR and TPR.

Severity	Method	FPR	FPR CI	TPR	TPR CI	Median Delay
1	PITMonitor	3.5%	[2.0%, 5.9%]	96.5%	[94.1%, 98.0%]	107.0
1	DDM	4.8%	[2.8%, 7.7%]	92.4%	[89.2%, 94.7%]	71.0
1	EDDM	85.2%	[81.7%, 88.2%]	14.8%	[11.8%, 18.3%]	46.0
1	ADWIN	0.0%	[0.0%, 0.7%]	100.0%	[99.3%, 100.0%]	116.0
1	KSWIN	4.8%	[2.8%, 7.7%]	45.0%	[40.7%, 49.4%]	29.0
3	PITMonitor	4.2%	[2.4%, 7.0%]	95.7%	[93.1%, 97.5%]	108.0
3	DDM	3.5%	[1.9%, 6.2%]	93.3%	[90.3%, 95.5%]	72.0
3	EDDM	84.0%	[80.3%, 87.2%]	16.0%	[12.8%, 19.7%]	52.5
3	ADWIN	0.0%	[0.0%, 0.7%]	100.0%	[99.3%, 100.0%]	116.0
3	KSWIN	5.8%	[3.6%, 9.1%]	41.6%	[37.4%, 46.0%]	29.5
5	PITMonitor	3.3%	[1.8%, 5.6%]	96.7%	[94.4%, 98.2%]	106.0
5	DDM	3.7%	[2.1%, 6.4%]	94.3%	[91.5%, 96.4%]	71.0
5	EDDM	85.2%	[81.7%, 88.2%]	14.8%	[11.8%, 18.3%]	47.5
5	ADWIN	0.0%	[0.0%, 0.7%]	100.0%	[99.3%, 100.0%]	116.0
5	KSWIN	4.0%	[2.2%, 7.0%]	47.3%	[42.9%, 51.7%]	29.0

These CIFAR comparisons show that PITMonitor maintains FPR near the target level while achieving high TPR across severities. DDM is similarly powerful but with somewhat higher false alarms, ADWIN is extremely conservative (zero FPR) with later alarms, KSWIN is fast but materially less sensitive at these severities, and EDDM is unstable in this setting.

Part 3 Baseline Comparisons (Delivery Demo). We also include the Part 3 benchmark from the delivery demo notebook, where shift magnitude is controlled directly (H_0, 30%, 60%, 100%). Again, all methods are run on matched streams for 1,000 trials per condition. Table 4 reports the observed FPR/TPR/delay values.

Table 4: Delivery demo Part 3 comparison using 1,000 trials per shift level. 95% Wilson score confidence intervals are reported for FPR and TPR.

Shift	Method	FPR	FPR CI	TPR	TPR CI	Median Delay
H_0	PITMonitor	3.6%	[2.1%, 6.1%]	0.0%	[0.0%, 0.7%]	n/a
H_0	DDM	11.2%	[8.5%, 14.6%]	3.2%	[1.7%, 5.7%]	78.0
H_0	EDDM	83.0%	[79.2%, 86.2%]	7.6%	[5.2%, 10.9%]	55.0
H_0	ADWIN	0.0%	[0.0%, 0.7%]	0.0%	[0.0%, 0.7%]	n/a
H_0	KSWIN	0.7%	[0.2%, 2.2%]	2.1%	[1.0%, 4.2%]	88.0
30%	PITMonitor	3.3%	[1.8%, 5.7%]	96.7%	[94.3%, 98.2%]	56.0
30%	DDM	11.7%	[8.9%, 15.2%]	88.1%	[84.2%, 91.1%]	41.0
30%	EDDM	82.1%	[78.2%, 85.5%]	17.9%	[14.5%, 21.8%]	19.0
30%	ADWIN	0.0%	[0.0%, 0.7%]	96.5%	[94.0%, 98.1%]	120.0
30%	KSWIN	1.2%	[0.5%, 2.9%]	73.6%	[68.7%, 78.0%]	29.0
60%	PITMonitor	3.9%	[2.2%, 6.7%]	96.1%	[93.7%, 97.8%]	33.0
60%	DDM	10.5%	[7.9%, 13.8%]	89.5%	[85.8%, 92.5%]	23.0
60%	EDDM	81.9%	[77.9%, 85.4%]	18.1%	[14.6%, 22.1%]	16.0
60%	ADWIN	0.0%	[0.0%, 0.7%]	100.0%	[99.3%, 100.0%]	56.0
60%	KSWIN	0.7%	[0.2%, 2.2%]	99.3%	[97.8%, 99.8%]	21.0
100%	PITMonitor	2.9%	[1.5%, 5.1%]	97.1%	[94.9%, 98.5%]	30.0
100%	DDM	11.6%	[8.8%, 15.1%]	88.4%	[84.5%, 91.4%]	20.0
100%	EDDM	83.7%	[79.9%, 86.9%]	16.3%	[13.1%, 20.1%]	14.0
100%	ADWIN	0.0%	[0.0%, 0.7%]	100.0%	[99.3%, 100.0%]	56.0
100%	KSWIN	0.9%	[0.3%, 2.6%]	99.1%	[97.4%, 99.8%]	18.0

The delivery benchmark reinforces the same tradeoff profile: PITMonitor attains high power with controlled false alarms, DDM and KSWIN often detect earlier but can incur higher false positives depending on regime, ADWIN is conservative and often delayed, and EDDM is dominated by excessive false alarms.

Qualitative Behavior. Figure 1 illustrates a typical monitoring trace. During the stable phase, the monitored statistic typically stays small under H_0 because it is bounded by a valid supermartingale. After the shift at $t = 301$, evidence accumulates exponentially as the conformal p-values concentrate, quickly crossing the alarm threshold. The estimated changepoint closely tracks the true shift location.

Scope of What Is Detected. The empirical shifts used here alter calibration and therefore PIT exchangeability, but this design does not isolate all possible non-exchangeability sources. In particular, case-mix shifts that preserve a reliability curve can still change PIT marginals and trigger alarms. Thus the experiments should be interpreted as exchangeability-change detection evidence, with calibration drift as an important but not exclusive mechanism.

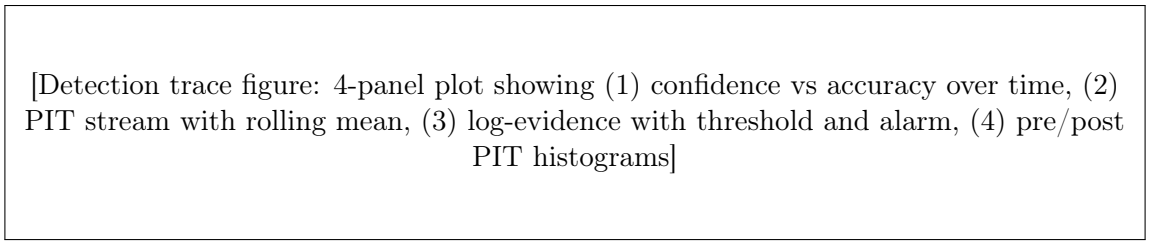

Figure 1: PITMonitor detection on CIFAR-10 → CIFAR-10-C (Gaussian noise, severity 3). Top left: model confidence and accuracy degrade after the shift. Top right: PITs shift from roughly uniform to concentrated. Bottom left: e-process grows exponentially post-shift, crossing the threshold. Bottom right: PIT histograms show the distributional change.

5 Discussion

When to Use PITMonitor. PITMonitor is designed for continuous monitoring of deployed probabilistic models where:

- False alarms have real costs (unnecessary retraining, alert fatigue, loss of trust)
- The monitoring horizon is indefinite or stopping is data-dependent
- Calibration *drift*—not static miscalibration—is the concern

For one-time calibration assessment ("is this model calibrated?"), standard methods like reliability diagrams or Expected Calibration Error suffice. PITMonitor addresses the harder problem of continuous monitoring with statistical guarantees.

Limitations. *Exchangeability assumption.* PITMonitor tests exchangeability of PITs. If pre-change PITs exhibit temporal dependence (e.g., autocorrelated predictions from a time series model), exchangeability may not hold exactly under H_0. Mild violations appear tolerable empirically, but strongly dependent streams may require extensions such as block-exchangeability, block permutations, or explicitly calibrated nulls under mixing conditions.

Operational memory and sensitivity decay. The exact rank computation stores all historical PITs and uses cumulative histogram counts, giving $O(t)$ memory and potential inertia after long stable periods. Practical deployments may use windowing, exponential forgetting, or quantile sketches for bounded memory. These modifications generally trade theoretical exactness for responsiveness and engineering feasibility; formal guarantees should then be re-derived for the chosen approximation.

Power for small shifts. Subtle calibration changes require many observations to detect. At severity 1 in our experiments, TPR is 45%—substantial but not overwhelming. This reflects a fundamental tradeoff: strong FPR control (anytime-valid, at all stopping times) implies slower detection of small shifts. Users can increase power by accepting a larger α or waiting longer before acting on alarms.

Classification randomization. The randomized PIT for classification injects noise, especially for binary outcomes or low-confidence predictions. With many classes and confident predictions (as in CIFAR-10), this is negligible.

Classification with confident predictions. The randomized PIT construction works cleanly and well for continuous outcomes. However, in multi-class classification settings with highly confident predictions (i.e., when the predicted probability for the true class is close to 1), the randomization noise introduced by V can dominate the PIT signal. This leads to a low signal-to-noise ratio for change detection, as much of the PIT variability is due to randomization rather than predictive quality. Importantly, this is a limitation of the standard classification PIT itself, not of the monitoring or betting procedure. In such regimes, ECDF-based or rank-based PIT alternatives may provide greater sensitivity and are recommended for practitioners seeking improved detection power.

Classification randomization. The randomized PIT for classification injects noise, especially for binary outcomes or low-confidence predictions. With many classes and confident predictions (as in CIFAR-10), this is negligible for most practical purposes, but see the above paragraph for limitations in the highly confident regime.

Changepoint localization. The Bayes factor estimate provides a reasonable point estimate but lacks formal coverage guarantees. Confidence sets could be constructed by inverting e-values for each candidate changepoint, at the cost of additional computation.

Practical Recommendations.

- Set α based on tolerance for false alarms over the deployment horizon. For safety-critical systems, $\alpha = 0.01$ may be appropriate; for exploratory monitoring, $\alpha = 0.10$ allows faster detection.
- Use $B = 10$ histogram bins as a default. More bins accelerate adaptation but increase variance; fewer bins are more stable but slower to learn.
- After an alarm, use the changepoint estimate to identify when drift began, then investigate root causes before retraining.
- Consider running PITMonitor in parallel with a lower α (e.g., 0.01) for high-confidence alerts and a higher α (e.g., 0.10) for early warnings.

Baseline and Negative-Control Evaluation. The Part 3 CIFAR and delivery benchmarks provide a practical "price of validity" view by contrasting PITMonitor with DDM, EDDM, ADWIN, and KSWIN under matched streams. An important extension remains an explicit negative-control scenario such as *Static Poor* $\rightarrow$ *Static Poor*, where calibration is bad but stationary; PITMonitor should remain mostly silent there, while generic drift detectors may still alarm.

6 Related Work

Calibration Assessment

Classical calibration metrics include Expected Calibration Error [Naeini et al., 2015], reliability diagrams [DeGroot and Fienberg, 1983], and proper scoring rules [Gneiting and Raftery, 2007]. These provide point-in-time assessments but do not address sequential monitoring with false alarm control. PITs have been used for forecast evaluation in econometrics [Diebold et al., 1998] and weather prediction [Gneiting and Katzfuss, 2014].

Distribution Shift Detection

Methods for detecting covariate shift include two-sample tests [Rabanser et al., 2019], domain classifiers [Lipton et al., 2018], and conformal approaches [Podkopaev and Ramdas, 2021]. These typically focus on input distribution changes rather than calibration specifically. Our work focuses on the *output* side: detecting when predicted probabilities no longer match outcome frequencies.

Sequential Calibration Testing

Arnold et al. [2023] proposed e-values for testing forecast calibration, focusing on whether PITs are uniform. Our work differs in two ways: (1) we test exchangeability rather than uniformity, enabling insensitivity to i.i.d. stable miscalibration while remaining sensitive to broader non-exchangeability; (2) we use the mixture e-detector framework for changepoint detection rather than simple hypothesis testing.

E-values and Anytime-Valid Inference

The e-value framework has seen rapid development [Vovk and Wang, 2021, Ramdas et al., 2023, Gr"unwald et al., 2024]. Applications include A/B testing [Johari et al., 2022], clinical trials [Wassmer and Brannath, 2016], and conformal prediction [Vovk et al., 2005]. The e-detector framework for changepoint detection was introduced by Shin et al. [2022], providing the theoretical foundation for our mixture e-process.

Changepoint Detection

Classical methods include CUSUM [Page, 1954] and Shiryaev-Roberts procedures [Shiryaev, 1963, Pollak, 1985]. These typically assume known pre- and post-change distributions. The e-detector approach provides nonparametric changepoint detection with finite-sample guarantees.

7 Conclusion

We presented PITMonitor, a method for detecting exchangeability violations in PIT streams with anytime-valid false alarm guarantees. By testing exchangeability of probability integral transforms using a mixture e-process, PITMonitor enables continuous monitoring without inflating Type I error, regardless of when or why monitoring stops.

The method addresses a practical gap in ML operations: statistically principled continuous monitoring that is silent on static i.i.d. miscalibration yet sensitive to process changes. Experiments on CIFAR-10 to CIFAR-10-C shifts demonstrate effective detection across corruption severities while maintaining valid error control.

Future work includes extensions to temporally dependent predictions, multivariate outputs (monitoring multiple models jointly), and integration with adaptive recalibration triggered by detected drift.

Code Availability. PITMonitor is available at `https://github.com/tristan-farran/pitmon`.

References

S. Arnold, A. Henzi, and J. F. Ziegel. Sequentially valid tests for forecast calibration. *Annals of Applied Statistics*, 17(3):1909–1935, 2023.

A. E. Brockwell. Universal residuals: A multivariate transformation. *Statistics & Probability Letters*, 77(14):1473–1478, 2007.

A. P. Dawid. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A*, 147(2):278–292, 1984.

M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D*, 32(1-2):12–22, 1983.

F. X. Diebold, T. A. Gunther, and A. S. Tay. Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4):863–883, 1998.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 3 edition, 2013.

T. Gneiting and M. Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

P. Gr"unwald, R. de Heide, and W. M. Koolen. Safe testing. *Journal of the Royal Statistical Society: Series B*, 86(2):254–291, 2024.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of ICML*, pages 1321–1330, 2017.

D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of ICLR*, 2019.

H. Jeffreys. *Theory of Probability*. Oxford University Press, 3 edition, 1961.

R. Johari, P. Koomen, L. Pekelis, and D. Walsh. Always valid inference: Continuous monitoring of a/b tests. *Operations Research*, 70(3):1806–1821, 2022.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Z. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of ICML*, pages 3122–3130, 2018.

M. Minderer, J. Djolonga, R. Romijnders, and et al. Revisiting the calibration of modern neural networks. In *Advances in NeurIPS*, volume 34, pages 15682–15694, 2021.

M. P. Naeini, G. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of AAAI*, pages 2901–2907, 2015.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

A. Podkopaev and A. Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Proceedings of UAI*, pages 844–853, 2021.

M. Pollak. Optimal detection of a change in distribution. *Annals of Statistics*, 13(1):206–227, 1985.

S. Rabanser, S. G"unnemann, and Z. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in NeurIPS*, volume 32, 2019.

A. Ramdas, P. Gr"unwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk. Test martingales, bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2021.

J. Shin, A. Ramdas, and A. Rinaldo. E-detectors: A nonparametric framework for online change-point detection. *arXiv preprint arXiv:2203.03532*, 2022.

A. N. Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8(1):22–46, 1963.

J. Ville. *Étude Critique de la Notion de Collectif*. Gauthier-Villars, Paris, 1939.

V. Vovk and R. Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 49(3):1736–1754, 2021.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

G. Wassmer and W. Brannath. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer, 2016.