processes

MDPI

Article

Valid but Uninformative: AI Process Alarms Beyond the Commissioned Operating Range

Norkobil Saydirasulov [1,2,3] and Young Im Cho [1,*]

[1] Department of Computer Engineering, Gachon University, Seongnam 13120, Republic of Korea; saydirasulov@gachon.ac.kr
[2] Department of Computing Systems Engineering, Samarkand State University Named After Sharof Rashidov, Samarkand 140104, Uzbekistan
[3] Department of Engineering and Communications, Samarkand Campus, ZARMED University, Samarkand 140100, Uzbekistan
* Correspondence: yicho@gachon.ac.kr

Abstract

We ask whether an artificial intelligence (AI)-enabled process monitor's alarms remain informative beyond its commissioned operating range. Conformal alarm rules rely on exchangeability, which deployment shifts can break. Conditioning calibration on the operating point restores validity, not necessarily informativeness: the certificate constrains only the fault-free calibration law, so a diagnostic relying only on that law cannot identify the loss without an additional assumption. Across ten one-class detectors in a seed-controlled Tennessee Eastman campaign with counterfactual twins, validity holds at every evaluated level while informativeness decays. Resolved at ×0.001, the loss is a ramp, not a step, crossing a declared 0.70 floor at ×0.825. There the plant is bistable: fault-free episodes settle into a steady mode or an unsteady one with the purge valve on its stop, the detectors cannot rank the fault inside the unsteady mode, and informativeness above chance scales with the square of the steady share. Disabling that protective override removes the unsteady mode and lifts far-edge receiver operating characteristic (ROC) area from 0.47 to 0.76. Supervised classifiers recover the ranking only when trained on labelled faults in that settled state. Fault-free commissioning measurements, mapped by a curve fitted on labelled faults, predict the loss to a mean absolute ROC-area error of 0.041.

Keywords: process monitoring; fault detection; conformal prediction; operating envelope; one-class detectors; distribution shift; trustworthy artificial intelligence; Tennessee Eastman process

Check for updates

Academic Editor: Jie Zhang

Received: 4 September 2026
Revised: 13 October 2026
Accepted: 26 October 2026
Published: 30 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

An industrial fault-detection system earns its place by how often it is wrong. Alarm floods are a recognised contributor to loss of situational awareness during upsets, so a guarantee on the false-alarm rate is worth more to a plant than another point of benchmark accuracy: in a cyber–physical plant it is the decision-support artefact the operator and the safety case inherit from the monitoring layer. Conformal prediction offers one without distributional assumptions: given a calibration set exchangeable with what deployment will bring, a quantile of the calibration scores bounds the probability that a nominal observation is flagged [1,2]. It has begun to reach process monitoring, and Section 2 places that work beside the monitoring and distribution-shift literatures this paper draws on.

The assumption is the problem. A plant follows load, switches grades and drifts with ambient conditions, and each of those moves the joint distribution of its measurements, so a threshold inherited across such a change may no longer carry a valid guarantee. The theory is equally unforgiving: distribution-free coverage conditional on an arbitrary covariate value is impossible without further assumptions [3], and an operating point is one such covariate. Validity under departures from exchangeability then has to be bought with reweighting, online adaptation or an explicit accounting of the departure. What remains achievable without any of those is coverage conditional on a discrete, observable category, through the Mondrian construction [1], whose conditional validity for inductive conformal predictors is established by Vovk [4].

That construction fits process monitoring unusually well, because a plant's operating point is not a hidden state: the production-rate setpoint, the grade code and the feed selection are quantities the control system already knows and logs, so conditioning on them is a bookkeeping decision rather than an inference problem. We call the resulting key an *eligibility* key (the declared condition under which a particular calibration set is eligible to supply a threshold) and the assurance layer built on it eligibility-conditioned alarm control (ECAC). Figure 1 shows what it adds to a monitoring stack and what it leaves alone.

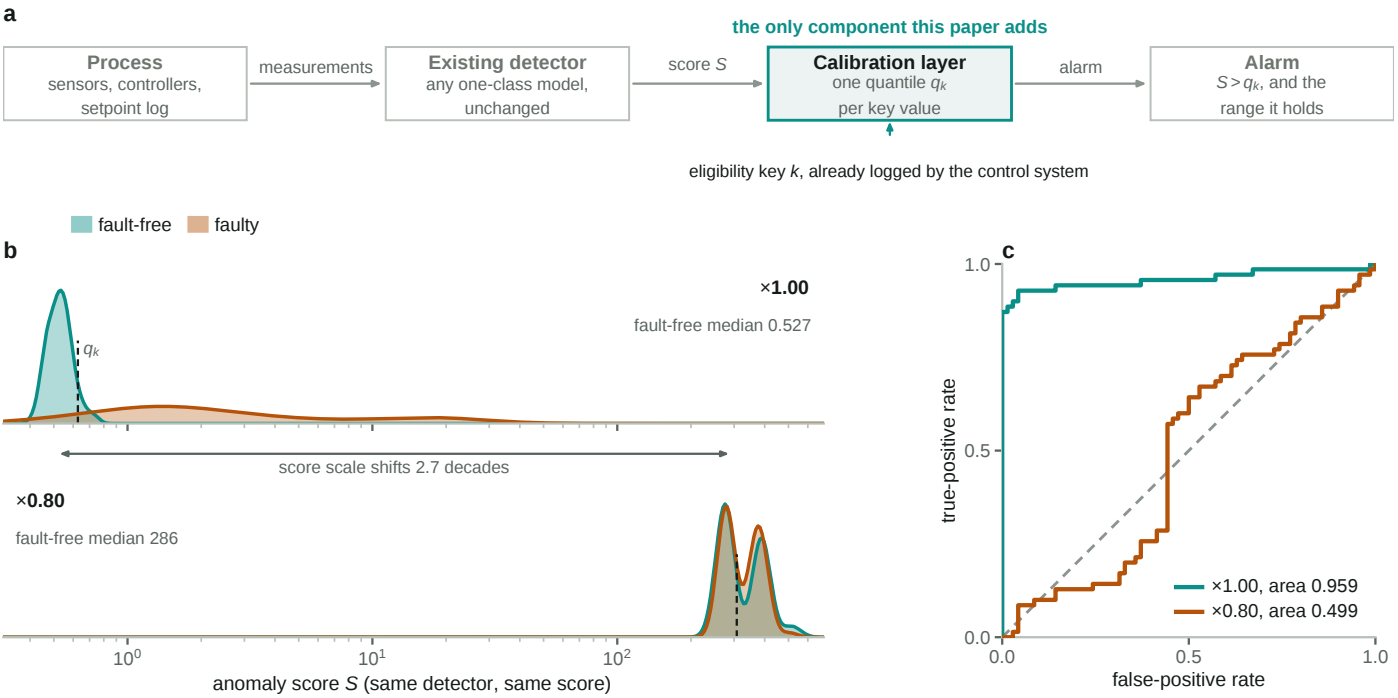


Figure 1. Validity does not imply informativeness once deployment leaves the commissioned operating range. (**a**) The layer's place in a monitoring stack: everything added here operates on the score an existing detector already emits and on an operating variable the control system already logs, so no part of it is a new detector. (**b**) Score distributions for one autoencoder at the commissioned setpoint and at ×0.80, on one absolute axis. Each faulty episode is set against its counterfactual twin (Section 3.6); the dashed line is the level-α quantile q_k of the commissioning calibration pool at that setpoint. The mechanism is in the axis: the fault-free median moves from 0.527 to 286 between the two. (**c**) The ranking those distributions support, for that one detector on deployment-role episodes of the shift campaign, with no averaging over detectors or over levels; each operating point contributes 70 faulty episodes and the 70 twins paired with them. The dashed diagonal marks chance ranking, a receiver operating characteristic (ROC) area of 0.5.

One distinction carries the argument and is fixed precisely in Section 3.1: a calibration layer can be conditioned only on the *operating point*, the condition the control system

commands and logs, while what decides whether the monitor still works is the *operating state*, the dynamical condition the plant occupies as a result.

Conditioning buys validity, and only validity. A conformal threshold is a statement about a scalar score; it says nothing about whether that score still tracks what it was built to detect. Far enough from where the model was fitted, the score may increasingly reflect displacement from the calibration distribution rather than fault-related deviation, so the two questions (is this observation unusual, is this observation faulty) come apart. What costs informativeness is displacement into an operating state whose score distribution no longer supports a fault ranking, and not a change of operating point in itself; how far is far enough we answer with a control rather than an assumption.

What this paper contributes. We are exact about what is new, because the layer is not: ECAC is Mondrian conformal prediction [1] with the taxonomy handed over by the control system instead of learned, and a reader who knows Vovk [4] will find no unfamiliar mathematics in Section 3.7. What is new is not a detector architecture but a measurement about the learned monitor underneath that construction, and a question about trustworthy artificial intelligence (AI) in process monitoring: whether statistical validity stays aligned with operational informativeness under deployment shift. It comes in four parts that depend on each other. *Why the certificate cannot warn:* an elementary proposition formalises that the guarantee constrains only the fault-free calibration law, so no monitoring diagnostic built on that law alone can reveal a loss of fault-ranking informativeness without an additional assumption. *A control that makes the loss measurable:* every faulty episode is re-simulated with the fault switched off and nothing else changed, so informativeness is measured against a fault-free run at the same operating point, the same phase and the same noise realisation; neither the pairing nor the conditioning is new on its own, and the combination is what allows validity and informativeness to be read off the same episodes. *Where the loss opens, and why:* a twenty-six-level warm-started sweep brackets it to a single one-percent increment of the production setpoint, and twenty-one levels placed inside that increment show that what the coarse grid reads as a step is a ramp, crossing the declared floor at ×0.825. Characterising the plant there finds a bistable region: the same plant, at the same setpoint, settles into a steady mode or into an unsteady one in which a second control loop has lost its authority, and an exact decomposition of the pooled receiver operating characteristic (ROC) area shows why the ranking follows the square of the steady share. Disabling the one protective override that the unsteady mode tracks, with every seed unchanged, removes the mode and returns the ranking at the far edge to the steady plateau, so the loss follows the operating state and not the operating point. Which loop does this is a property of this plant and its control structure; what we claim beyond it is the separation itself, that a state no calibration key records can remove the ranking while the certificate holds, and a procedure that finds where. Ten one-class detectors and four supervised classifiers trained in the band lose the pooled ranking across the region alike; the same supervised families recover it at the far edge only when trained on labelled faults in the settled state itself, and not when trained on faults at the same setpoints reached by a step, so what restores informativeness is coverage of the operating state, not of the operating point. *An instrument:* since the loss is not identifiable from the certificate, it must be predicted, and two quantities already collected during commissioning, distribution displacement and fault-free variability, estimate a level's ROC area, through a curve fitted on levels where faults are labelled, with a mean absolute error of 0.041 when extrapolated into the displaced half of the range.

The practical consequence is a commissioning decision rather than a modelling one: in this campaign, extending the operating range represented in the in-control data produced a larger improvement in informativeness than the architectural changes evaluated here. Explainable AI is advanced as the route to trustworthy intelligent process monitoring [5];

we find that the same limitation reaches it, since an explanation cannot be more informative than the alarm it explains, and a real testbed provides an external check.

2. Related Work

This paper sits between five lines of work and takes something from each. Table 1 states what each asks and what it leaves unmeasured, with representative sources [1,5–8]; the paragraphs below give the detail and the full citations.

Table 1. What the lines this paper draws on ask, and what none of them measures. The last row is the question taken up here.

Line of Work	What It Asks	What It Does Not Measure
Process monitoring [6,9]	How many faults a detector catches, and how quickly	Whether the score still ranks faults away from the operating point it was fitted on
Multimode monitoring [7]	How to keep detection working across the modes a plant is known to visit	What happens at an operating point, or in a mode, the commissioning data never covered
Conformal prediction [1,2]	How to guarantee a false-alarm rate without distributional assumptions	What the certified statistic is worth once the rate is guaranteed
Performance estimation under shift [8,10]	How to predict accuracy on unlabelled target data	Whether a calibrated alarm's guarantee and its usefulness come apart, and where
Trustworthy and explainable monitoring [5]	What should be built around an alarm so that it can be trusted	The informativeness of the alarm underneath, which it takes as given
This paper	Whether an alarm that keeps its guaranteed rate under a change of operating point also keeps ranking faulty operation above fault-free operation; where it stops, and which change in the plant's state stops it; and whether that can be estimated in advance from fault-free commissioning data, through a curve fitted where faults are labelled	

Data-driven monitoring of chemical plants grew out of multivariate statistical process control, where a principal-component model of in-control operation supplies Hotelling's T^2 and a squared prediction error, and an alarm is raised when either exceeds a control limit [6]. The Tennessee Eastman problem [11], run under a published decentralised control schemes [6,12], became the common testbed for that line of work, and comparative studies on it report detection rates fault by fault for the classical estimators, which is how the difficulty tiers of its fault set became standard [9]. Machine-learning and deep architectures followed, with recurrent networks applied to detection and diagnosis on the same benchmark [13] and autoencoders scored by reconstruction error becoming a common one-class alternative to the principal-component statistics [14], supported by simulation data prepared specifically for anomaly-detection evaluation [15]. A recent survey maps that landscape in full [16], including the latent-variable methods it grew from [17], and its one-class branch remains active on the same benchmark, each variant scoring how far an observation departs from a model of in-control operation [18,19], down to how much fault-free data such a model needs before its detection rate stops improving [20]. This literature has largely emphasised detection performance (how many faults a monitor catches, and how quickly) and its evaluations are typically conducted at the operating point the model was fitted on.

That a plant changes operating point is not a new observation, and one literature takes it as its subject. Multimode process monitoring treats the plant as occupying several distinct operating modes with transitions between them and builds monitoring that knows which mode is current; its reviews classify methods by whether they address the steady modes or the transitions between them [7,21], place the multimode case among the non-Gaussian,

nonlinear, dynamic and time-varying characteristics that data-based monitoring has to handle [22], and treat the choice of which stretch of plant history to model as a problem in itself [23]. That work and this paper share a premise (a monitor fitted at one operating point need not work at another) and differ in the question. Multimode monitoring asks how to keep detection working across the modes a plant is known to visit, typically by giving each mode its own model or by identifying the current mode before scoring. What is asked here is what happens at an operating point the commissioning data never covered, and what a statistical guarantee is worth there. The two questions meet in Section 4.4.1, where the plant turns out to occupy two modes at one operating point. In that literature's terms those are modes the plant visits; what sets them apart is that neither is commanded or logged, so a mode-keyed model would first have to estimate the mode, and that inside one of them the one-class statistics carry no fault ranking, which only a model trained on labelled faults in that mode recovers (Section 4.5).

Conformal prediction addresses the complementary requirement of a monitor: a false-alarm rate with a distribution-free finite-sample guarantee [1], in the inductive or split form that trains once and calibrates the threshold on held-out data [2,24,25]. The inductive conformal anomaly detector carries the construction to unlabelled data with alarm-rate calibration as its primary statistical guarantee, and to sequential detection within a trajectory [26]; in process fault detection it has recently been compared against classical control limits on principal component and autoencoder detectors [14], and proposed, in preprint, as a p-value chart for multivariate monitoring [27]. Where coverage is wanted separately within known categories rather than only on average, the Mondrian construction calibrates within each category [1]; Vovk [4] establishes which notions of conditional validity an inductive conformal predictor so constructed attains, and the taxonomy has since been carried to conformal predictive distributions with categories formed by binning a regressor's own predictions [28].

That the guarantee rests on exchangeability is well understood, and so are the limits of repairing it: distribution-free coverage conditional on a prescribed value of a continuous covariate is not attainable without further assumptions [3]. The available responses are to reweight the calibration scores when the covariate law changes in a known way [29], to adapt the threshold online from realised coverage [30], or to bound the coverage loss by a weighted total-variation measure of how far exchangeability has been violated [31]. The aim throughout is to restore coverage, or to account for what is lost of it; how much the covered statistic is worth once coverage has been restored is not what these methods are built to track.

A parallel line asks what a monitoring system must offer beyond detection performance to be trusted in service. Explainable-AI methods make a decision inspectable, and their role in AI-based control charts for process monitoring has been reviewed recently [5], and attention has been used to make a convolutional fault classifier's evidence inspectable on the same benchmark [32]; other work approaches trustworthiness through benchmark validity and calibration [33] or through runtime governance of a closed loop [34]. What these have in common is that they take the informativeness of the underlying alarm as given and ask what should be built around it.

One criticism cuts across all of these and lands on this paper as well. A recent review of chemical process monitoring argues that the field's research is algorithm-centric, and that its reliance on simulated benchmarks, the Tennessee Eastman problem chief among them, is part of why methods that perform well in publication do not survive industrial deployment [35]. The campaign reported here is simulated, and that objection applies to it; Section 4.7.2 is a partial answer rather than a complete one, since it carries the cali-

bration question to a physical rig but addresses calibration ageing rather than operating-point displacement.

The directions above are separate and each is well motivated on its own terms. The question taken up here sits between them: does an alarm that keeps its guaranteed rate under a change of operating point also keep ranking faulty operation above fault-free operation, how far can the point move before it stops doing so, and can that distance be estimated in advance from data a plant already collects while commissioning the monitor? It takes the guarantee from the conformal and shift literatures and the detectors from the monitoring literature, and adds what neither supplies alone: a campaign in which validity and informativeness are measured on the same episodes, against a fault-free control that holds the operating point fixed.

3. Materials and Methods

The question of Section 2 is answered by one measurement and a sequence of controls that rule out the alternative explanations of it. This section describes the plant, the campaigns, the detectors and the calibration layer they share. Table 2 maps each control to the objection it answers and to where its result is reported, so the argument can be followed as an argument rather than as an inventory of experiments.

Table 2. Controls used to distinguish the observed loss of informativeness from alternative explanations, read as a reader would: the objection first, then what answers it. The four groups are the measurement, the explanations it could have had, what repairs it, and how far it reaches. SKAB, Skoltech Anomaly Benchmark.

Objection Addressed	Control	Evidence
The measurement		
Where does validity part from informativeness?	Operational-limits map	Section 4.2
Where exactly, and how sharp?	Dense sweep	Section 4.4
Alternative explanations, each ruled out		
Is it the transition rather than the point?	Warm-started control	Section 4.3
Is it an artefact of when the fault arrives?	Fault-onset phase sweep	Section 4.3
Is it a weak-model artefact?	Capacity control	Section 4.5
Is it a property of one-class training?	Supervised control	Section 4.5
Does it coincide with a change in the plant?	Regime characterisation	Section 4.4
What can be done about it		
Can a moving threshold repair it?	Adaptive conformal	Section 4.6.1
What does repair it?	Conditioning and coverage	Section 4.6
Can the loss be predicted before deployment?	Displacement metric	Section 4.4.3
How far the observation reaches		
Does the explanation survive the alarm?	Attribution study	Section 4.7.1
Does any of it hold on real data?	SKAB testbed	Section 4.7.2

3.1. Terms Used Precisely

Four terms carry weight throughout, and the paper turns on the first two being different. An *operating point* is a commanded condition, here the production-rate setpoint multiplier, a number the control system sets and logs. An *operating state* is the dynamical condition the plant actually occupies as a result: steady, transient, or cyclic. A calibration layer can be conditioned only on the point, because that is what a control system can supply, while what decides whether the monitor still works is the state that point puts the plant in. The *commissioned operating range* is the set of operating states represented in the data a detector was fitted and calibrated on, and the *operating envelope* is the part of that range over which the resulting certificate is measured to carry information: the first a property of the commissioning data, the second a measured result. Operationally the range is indexed

by the operating point that produced those data. Unless stated otherwise the detectors are fitted on the routine band, production-rate multipliers ×0.98 to ×1.02 (Section 3.2); the two settings of Section 4.1, which fit at the nominal point alone (A) and on all three regimes (B), are the stated exceptions. "Beyond the commissioned range" therefore means a commanded point outside the band. Whether the plant's *state* there also lies outside the states the band produced is not assumed from the point; it is measured, from fault-free data alone, by the displacement of Section 4.4. *Regime* is used only for the campaign's three named operating points, R_{low}, R_{nom} and R_{high}, and for the *regime-blind* policy named after them: one threshold, fitted at the nominal point and applied unchanged at every other, which is standard practice and the object under test.

3.2. Plant and Simulation Campaigns

The Tennessee Eastman process [11] is the benchmark, run closed loop under the decentralised plant-wide control scheme of the Russell–Chiang–Braatz implementation [6], in the Fortran code distributed with Xavier and de Seixas [13]. Every controller, override and set point referred to below is that implementation's. Two capabilities absent from the published source were added, both inert at their defaults, so the plant reproduces the published behaviour when they are not used. The first exposes the pseudo-random seed of the measurement noise and random disturbances, which the original code fixes inside the initialisation routine; without it there is no way to draw independent episodes. The second applies a multiplier to a controller setpoint after initialisation. Of the twenty setpoints most are cascade slaves, overwritten by an outer loop within one reset time, so they cannot define an operating point at all; `SETPT(11)`, the production-rate loop, is the one that can, and it tracks its setpoint linearly and drives the product stream.

Twelve campaigns were generated, all on central processing units (CPUs), totalling 23,223 episodes; Table 3 lists every one, with its composition, its trip accounting and its role split. This subsection describes the first five, on which Sections 4.1 and 4.2 rest; the others are introduced where they are used.

Table 3. Every campaign in the paper, with the section that uses it, its composition, the trip accounting (episodes that ran to completion and are therefore usable, counted rather than silently dropped) and the split of episodes between detector training, calibration and deployment. Seeds are disjoint across campaigns with two deliberate exceptions: a twin carries its parent's seed, and the intervention re-uses the dense and fine seeds so that each of its episodes has an exact counterpart. The phase and stream campaigns are analysed within themselves and carry no role split.

Campaign	Used in	Composition	Episodes	Usable	Train/Cal./Deploy.
Main	Section 4.1	3 regimes × 8 conditions × 30 clusters	720	714	288/192/240
Band	all	setpoints drawn from [×0.98, ×1.02]	240	239	96/64/80
Shift	Section 4.2	8 levels × 8 conditions × 18 clusters	1152	1110	0/512/640
Commissioning	Section 4.2	64 fault-free runs at each of 8 levels	512	512	0/512/0
Twins	Section 3.6	every faulty episode of the three above, fault switched off	1799	1799	333/652/814
Warm-started	Sections 4.3 and 4.5	8 levels, held 48 h; 1008 faulty, 1008 twins, 512 commissioning	2528	2528	0/1408/1120
Phase	Section 4.3	onset phase swept across the cycle, 2 levels	828	828	none
Dense	Section 4.4	26 levels × (70 faulty, 70 twins, 64 commissioning)	5304	5304	0/1664/3640
Fine	Section 4.4	21 levels at ×0.001, same composition	4284	4284	0/1344/2940
Intervention	Section 4.4.2	the 21 fine and 7 dense levels, purge override disabled	5712	5712	0/1792/3920
Wide	Section 4.6	training set spread across the 8 levels	96	96	96/0/0
Stream	Appendix C	twenty-day fault-free streams, 12 at each of 4 levels	48	48	none
Total			23,223	23,174	

Every episode carries a distinct seed. No seed is shared between the main, shift and band campaigns. By construction each twin deliberately reuses its parent's seed, which is what holds the noise realisation fixed. Twins are generated for calibration-role faulty episodes as well as deployment-role ones: the deployment twins are the fault-free controls, the calibration twins are post-onset units. No episode belongs to both roles. The

commissioning runs are ordinary fault-free episodes at each setpoint, and they, not the twins, calibrate every post-onset threshold reported here.

Every episode is 30 simulated hours at the standard three-minute sampling; the first four hours are discarded as start-up transient, a fault where present begins at hour ten, and each episode contributes windows of identical length so that statistics are comparable. Eight fault conditions (IDV(n) denotes the nth disturbance of the Tennessee Eastman model) span the difficulty tiers of the benchmark literature [6,9]: fault-free operation, the step faults IDV(1), IDV(4) and IDV(5), the random-variation faults IDV(8), IDV(10) and IDV(11), and the slow kinetics drift IDV(13).

A production-rate multiplier is easy to apply and easy to over-interpret, so the range is characterised physically before any statistic is computed on it. Two properties agree on where the plant is genuinely operable. The first is whether the control system still has authority: the stripper steam valve `XMV(9)` (XMV(n), the nth manipulated variable), the loop that runs out of travel first, sits at 47.9% of its range at the nominal rate but is on a stop at every level outside [×0.98, ×1.02], leaving the stripper temperature cascade above it with no authority. The second is product quality: the G/H split moves from 53.7/43.9 at the nominal rate to 58.9/38.9 at ×0.80, a five-point shift a customer would notice. Outside that interval the plant is not running at a different operating point so much as running with a saturated loop, which is worth keeping in view when the monitoring results are read. Appendix B reports the characterisation level by level. The sweep over this range is used as a controlled displacement experiment, and not as a claim about a universal industrial production range. What it fixes is the distance the plant has moved from the conditions the monitor was fitted under, which is the variable this study needs; how much such displacement a particular plant meets depends on its own operating envelope.

Fault-free, the plant completes every episode at every level from ×0.80 to ×1.05. Under fault the envelope is narrower and asymmetric, and the tripped episodes are counted rather than silently dropped: six in the main campaign, one in the band campaign and 42 in the shift campaign, and none in any other campaign of Table 3, all excluded afterwards, because a trajectory that has left the operating envelope is not comparable with one that has not. *All reported rates are therefore conditional on episodes that completed without a trip*. That conditioning is not uniform along the shift axis, and it does not touch the result this paper turns on: no episode trips at any level at or below nominal, so the 576 episodes at ×0.80 through ×0.95 enter complete. Trips occur only at ×1.02 (2 of 144) and ×1.05 (40 of 144), on IDV(1), IDV(5), IDV(8) and IDV(13); Appendix B gives the breakdown by fault.

One design decision about the in-control data governs everything measured later. Fitting the detector at a single production setpoint gives it a zero-width notion of normal, so every off-nominal level breaks the threshold built on it, including ×0.98, a two percent change. That is an artefact of the design rather than a property of the plant: multivariate process monitoring in service is fitted on months of load-following data, so its in-control class has genuine spread. The band campaign therefore draws each episode's setpoint uniformly from [×0.98, ×1.02], the interval Table A2 identifies as the one the control system can hold, and the plant model and its calibration are built on it. The point-nominal design is reported as an ablation in Section 4.7.3.

3.3. A Warm-Started Fixed-Setpoint Campaign

The main, band and shift campaigns all apply the setpoint multiplier at initialisation, so an off-nominal episode is measured while the plant is still answering the change. How long that lasts is a measurement, not a guess: on long fault-free runs the unit statistic falls from 44.5 to 11.9 at ×0.90 and from 10.1 to 2.8 at ×1.05, reaching a level indistinguishable

from the run's own final day only around hour 40 to 46. The six-hour window at hours 10 to 16 therefore sits inside the transient at four of the eight levels.

A further campaign, the warm-started campaign of Table 3, controls for this. The plant is held at the setpoint for 48 h before anything is recorded; the fault, where present, begins at hour 54; and the calibration and deployment windows are the same six-hour windows at the same phase, of a plant that has been at that setpoint for two days. We call this *warm-started* rather than settled, for the reason Section 4.3 gives. Levels, conditions, cluster split and window geometry match the shift campaign, so the only difference between the two is the warm-up, and their seeds are disjoint. It contributes 2528 episodes: 1008 faulty, their 1008 twins, and 512 fault-free *commissioning* runs, 64 per level, on seeds shared with nothing else.

Those commissioning runs matter beyond this one campaign. The post-onset threshold used everywhere else in this paper is calibrated on twins, which no plant can collect; here it is calibrated on fault-free runs at the setpoint, which any plant can. Section 4.3 therefore reports both a warm-started result and one whose calibration path is a deployed ECAC could actually follow.

3.4. Detectors and Scores

Three detector families are used, all fitted on standardised fault-free training episodes and all producing a single scalar per sample, larger meaning more anomalous. Principal component analysis (PCA) retaining 90% of variance supplies Hotelling's T^2, the squared prediction error (SPE) [6], and a scale-normalised sum of the two. A k-nearest-neighbour (kNN) score takes the mean distance to the five nearest training samples, in a Euclidean and a squared variant. A shallow autoencoder (AE) with a 24–8–24 tanh architecture supplies the mean squared reconstruction error.

The conformal layer never inspects the detector; it consumes the score alone.

Four deep temporal detectors are added for the capacity control of Section 4.5: long short-term memory (LSTM) and gated recurrent unit (GRU) sequence autoencoders, a dilated temporal convolutional autoencoder, and a Transformer encoder autoencoder, each reconstructing a one-hour window of all 52 variables and scoring the window's final sample by its reconstruction error. They are trained on exactly the same in-control episodes as the classical detectors (6240 samples from one commissioning period, which is what a plant actually has), so that the comparison is not confounded by a larger training set, and each is fitted from three independent initialisations so that a difference between architectures can be read against the spread within one.

Every architectural and optimisation setting is frozen in the configuration and is the same at every operating level and in every experiment. For the classical detectors: 90% retained variance for the principal-component scores, $k = 5$ for the nearest-neighbour scores, and a 24–8–24 tanh network for the shallow autoencoder. For the four deep detectors: a hidden width of 32, a latent width of 16, Adam at 10^{-3}, batches of 128, and at most 60 epochs with early stopping after 8 epochs without improvement. The stopping criterion is the one place a fitted model consults held-out data, and the 15% split it monitors is drawn from the same fault-free training episodes. Consequently, no hyperparameter and no stopping decision in this paper is taken on a faulty episode, a counterfactual twin, a calibration set, or an episode at an operating level outside the band the models are fitted on. These configurations were fixed for the reported campaign and were not adjusted per operating level or evaluation outcome; a plant cannot retune a monitor for an operating point it has not reached. Which detectors enter which experiment follows the claim under test: the calibration-policy comparison uses five classical variants and the exchangeability diagnostic six, since there the claim concerns the calibration rather than the model, while the limits map, the informativeness measurement and the external check use one detector per family.

Four discriminative models are fitted on labelled faults as a paradigm control. The first is an L_2-regularised logistic regression whose penalty is tuned, not fixed: five-fold cross-validation over twelve values of the inverse penalty, log-spaced from 10^{-3} to 10^{2}, scored by ROC area and run within the training set alone. The second is a random forest of 500 trees with at least two samples per leaf, which is not the library default for either setting. The third is a histogram-based gradient-boosted tree ensemble whose learning rate, depth and number of iterations are chosen by five-fold cross-validated grid search on ROC area over eighteen combinations, and the fourth a multilayer perceptron with one or two hidden layers whose width and L_2 penalty are chosen the same way over nine combinations, with early stopping; both searches run within the training set alone. Each window is presented to all four as the per-variable mean and standard deviation of the standardised measurements over the window, 104 features. That is the information a residual detector draws on, without the reduction to a distance from a fitted manifold, so the contrast with the one-class arm is the training principle and not the input. All four see the same band training episodes as the one-class detectors and the same standardiser, fitted on fault-free training samples; the pre-onset window of every episode supplies the fault-free class and the post-onset window of every faulty episode the faulty one. All are then evaluated on the warm-started campaign against the same counterfactual twins, so the only thing that differs from the one-class arm is what the model was allowed to learn from. The same four families are fitted a second time on a deliberately different training set, the *all-levels* control: labelled windows from the calibration-role episodes of the shift campaign at all eight operating levels, so the models have seen faults at every setpoint they are evaluated at. Those episodes are disjoint from the evaluation campaign, and they reach each setpoint by a step rather than being held at it for two days. Section 4.5 reports the outcome.

3.5. The Episode as the Exchangeable Unit

Consecutive process samples are strongly autocorrelated. A pooled bag of samples drawn from a handful of runs is therefore not exchangeable with a sample from a fresh run, and a per-sample conformal quantile inherits a guarantee it cannot honour. Every episode here contributes exactly one unit: a fixed-length window, summarised by the statistic an operator would actually alarm on.

How strong the dependence is was measured rather than assumed. On fault-free windows of the warm-started campaign, sixteen episodes per level, the lag-1 autocorrelation of the 52 standardised measurements averages 0.45 to 0.70 across the eight levels and reaches 0.95 to 0.99 for the slowest variable at every level, so a 120-sample window carries the information of only 25 to 49 independent samples by the integrated autocorrelation time, fewest at the two edges of the range. Nothing in the guarantee relies on samples within a window being independent: exchangeability is required of episodes, which are generated from distinct seeds, and each contributes one number. The same dependence bears on the displacement of Section 4.4.3, whose covariance is estimated from pooled samples; it does not bias that estimate but makes its nominal sample size optimistic, which is why no interval is attached to D_{shift} and the curve built on it is judged on held-out levels rather than on the sampling distribution of the distance.

Industrial alarm rules require persistence: an alarm fires when the score stays above threshold for m consecutive samples. The matching episode statistic is the m-persistent peak

$$S = \max_{t} \min\left(s_{t-m+1}, \ldots, s_{t}\right), \tag{1}$$

so that $S > q$ holds exactly when the m-of-m rule would have fired somewhere in the window. We use $m = 3$, that is nine minutes. Calibrating the quantile of S therefore controls the probability that a fault-free episode produces a nuisance alarm, which is the quantity

a plant cares about, rather than a per-sample rate that no one operates on. The window length and m are conventions, and Appendix K measures what they contribute (Table A12).

3.6. The Counterfactual Twin as the Fault-Free Control

The two questions asked of the monitored statistic need different controls. Whether the alarm rate is at target concerns fault-free data alone, and the pre-onset window of every episode supplies it: hours four to ten precede fault onset, so that window is fault-free whatever the episode's label, and calibration and deployment units are drawn from the same window in both arms.

Whether the statistic can tell a fault from no fault is a different question, and it needs a control that differs from the faulty window in the fault and in nothing else. Pairing a treated and an untreated run on the same random seed is the classical variance-reduction and paired-comparison design of stochastic simulation, common random numbers [36,37], and the same construction appears in fault-injection assessment as the fault-free "golden run" against which an injected run is judged [38]. We claim no novelty for it; what it is doing here is removing a specific confound, and the pairing is not free: it can in principle induce negative correlation and inflate variance rather than reduce it [39]. We therefore measured it. On a log scale the within-pair correlation between a faulty episode's statistic and its twin's is near zero at the commissioned setpoint (-0.05, -0.00 and 0.06 for the three headline detectors) and moderate at ×0.80, at 0.47, 0.33 and 0.33. Six of the twenty-four detector-level cells are slightly negative: the pairing removes a confound rather than tightening an interval. The pre-onset window will not serve either, because it also differs in elapsed time. The setpoint multiplier is applied once at initialisation while the manipulated variables start from their nominal steady state, so every off-nominal episode is a slow step response: on fault-free episodes the hours 10–16 statistic separates from the same episode's hours 4–10 statistic almost completely at five of the eight setpoints, and still does at hours 22–28. Only at ×1.00 are the two windows interchangeable, so a contrast between them would measure elapsed time as much as the fault. The size of what that leaves in is measurable. For the autoencoder the twin control gives ROC areas of 0.959 at the commissioned setpoint and 0.499 at ×0.80; the pre-onset contrast on the very same episodes gives 0.956 and 0.534. The two agree at the nominal setpoint, as they should, and the gap between them at ×0.80 is the confound.

The campaign therefore carries its own control. Every faulty episode is re-simulated with the fault flag switched off and everything else held fixed (the same seed, the same setpoint, the same noise realisation), giving a *counterfactual twin*. The plant's stochastic stream is driven entirely by the seed, and switching a fault flag does not change how much of it is consumed: the pre-onset trajectories of a faulty run and its twin are bit-identical for all seven fault types used. Every fault-discrimination analysis in this paper is time-matched through these twins, though not all in the same way. The threshold-free measures (ROC area, the paired win rate and the attribution scores) compare a faulty post-onset window directly against its own twin's post-onset window, so operating point, elapsed time and noise are common to both arms and cancel. The detection rate is a threshold comparison rather than a paired one: a faulty post-onset window is scored against a threshold, and the twins enter it as the fault-free arm that threshold is measured on. Both are free of the elapsed-time confound, because every window in every arm sits at the same phase.

A false-alarm rate on the post-onset contrast needs a threshold, and where that threshold comes from decides whether the rate means anything. It cannot be taken from the units the rate is then measured on: for a calibration sample of size n the conformal quantile (2) is the r-th order statistic, so exactly $n - r$ of the n points lie above it and the reported rate is $(n - r)/n$ by construction, whatever the plant is doing.

Nor should it come from twins, though it can. A twin is a run re-simulated with the fault switched off; no plant can collect one, so a threshold built on twins is an evaluation device rather than a calibration procedure, and a result resting on it would not transfer to a real commissioning exercise. A fifth group of episodes therefore exists for this purpose alone: 64 fault-free *commissioning* runs at each setpoint, on their own seeds, at the same episode phase as the units the threshold is applied to. Together with the 8 fault-free episodes already in the calibration role they give a *calibration pool* of 72 fault-free units per setpoint; "commissioning run" refers only to one of the 64, "calibration pool" the 72. Every false-alarm rate and detection rate on the post-onset contrast in this paper is computed from a threshold set on that pool and applied to held-out deployment units, so the two are measurements of one rule, on a calibration path a deployed system could reproduce. The twins are then used only where a counterfactual is the point: the ROC contrast, the paired win rate and the causal attribution reference.

3.7. Split Conformal Alarm Control and Eligibility Conditioning

Given calibration statistics $S_1, \ldots, S_n$ from fault-free windows and a target level α, the threshold is the order statistic

$$q = S_{(r)}, \qquad r = \left\lceil (n+1)(1-\alpha) \right\rceil, \tag{2}$$

which satisfies $\mathbb{P}(S_{\text{test}} > q) \leq \alpha$ when the calibration units and the test unit are exchangeable [1,25]. The probability is over the calibration draw and the test unit jointly, not conditional on the calibration set in hand, and the distinction matters later. When $r > n$ no finite threshold can carry the claim; the implementation returns $q = \infty$ rather than a threshold that would misrepresent the guarantee, and the smallest certifiable level for a calibration set of size n is $1/(n+1)$. The same refusal covers a key that was never calibrated: a value of the eligibility key with no calibration units is not a key with $n = 0$ and a borrowed threshold, but a state the certificate does not reach. The procedure then reports the window as *uncertified* rather than scoring it against the threshold of another key (Algorithm A1). A threshold pooled over all calibrated keys can still be applied as an operational fallback, and Algorithm A1 allows it, but any alarm it raises is labelled uncertified, because the pooled guarantee is marginal over the calibrated keys and says nothing about the one that was not among them. In the campaigns reported here every deployment key is present in calibration by construction, so the branch is never taken.

What is specific to the setting is where the Mondrian taxonomy comes from. A Mondrian category is normally derived from the data or the model (a label, a bin of the model's own predictions) and must be justified as a modelling choice. Here it is read off a variable the control system commands and logs, so the taxonomy is declared before any data is seen and costs no inference.

What the Certificate Constrains, and What It Leaves Free

The separation this paper measures has an elementary reason, which is worth stating precisely because the two quantities are so often treated as one. Write P_0 for the law of the unit statistic on fault-free operation and P_1 for its law under the fault mixture, and define the alarm's *informativeness*, or equivalently its *utility*, as the probability that it orders a faulty unit above a fault-free one,

$$U = \mathbb{P}\big(S_1 > S_0\big) + \tfrac{1}{2}\,\mathbb{P}\big(S_1 = S_0\big), \qquad S_1 \sim P_1,\ S_0 \sim P_0 \text{ independent}, \tag{3}$$

which is the ROC area reported throughout. The tie term is the usual Mann–Whitney convention and is what the estimator computes: every empirical ROC area in this paper is obtained from mid-ranks, so a tied faulty–fault-free comparison counts one half. For

continuous scores the term vanishes. Validity, in the sense of (2), is a statement about P_0 alone: the calibration units and the test unit are both fault-free, and P_1 appears nowhere in it.

ROC area is therefore used throughout as a threshold-free measure of ranking informativeness: it evaluates the separation between P_0 and P_1 independently of any alarm threshold, which is what lets the two properties be read against each other at a single operating level.

Two estimands of informativeness appear, and they are kept apart. The *pooled* ROC area, Equation (3) with P_0 the law of the counterfactual twins at that level, compares every faulty episode with every fault-free one: it is the probability that a faulty episode drawn at random outscores a fault-free episode drawn at random, which is the comparison a deployed alarm faces, since it never sees the counterfactual. The *paired win rate* compares each faulty episode only with its own twin, $\mathbb{P}\big(S_1^{(i)} > S_0^{(i)}\big)$, and asks whether the fault moves the score at all when the operating point, the phase and the noise realisation are held fixed. Both are estimated on the same pairs, and the bootstrap resamples whole pairs for both, so the pooled estimate's interval respects the pairing although its estimand does not use it. Where the text says "ROC area" without qualification it means the pooled estimand, the area under the empirical ROC curve of faulty episodes against their twins, and the paired win rate is named as such wherever it is used; the two can differ widely, and where they do (Section 4.5) the difference is reported as a finding.

Proposition 1. *Fix $\alpha \in [1/(n+1), 1)$, $n \geq 1$ and $u \in [0,1]$. There exist P_0 and P_1 such that the split-conformal rule (2) has marginal false-alarm rate exactly $(n+1-r)/(n+1) \leq \alpha$, which equals α whenever $(n+1)\alpha$ is an integer, and $U = u$. In particular a rule that meets its false-alarm guarantee, with equality, may have $U = 1/2$, ranking faulty and fault-free units no better than chance, or $U < 1/2$, ranking them backwards. We call such a rule* valid but uninformative*: it delivers the certificate it claims, over a statistic that carries no information about the event the alarm exists to catch.*

Proof. The guarantee in (2) is a statement about the joint law of $n+1$ exchangeable draws from P_0, so it holds for every P_0 and is unaffected by P_1. If P_0 is continuous the rank of the test statistic among the $n+1$ draws is uniform, so $\mathbb{P}(S_{\text{test}} > S_{(r)}) = (n+1-r)/(n+1)$ exactly; with $r = \lceil (n+1)(1-\alpha) \rceil \leq n$ this lies in $(\alpha - 1/(n+1), \alpha]$ and equals α when $(n+1)\alpha$ is an integer. It remains to choose a continuous P_0 and a P_1 with $U = u$. For $u \in (0,1)$ take $P_0 = \mathcal{N}(0,1)$ and $P_1 = \mathcal{N}(\delta, 1)$ with $\delta = \sqrt{2}\,\Phi^{-1}(u)$; then $S_1 - S_0 \sim \mathcal{N}(\delta, 2)$, ties have probability zero, and $U = \Phi(\delta/\sqrt{2}) = u$. Setting $u = 1/2$ gives $\delta = 0$, that is $P_1 = P_0$. For the endpoints, where Φ^{-1} is not finite, take P_0 uniform on $[0,1]$ and P_1 uniform on $[2,3]$, which gives $U = 1$, or on $[-2,-1]$, which gives $U = 0$. In every case the false-alarm rate is the one computed above, because it does not involve P_1. □

The distinction the proposition turns on is not new, and it has a name. The conformal literature separates *validity*, which is automatic and holds whatever the underlying model, from *efficiency*, which is not guaranteed and has to be optimised separately: conformal predictors "are guaranteed to have a prespecified coverage probability" but "widely differ in their efficiency" [40]. What follows is that distinction carried into one-class monitoring, where the quantity at stake is not the width of a prediction set but the ability to rank faulty operation above fault-free operation. It is then pushed one step further: the loss of informativeness cannot be identified from the certified law alone, so it cannot be detected and must be predicted. Its content is entirely in the first sentence of the proof: P_1 is a free parameter of the deployment and the certificate does not touch it. Two consequences shape the rest of the paper.

Throughout, a *diagnostic* means a monitoring signal computed from the deployed scores and the fault-free calibration law alone (an alarm rate, a conformal p-value, a drift statistic on the score distribution) and not any procedure that has access to fault labels or to an assumed model of what a fault does to the score.

Corollary 1. *No monitoring statistic computed solely from the fault-free calibration law identifies U without an additional assumption linking P_1 to P_0. Any two deployments sharing P_0 are indistinguishable to such a statistic, and by Proposition 1 they may have any pair of informativeness values.*

So the loss is silent by construction. An instrument built solely on the fault-free calibration law can report that P_0 has moved, which is a different statement and says nothing about whether the ranking survived. What remains is to measure U where labels exist and to predict it elsewhere from something observable. Section 4.4 does both: it measures U against counterfactual twins along a dense operating axis, and asks whether the movement of P_0 itself (computable from fault-free commissioning runs, before any fault is seen) carries it.

Algorithm A1 states the whole procedure in one place. Three calibration policies are compared, and they differ only in how the key is defined. The *regime-blind* policy calibrates on the nominal regime and deploys the resulting threshold everywhere, which is standard practice and the object under test. The *pooled* policy merges all regimes into one quantile. The *eligibility-conditioned* policy computes one quantile per value of the eligibility key, in the Mondrian sense [1,4]; the key must be an observable operating condition, and here it is the production-rate setpoint, which the control system sets and logs. That is the whole of the construction.

The key is not the setpoint alone, and the paper's own diagnostic is what shows it. Section 3.6 reports that the unit statistic at hours 10–16 differs systematically from the same episode's hours 4–10 at five of the eight setpoints: within one setpoint the plant is not stationary in the elapsed time since the operating change. Exchangeability of the calibration and deployment units therefore rests on the setpoint *and* on their sharing an episode phase, which this design fixes by construction: every unit in every arm is the same six-hour window at the same offset from the same kind of transition. The honest statement of the guarantee is conditional on both: $K = (\text{operating point},\ \text{phase})$, with phase held constant here rather than varied. A plant deploying this layer on windows drawn at arbitrary times after a load change would need the elapsed time, or a steady-state flag, in its key as well; a single logged scalar is enough only because the phase is pinned. Section 4.3 measures what happens when the phase is moved to steady state instead.

Locating a boundary. Two estimators the results use are fixed before any number is quoted, and both are stated where they are first used rather than here, so that the definition and the number a reader checks it against sit on the same page: the *boundary* on the operating axis in Section 4.4 (a declared floor of 0.70 on the detector-averaged ROC area, a point-estimate crossing found by linear interpolation between adjacent levels, and a separate interval from resampling the counterfactual pairs) and the *displacement* that predicts it in Section 4.4.3, computed on fault-free commissioning runs alone. The floor sits above the chance value so that the boundary marks where the ranking stops being *usable* rather than where it stops existing, and it is a convention whose influence on the located crossing is measured in Section 4.4 rather than assumed away.

3.8. Diagnostics and Baselines for the Calibration Policy

Two things are needed before the policy can be judged: a test of whether the eligibility key is fine enough, and a baseline that holds a rate without assuming exchangeability at all.

A single realised false-alarm rate above α proves nothing. For a given calibration set the conditional false-alarm probability is a fixed number; across repeated calibration draws it follows a Beta law centred just below α and exceeds α in a substantial fraction of them by construction: 0.37 at a pool of 64, rising towards one half as the pool grows. We therefore compare the observed mean realised false-alarm rate, averaged over many calibration resamples, against its permutation null, obtained by reshuffling which fault-free units are calibration and which are deployment. A significantly larger observed value is evidence that conditioning left structure behind. It is a Monte-Carlo permutation test with $B = 1000$ permutations, so p-values carry the usual $(B + 1)$ correction and a Holm adjustment across the table.

The resampling counts are the same throughout. Every interval on a ROC area is a 90% percentile interval over 1000 pair-preserving bootstrap resamples, in which a counterfactual pair is resampled whole so that a faulty episode and its twin are never separated. The crossing of Section 4.4 carries no bootstrap interval: it is located between two simulated levels, for the reason given there. The confidence level is a convention, so its effect is measured: every one of the 165 detector–level intervals on the shift, dense and fine campaigns was recomputed at 95% from the same pairs. Eleven change a binary reading (whether the interval contains one half, or on which side of the 0.70 floor it lies), every one of them an interval whose 90% end lay within 0.012 of one half or of the floor, and none of them is a reading this paper draws a conclusion from: the floor is applied to the detector-averaged point estimate, and the chance statements concern levels where all three detectors sit near one half. The dense step from ×0.82 to ×0.83 remains non-overlapping at 95% for each of the three detectors (Appendix M, Table A14). The number of detectors, levels and comparisons bears on how the intervals are read, and the paper is built so that no conclusion rests on one of them. Three kinds of statement are kept apart. The formal hypothesis test on which a conclusion rests is the exchangeability diagnostic below, a permutation test with a Holm correction; the rank-correlation p-value of Section 4.4.3 is reported as a description of strength, not as a test. The four outcomes of the intervention of Section 4.4.2 (control, mode, informativeness and safety) are prespecified validation checks: each was stated as a directional prediction before its campaign ran and is reported as met or not met, with no significance level attached. Everything else is estimation, including the paired intervals on the intervention effect, which were computed afterwards: intervals are reported as descriptions of precision, and each conclusion is drawn from a pattern across levels and detectors or from a point-estimate crossing defined in advance, not from a single interval. In the exchangeability diagnostic a calibration subsample of 40 units is drawn to compute each realised rate; the observed value averages 200 such draws and each permutation replicate averages 50, an asymmetry that makes the observed statistic the less variable of the two and the test correspondingly conservative in the direction that matters.

Adaptive conformal inference abandons exchangeability outright and updates the level online from realised errors,

$$\alpha_{t+1} \;=\; \alpha_t + \gamma\big(\alpha - \mathrm{err}_t\big), \qquad \mathrm{err}_t = \mathbf{1}\{S_t > q(\alpha_t)\}, \tag{4}$$

which drives the long-run empirical error frequency to α for an arbitrary sequence [30].

Both implementations of (4) are run. The *bounded* form clips α_t to $[0, 1]$ at every step, which is the defensive choice a practitioner reading α_t as a miscoverage level is likely to write. The *unbounded* form lets α_t drift negative, in which case the quantile stays at $+\infty$ and the rule is silent until the recursion climbs back; it is the form the convergence argument of Gibbs and Candès [30] is stated for. The two agree when some threshold inside the calibration range can meet the target and diverge when none can, which is the post-change situation studied here. Weighted conformal prediction [29] is not compared separately:

with a discrete, observed key its likelihood ratio is constant within each key value, so it reduces to the conditional construction already under test.

The comparison runs on continuous fault-free streams rather than episodes, an operating change modelled as a switch from a nominal run to a shifted one, with recovery judged *sustained*: a trailing twenty-window alarm rate below 1.5α held for forty windows. That criterion needs sixty windows of headroom, so the held-out runs at each level are concatenated into one fault-free stretch, giving a post-change horizon of 474 windows, or four months, on which a late recovery is confirmable. Appendix C gives the stream construction, the two ways of supplying err_t without fault labels, and the injection used for the fault-prevalence variant.

3.9. Attribution and Counterfactual Twins

For the squared-error detectors the score is a sum of squared deviations, so the summand for variable j is its contribution and attribution is an exact decomposition; Hotelling's T^2 and the combined statistic admit no such decomposition and are excluded (Appendix D).

The attribution is scored from two baselines, and the difference between them is part of the result. The *deployable* attribution differences the per-variable contributions against the episode's own pre-onset window,

$$\Delta_j \;=\; \overline{c_j}\big|_{\mathrm{post}} - \overline{c_j}\big|_{\mathrm{pre}}, \tag{5}$$

so that a variable which was always noisy gets no credit for staying noisy; only observable data enters, and this is what a plant can compute. The *twin-baseline* attribution differences the same contributions against the counterfactual twin's post-onset window. It is not deployable, but it is time-matched in the sense of Section 3.6, so the gap between the two separates detector from baseline.

The reference comes from the twin. Holding the noise realisation fixed, the post-onset difference is the causal effect of the intervention on each process variable,

$$\tau_j \;=\; \overline{\big|x_j(\text{fault}) - x_j(\text{no fault})\big|}\big/\sigma_j, \tag{6}$$

with σ_j the training scale. Attribution is scored by top-k agreement, the appropriate measure when the true effect is sparse; with 52 variables and $k = 3$, chance top-3 precision is $3/52 = 0.058$.

3.10. A Regime-Conditioned Detector

The regime-conditioned autoencoder (RC-AE) is a temporal convolutional autoencoder whose feature maps are modulated by the operating point through feature-wise linear modulation (FiLM) conditioning: a small network maps the eligibility key to per-channel scale and shift parameters applied after each convolution, initialised at the identity. The model is asked to learn one in-control manifold per operating point, with the operating point supplied, not inferred. Its unconditioned counterpart is Conv-AE, the same backbone with the conditioning removed, so any difference is attributable to the conditioning and not to the architecture. The standardisation is deliberately frozen: every variant uses the scaling fitted on the band training episodes, so the comparison isolates the effect of what the *model* has seen rather than mixing in a change of preprocessing.

Three training conditions are compared. Under *band* training the model receives exactly the in-control data every other detector receives. Under *wide* training it receives fault-free episodes at every level of the shift axis, the position of a plant that could afford commissioning runs across its operating range. Because the wide set is also larger, a third condition holds the episode count at the band value and spreads those episodes across the levels.

3.11. External Check on a Real Testbed

The Skoltech Anomaly Benchmark (SKAB) [41] is a water-circulation rig with eight sensors at 1 Hz, one anomaly-free reference session and 34 labelled experiment files. Its structure mirrors the simulation design closely enough to test the same claims. The pump's volume flow rate is an observable operating variable taking three clearly separated levels across sessions, and the anomaly-free reference session exists only at the highest of them, precisely the position of a plant that calibrates once, at whatever point it happened to be running, and then deploys everywhere. The machinery is unchanged: 120-sample windows, the same persistence statistic, the same split conformal construction, the same monitor. Files, not windows, are split between calibration and deployment, so no session straddles the two. Counterfactual twins do not exist for real data, so attribution is not scored there.

One property of the dataset bounds what the external check can show. The three flow levels line up almost exactly with the three experiment campaigns the rig was run in (the low-flow files are the valve-manipulation experiments and nothing else), so conditioning on flow is close to conditioning on the campaign. The comparison this paper relies on is therefore the one that needs no such separation: a single flow level, reference session against other sessions at the same level.

4. Results

Table 2 is the map of what follows: each row names an objection, the control that answers it, and the section where its result is reported.

Throughout, $\alpha = 0.10$, and every fault-discrimination number is measured against counterfactual twins as Section 3.6 describes. *Informativeness* means one thing in this paper: the ability of the score to rank faulty units above fault-free ones in aggregate, over the prespecified mixture of seven fault conditions, measured as the area under the ROC curve. It is not fault-specific detectability: a score at chance on the mixture can still rank one condition well. The reported rate is the share of fault-free deployment episodes in which the persistence rule fired at least once: an episode-level nuisance-alarm probability, not a per-sample false-alarm rate.

4.1. Calibration Policy Decides Validity, Not the Model

Table 4 and Figure A1 give the comparison. A regime-blind threshold raises an alarm on essentially every fault-free episode away from the calibration regime: 0.992 at low rate and 0.747 at high rate, averaged over the five detectors. The striking part is that this holds when the detector has been trained on all three regimes: the model is not what fails, the threshold is.

Pooling is not a remedy. It violates the target in one regime (0.275 at low rate) and at the same time drives the false-alarm rate to 0.000 at nominal and 0.015 at high rate. The pooled quantile is dragged upward by the most extreme regime, and takes the detection rate with it: 0.646 at nominal and 0.623 at high rate, against 0.960 and 0.713 for the conditioned policy. Invalid in one regime and insensitive in the others, it combines both failure modes rather than trading between them.

Eligibility conditioning, the ECAC layer of Figure 1, returns the false-alarm rate to target in every regime: 0.095, 0.088 and 0.089.

One quantity in this comparison does not depend on the calibration policy at all. If the threshold is instead calibrated on the twins, on fault-free data at the same operating point and the same point of the transient, the detection rate is 0.666, 0.854 and 0.817 in the three regimes at a realised false-alarm rate of 0.086, whichever policy is nominally in force. That is the detector's own capability: the policies differ in validity, not in what the detector can see.

Table 4. False-alarm and detection rates at $\alpha = 0.10$, averaged over the five detectors, for the main campaign. Setting A trains the detector on the nominal regime only; setting B trains it on all three. False-alarm rates are episode-level nuisance-alarm probabilities. Table A1 gives an episode-level 95% interval for every cell and for the paired differences between policies.

		False-Alarm Rate			Detection Rate		
Training	**Calibration Policy**	**Low**	**Nominal**	**High**	**Low**	**Nominal**	**High**
Nominal only	regime-blind	1.000	0.075	1.000	1.000	1.000	1.000
Nominal only	pooled	0.258	0.000	0.000	0.329	0.000	0.003
Nominal only	eligibility-conditioned	0.055	0.075	0.132	0.286	1.000	0.594
All regimes	regime-blind	0.992	0.088	0.747	1.000	0.960	0.939
All regimes	pooled	0.275	0.000	0.015	0.606	0.646	0.623
All regimes	eligibility-conditioned	0.095	0.088	0.089	0.551	0.960	0.713

The Diagnostic Detects No Calibration–Deployment Discrepancy

Across six detectors and three regimes, the diagnostic of Section 3.8 rejects in none of the eighteen cells. The smallest raw p-value is 0.074 (PCA-combined at high rate); after the Holm step-down correction every adjusted value is 1.000. The permutation null has mean between 0.0945 and 0.0998 and a 95th percentile between 0.171 and 0.191, which is the width a single realised rate has at these calibration sizes, and the reason individual cells above α are expected rather than alarming. Failing to reject is not evidence of exchangeability: the test is directed at one alternative only, a calibration threshold systematically too permissive for the units it will be applied to.

4.2. Validity Is Restored; Informativeness Is Not

Figure 1 established the mechanism on one detector at two setpoints. Figure 2 asks how far and for which detectors, walking the production setpoint away from the band the plant model and its calibration were built on and reporting two quantities against that one axis: the false-alarm rate a plant would experience, and the area under the ROC curve of the statistic that rate is attached to.

Validity behaves as the theory promises (Figure 2a). Both rows of that figure are computed on post-onset units, so the vertical reading compares two properties of one population. Inside the band the whole scheme works: on held-out band episodes the realised false-alarm rate is 0.133 against a target of 0.10, with a detection rate of 0.900 and an ROC area of 0.954. Outside it the frozen threshold is void, flagging every post-onset fault-free unit at ×0.80 through ×0.90 and at ×1.05; only inside ×0.98 to ×1.02 does it stay near target (0.000 to 0.127). Re-calibration restores the finite-sample conformal guarantee at every level. That guarantee is marginal over the calibration draw, not a promise about the set in hand, and the realised rates are draws from it: mean 0.067 across the twenty-four detector–level cells, from 0.013 to 0.175, with two above the nominal α. That is the spread permitted by a calibration set of 72 units, and neither the operating level nor the detector determines which way a given cell falls.

Directly beneath those corrected rates, the statistic they certify is losing its grip on the fault (Figure 2b). Inside the band the statistic is strongly informative: 0.945 to 0.963 on the band's own held-out episodes and 0.920 to 0.991 at ×0.98 through ×1.02. Away from the band it decays, monotonically and in step across all three detectors: 0.875 to 0.957 at ×0.95, 0.695 to 0.777 at ×0.90, 0.612 to 0.752 at ×0.85, and at ×0.80 it reaches 0.491 to 0.510, with pair-preserving 90% bootstrap intervals of $[0.443, 0.580]$, $[0.418, 0.565]$ and $[0.424, 0.575]$ that all contain one half. On the upper side the decay is shallower, 0.576 to 0.786 at ×1.05. The paired win rate, the probability that a faulty episode outscores its own twin, agrees: 0.514, 0.479 and 0.550 at ×0.80 against 0.979 to 1.000 at nominal.

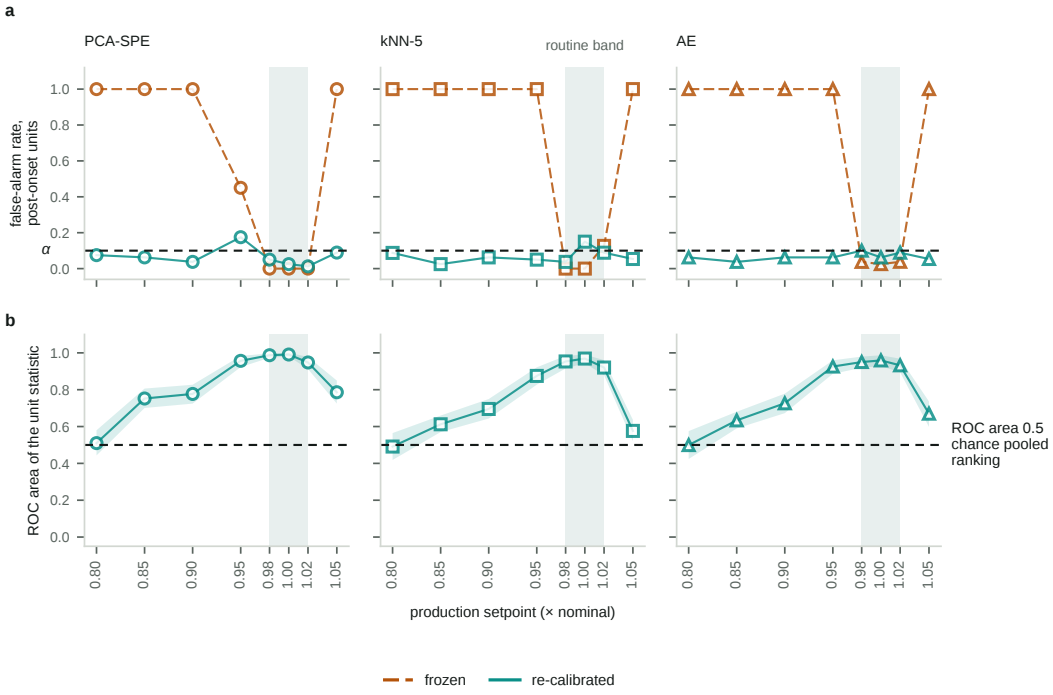


Figure 2. The paper's central measurement, one column per detector and one setpoint axis for both rows, so the divergence is read vertically. The shading marks the routine band, ×0.98 to ×1.02. (**a**) False-alarm rate, with the frozen policy dashed and the re-calibrated one solid, so the two are separated by more than hue; the marker identifies the detector. (**b**) Area under the ROC curve of the re-calibrated statistic, the solid series of panel (**a**) and only that one, measured against counterfactual twins; the band is the 90% pair-preserving bootstrap percentile interval over 1000 resamples of the pairs. Each level uses 64 calibration episodes and 80 fault-free deployment episodes, and 70 counterfactual pairs, 69 at ×1.02 and 46 at ×1.05 after the trip exclusions of Section 3.2. Three of the ten one-class variants are drawn, one from each family; the full set is in Table A4 and Section 4.5. The horizontal dashed lines mark the target rate $\alpha = 0.10$ in (**a**) and chance ranking, an ROC area of 0.5, in (**b**). PCA-SPE, squared prediction error of a principal component analysis model; *k*NN-5, five-nearest-neighbour distance; AE, autoencoder.

A pooled figure at chance does not mean the score has gone blind to every fault. At ×0.80 the seven conditions separate: IDV(1), the A/C feed ratio step, is still ranked at 0.80 to 0.82, while IDV(8), IDV(11) and IDV(13) sit at or below 0.46 (Table A6). Averaging over conditions rather than pooling over episodes gives 0.526 to 0.538, so the aggregate is not a uniform collapse but a mixture in which one condition survives and six do not. *Informativeness* is defined over the prespecified mixture (Section 3.7), because a plant does not choose which fault it gets; a scheme aimed at IDV(1) alone would still work at ×0.80.

Table 5 puts the two quantities on one line per setpoint, and the last row of the low side is the paper. At ×0.80 the rule fires on 7.5% of fault-free episodes and 9.5% of faulty ones: the ROC area is 0.500 and its interval contains one half, so the two rates are one rate. At the nominal setpoint the same construction fires on 7.9% of fault-free episodes and 94.3% of faulty ones. The certificate reads the same on both rows.

The mechanism has two parts. The first is displacement: far from the band the score is dominated by the distance between the plant's operating state and the manifold the detector was fitted on. On warm-started fault-free runs the median unit statistic is 23 to 757 times its nominal value at ×0.80, depending on the detector, against 1.0 to 1.2 inside the band, so a fault's contribution is a rounding error on that scale. The second is variability at the far edge, which Section 4.3 characterises.

Table 5. The operational-limits map, averaged over the three detectors. The threshold at each level is calibrated on the 72-unit fault-free calibration pool (64 commissioning runs and 8 calibration-role episodes) at the same setpoint and episode phase, and applied to held-out deployment units. ROC area is pooled over the seven fault conditions; the macro column averages over them instead.

Setpoint	False-Alarm Rate	Detection Rate	ROC Area	Macro ROC
×0.80	0.075	0.095	0.500	0.533
×0.85	0.042	0.190	0.666	0.680
×0.90	0.054	0.405	0.733	0.729
×0.95	0.096	0.819	0.919	0.923
×0.98	0.062	0.910	0.964	0.966
×1.00	0.079	0.943	0.973	0.973
×1.02	0.063	0.865	0.934	0.939
×1.05	0.065	0.283	0.677	0.739
band (held out)	0.133	0.900	0.954	0.951

It is tempting to stop at control-loop saturation, because the levels that fail are levels at which the stripper steam valve is on a stop. Our own control refutes it: the valve is equally on a stop at ×1.05, at the opposite end of its travel, yet the warm-started statistic there is only 3 to 6 times nominal and the ROC area is 0.847. Saturation marks where the plant stops being controllable, not where the monitor stops being informative.

The claim this section supports is correspondingly narrow, and what should not be read from it is a number: the setpoint at which informativeness goes depends on the plant, its loops and its in-control data, and is not a constant.

4.3. The Far Edge Is a Limit Cycle, and the Loss Survives It

Everything above is measured on episodes in which the setpoint multiplier is applied at initialisation, so the window under test sits inside the plant's response to the operating change. That leaves one reading the shift campaign cannot rule out: that the loss belongs to a prolonged load-change transient rather than to operation at a different setpoint. The distinction matters, because a plant that changes rate twice a day and a plant that runs for months at a lower rate are different customers for this result.

Section 3.3 describes the control: the plant is held at each setpoint for 48 h before anything is recorded, with calibration, deployment and window geometry otherwise identical. Table 6 places the two campaigns side by side, and the answer is not the same everywhere.

Table 6. The same map on a warm-started plant, against the step-transition campaign the rest of the paper uses; ROC area against counterfactual twins, averaged over the three detectors. The warm-started false-alarm and detection rates come from a threshold calibrated on independent fault-free commissioning runs at the same setpoint.

	ROC Area		Warm-Started Campaign		
Setpoint	Step	Warm-Started	False-Alarm	Detection	Macro ROC
×0.80	0.500	0.457	0.090	0.133	0.465
×0.85	0.666	0.791	0.033	0.600	0.793
×0.90	0.733	0.850	0.048	0.643	0.863
×0.95	0.919	0.959	0.029	0.876	0.951
×0.98	0.964	0.972	0.086	0.924	0.970
×1.00	0.973	0.972	0.062	0.919	0.966
×1.02	0.934	0.972	0.071	0.910	0.972
×1.05	0.677	0.847	0.090	0.586	0.839

At every level except the lowest, warm-starting recovers a large part of what the transitioning plant had lost: 0.666 to 0.791 at ×0.85, 0.733 to 0.850 at ×0.90, and most

strikingly 0.677 to 0.847 at ×1.05. On those levels the shift campaign was measuring the transition as much as the operating point. At ×0.80 it recovers nothing: the ROC area is 0.462, 0.454 and 0.454 for the three detectors, with 90% pair-preserving intervals of $[0.421, 0.501]$, $[0.413, 0.496]$ and $[0.415, 0.492]$, and paired win rates of 0.414, 0.400 and 0.429.

What the plant is actually doing at ×0.80. Two days at a setpoint does not guarantee a steady one, and it is not steady here. On fault-free runs of ten days the dominant period of the score series at ×0.80 reproduces at 21.3, 24.0 and 21.3 h across three independent runs, 74% of the in-band spectral power sits in the top percentile of frequencies, the autocorrelation of the six-hour unit statistic reaches 0.87 in magnitude, and the fault-free statistic has a within-run coefficient of variation (CV) of 0.42; at ×0.95 through ×1.02 none of this holds, the power share being 0.04 to 0.09 and the coefficient of variation 0.09 to 0.11. The plant here is on a slow limit cycle rather than at a displaced steady state, which is why we call the campaign *warm-started* and not settled. A six-hour window is a quarter of that cycle, so the score a fault has to move is already swinging by more than a fault moves it. Section 4.4 locates where the cycle begins.

The loss is a property of the operating state, not of the injection phase. A six-hour window is about a quarter of a one-day cycle, so one onset time cannot separate the two. We varied it, injecting the fault at 48, 54, 60 and 66 h, four points six hours apart across most of a cycle. Three things are fixed by design. The seeds are *shared* between onsets: one seed per (cluster, level, fault) serves all four and the twin carries the same seed, so the onsets differ in phase and nothing else. The threshold is re-derived *separately at each onset* from the same 64 commissioning runs windowed there, so each phase is judged against fault-free data at its own phase. Detector, window and fault set are unchanged, and ×1.00, where there is no cycle, is the control. The campaign contributes 828 episodes and no trip; Table A7 reports the sweep. At ×0.80 the ROC area stays between 0.437 and 0.563 across all twelve detector–onset cells, with a mean of 0.499; no onset recovers informativeness, and the spread across phases (0.075 to 0.087 per detector) is of the same order as the sampling spread at ×1.00 (0.016 to 0.024), where the ROC area never leaves $[0.944, 0.987]$. The loss at ×0.80 is a property of the operating state.

Two readings follow. First, the operational envelope is wider than the step-transition map alone suggests: a plant two days into a new rate is in better shape than Table 5 implies. Second, what remains at the far edge is a displaced state and not a transition, since it survives two days at the setpoint and injection at any phase of the cycle. That is not a claim that operating-point change produces such a loss in general. Only the displaced state is a reason to stop trusting the alarm's meaning, and Section 4.4 is what separates the two.

4.4. Resolving the Loss: A Transition Region, Not a Boundary

Eight levels bracket the loss. Twenty-six resolve the bracket to a single one-percent increment. Twenty-one levels placed *inside* that increment resolve the increment itself, and in doing so withdraw the description the coarser grid invited: what looks like a step at ×0.01 is a ramp at ×0.001. We report the sequence in that order because the lesson is not specific to this plant. A crossing located by interpolating across a grid step is a property of the step until something is simulated inside it, and the width of an interval quoted on such a crossing is not evidence about where the crossing lies.

This section walks ×0.80 to ×1.05 at ×0.01 (twenty-six warm-started levels, 5304 episodes, no trips) and then ×0.815 to ×0.835 at ×0.001 (twenty-one levels, 4284 episodes, no trips: every level retains all 64 commissioning episodes and all 70 counterfactual pairs), with the calibration pool, the window, the pairing and the detectors unchanged throughout. The fine campaign draws seeds from a base disjoint from every other campaign, and refuses to run if any seed it would generate already appears in an existing manifest, so its two levels that

coincide with the dense grid are an independent reproduction rather than a consistency check on the same episodes.

Validity does not move along the axis. Across all 141 detector–level cells of the two campaigns, the independently calibrated rate at the 0.10 target runs from 0.000 to 0.229, mean 0.090, on a calibration pool of 64 and 70 deployment pairs per cell (Figure 3a). That spread is what a correctly calibrated rule produces at this pool size and not a sign that calibration is slipping: with 64 calibration scores the split-conformal threshold is the 59th order statistic, the exact marginal rate is $6/65 = 0.092$, and a single cell's realised rate has a standard deviation of 0.05 across calibration and deployment draws. Simulating 141 exchangeable cells at those sizes, neither extreme is unusual: 77% of null draws reach a minimum as low as the observed one and 86% a maximum as high. The dense campaign alone (78 cells, 0.029 to 0.186) is, if anything, narrower than the null, as the sharing of episodes among the three detectors at a level would make it. Nor is there a trend in the operating point: the rank correlation between level and realised rate is +0.11 on the dense axis and −0.31 on the fine one, of opposite sign, on cells that are not independent within a level.

Informativeness is a ramp. On the dense grid the detector-averaged ROC area is 0.529 at ×0.82 and 0.791 at ×0.83, and there is no simulated point between them. Inside that increment the rise is gradual (Table 7): eleven levels carry it in steps averaging 0.026, and it is present in each detector separately, PCA-SPE 0.545 → 0.833, kNN-5 0.523 → 0.778, autoencoder 0.522 → 0.767. PCA-SPE, the most sensitive of the three, increases strictly at every one of the eleven levels. Neighbouring levels share a plant state and are not exchangeable, so no significance level is attached to that ordering; the scatter comparison that follows is the evidence for the trend.

Table 7. The eleven levels inside the increment the dense sweep could not resolve. p is the share of fault-free commissioning episodes the frozen classification rule assigns to the unsteady mode; CV is the across-episode coefficient of variation of the fault-free unit statistic across the commissioning pool at that level. Table A9 in Appendix I carries all twenty-one fine levels.

Setpoint	PCA-SPE	kNN-5	AE	Mean	p	CV
×0.820	0.545	0.523	0.522	0.530	0.62	0.415
×0.821	0.557	0.539	0.536	0.544	0.66	0.383
×0.822	0.594	0.586	0.578	0.586	0.52	0.458
×0.823	0.602	0.595	0.589	0.595	0.41	0.518
×0.824	0.643	0.639	0.639	0.640	0.23	0.465
×0.825	0.713	0.700	0.687	0.700	0.24	0.448
×0.826	0.716	0.716	0.718	0.716	0.06	0.312
×0.827	0.748	0.750	0.737	0.745	0.04	0.166
×0.828	0.799	0.755	0.753	0.769	0.03	0.163
×0.829	0.832	0.740	0.742	0.772	0.01	0.072
×0.830	0.833	0.778	0.767	0.792	0.00	0.025

How much of the ramp is resolved, and how much is not. Above ×0.830 the curve is flat in expectation, and its level-to-level scatter there has a standard deviation of 0.022, against a mean ramp step of 0.026. The trend across the eleven levels is therefore unambiguous, at roughly twelve times that scatter, while no individual ×0.001 step is resolved by it, and the curve is not monotone when all three detectors are averaged. We report the ramp and decline to read its internal shape. The same arithmetic retires the interval this paper previously quoted on the crossing: a 90% interval of width ×0.003 computed by resampling pairs at two levels ×0.01 apart presumes the straight line drawn between them, and presuming it is what was at issue. What replaces it is coarser and means more: the crossing falls between two *simulated* levels, ×0.824 and ×0.825.

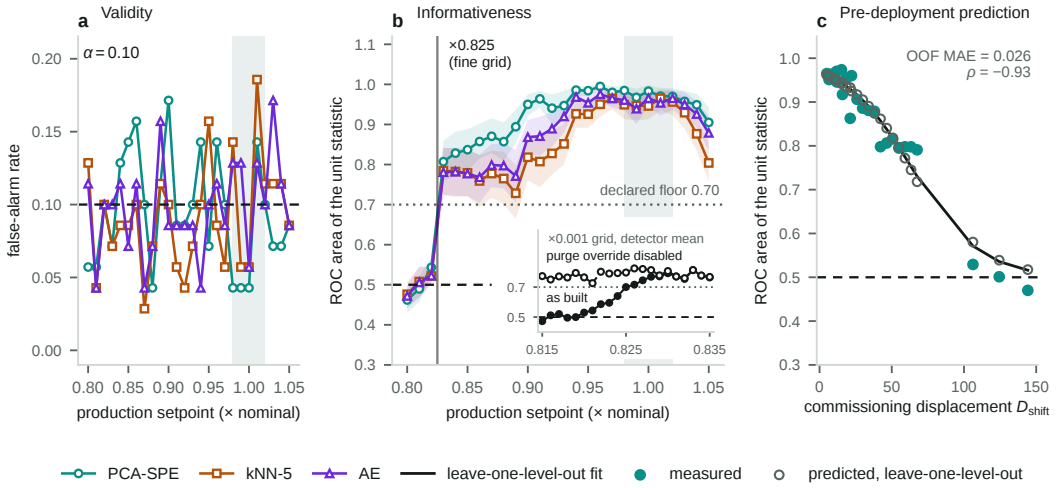


Figure 3. The validity–informativeness landscape on the dense warm-started axis. The pale shading in panels (**a**,**b**) marks the routine band, ×0.98 to ×1.02. (**a**) The independently calibrated false-alarm rate at each of the 26 levels. (**b**) The ROC area of the statistic that rate is attached to. The pale band around each curve is that detector's 90% pair-preserving bootstrap percentile interval over 1000 resamples. The vertical line is the crossing of the declared floor by the detector-averaged area, located on the fine grid between two simulated levels (Section 4.4); it is not the crossing of any individual curve. The inset is that fine grid, ×0.815 to ×0.835 in steps of ×0.001: the detector-averaged area of the plant as built (filled) and of the same plant with the low-pressure purge override disabled and every seed unchanged (open; Section 4.4.2). (**c**) The same ROC areas against a displacement computed from fault-free commissioning runs alone, each level predicted by a curve fitted on the other 25; the annotated error (OOF MAE, out-of-fold mean absolute error) is for the displacement term alone. Every level uses 64 calibration episodes and 70 counterfactual pairs.

The crossing, and what the floor and the window contribute to it. We define the *boundary* explicitly, as before and before any number is quoted: it is the setpoint at which the point estimate of the detector-averaged ROC area, linearly interpolated between adjacent levels, last crosses a declared floor of 0.70 descending towards the low edge. The floor is stipulated rather than derived, and it is 0.70 rather than the chance value 0.5 so that it marks where the ranking stops being usable rather than where it stops existing. On the fine grid the floor is crossed between ×0.824 (0.640) and ×0.825 (0.700), interpolating to ×0.8250. The dense grid's interpolation across its own single step gave ×0.8265.

The floor is a convention, and only the location depends on it. Validity at every level, the ramp, the mode decomposition of Section 4.4.1 and the intervention are stated without a floor; the floor turns the ramp into one number that can be reported next to a certificate. What it supplies to that number is now measured rather than argued. Floors of 0.55, 0.60, 0.65, 0.70 and 0.75 place the crossing at ×0.8211, ×0.8231, ×0.8242, ×0.8250 and ×0.8272: a span of ×0.006 over the tested range, on simulated levels throughout. On the dense grid every floor in that range fell inside one increment, so the stability it reported there was an artefact of interpolating within a single step; the fine grid reaches the same answer for a reason. A floor of 0.80 is a different matter and must be read off the dense axis, where it lands at ×0.890 because the curve carries a plateau at 0.79 to 0.82 from ×0.83 to ×0.89 that sits just under it. The fine window ends at ×0.835 and so contains none of that plateau; a crossing of 0.80 computed inside it is an artefact of the window and is not reported.

The scoring window is part of the definition rather than a nuisance parameter, and Table 8 says so quantitatively. At the declared six-hour window the crossing is ×0.8250; at four hours it is ×0.8271; at two hours there is no crossing at all, because the detector-averaged area never exceeds 0.647 anywhere on the fine axis. The persistence length, by contrast, is immaterial: at six hours, requiring the score to hold for one, three or five consecutive

samples moves the crossing by less than ×0.0003. A six-hour window is a quarter of the cycle the plant runs when it is unsteady, which is why it was declared; the two-hour row is the empirical reason the declaration matters.

The fine campaign reproduces the dense one where they meet. Levels ×0.820 and ×0.830 are walked again on fresh seeds. The detector-averaged ROC areas agree to 0.0008 and 0.0014 respectively (0.5292 against 0.5300, 0.7909 against 0.7923). Two levels is a narrow basis, and we report it as such; what it establishes is that the ramp is not an artefact of the seed stream that produced the dense step.

Table 8. What the scoring geometry contributes to the located crossing, on the fine axis. The declared pair is six hours with a persistence of three, shown in bold. A two-hour window places the whole curve below the floor, so no crossing exists to locate.

Window	Persistence	Crossing	Bracket	Lowest Area	Highest Area
2 h	1	none	no crossing	0.540	0.647
2 h	3	none	no crossing	0.537	0.636
2 h	5	none	no crossing	0.541	0.646
4 h	1	×0.8270	0.827–0.828	0.529	0.751
4 h	3	×0.8271	0.827–0.828	0.530	0.739
4 h	5	×0.8273	0.827–0.828	0.531	0.741
6 h	1	×0.8251	0.825–0.826	0.480	0.785
6 h	**3**	**×0.8250**	**0.824–0.825**	0.472	0.764
6 h	5	×0.8252	0.825–0.826	0.483	0.767

4.4.1. What the Plant Is Doing There: A Bistable Region

The ramp asks a different question of the plant than a step did. A step invites a change of regime at a point; a ramp over eleven levels invites a quantity that varies smoothly across them, and there is one.

The classification rule, declared on data frozen before the fine sweep. A fault-free episode is called *unsteady* when its unit statistic exceeds a cut fixed, per detector, as the geometric mean of the median commissioning statistic at ×0.84 and at ×0.81, one level clear of the transition on each side, so that the rule is a property of the two modes and not of anything measured between them. The cut is computed on the dense campaign, which was complete before the fine levels existed, and applied unchanged. For a counterfactual pair the mode is read off the *twin*: the twin carries the same seed, the same operating point and the same noise realisation with the fault switched off, so it reports which mode the episode occupied without being contaminated by the fault it is the control for.

The share of episodes in the unsteady mode falls across the whole region. It is 1.00 at ×0.80, 0.89 at ×0.81, 0.91 at ×0.815, and then declines to 0.62 at ×0.820, 0.41 at ×0.823, 0.23 at ×0.824, 0.06 at ×0.826 and 0.00 at ×0.830 and above (Table 7). At a pool of 64 the binomial standard error on such a share is about 0.06, which is the scale of the non-monotonicity within it, and of the difference between the dense and fine estimates at ×0.82 (0.72 against 0.62). What the region is, then, is not a boundary the plant crosses but a span of setpoints at which the same plant, warm-started and held for two days, settles into one of two modes with a probability that moves with the setpoint.

The fault-free variability says the same thing from the other side. If the plant simply became more oscillatory as the setpoint fell, the across-episode coefficient of variation would rise monotonically towards ×0.80. It does not: it is 0.025 at ×0.830, rises to a maximum of 0.518 at ×0.823, where $p = 0.41$, and falls again to 0.215 at ×0.815 and 0.202 at ×0.80. A variance that peaks where two modes are equally likely and is smaller within either of them is what a mixture produces, and it is visible in the dense campaign's own characterisation

(Table 9), where ×0.82 carries both the largest spread across independent runs in the table, 14.4 h against 0.9 h or less everywhere else, and the largest within-run coefficient of variation, 0.753, larger than at ×0.81 or ×0.80, where the limit cycle is clean. Three runs per level could not separate a mixture from an intermediate state; twenty-one levels with sixty-four episodes each can.

Table 9. Dynamic characteristics of the fault-free plant across the transition region, from 240-h runs after a 48-h warm-up, three independent runs per level. The ×0.82 row is the signature of the bistable region rather than of an intermediate state: it carries the largest spread across runs in the table and the largest within-run coefficient of variation (the fluctuation of the statistic along one run, not the across-episode CV of Table 7), exceeding both the clean limit cycle below it and the steady operation above it. Three runs per level cannot resolve that, and Section 4.4.1 measures it on twenty-one levels instead. The table reports where the two changes fall on the setpoint axis; it does not establish that one causes the other. ACF, autocorrelation function.

Setpoint	Dominant Period (h)	Spread Across Runs	Top-Decile Power	Max |ACF|	Within-Run CV
×0.80	24.0	0.0	0.84	0.88	0.448
×0.81	24.0	0.0	0.82	0.86	0.452
×0.82	33.6	14.4	0.52	0.63	0.753
×0.83	4.4	0.0	0.39	0.33	0.029
×0.84	4.5	0.3	0.39	0.41	0.029
×0.90	4.5	0.1	0.40	0.61	0.060
×1.00	1.0	0.9	0.05	0.33	0.104
×1.05	4.4	0.2	0.53	0.76	0.269

Inside the unsteady mode the fault is invisible to the detectors. Splitting the counterfactual pairs at each level by the mode their twin occupies, and computing the ROC area within each part, separates what the mode does from what the fault does. Write p for the share of pairs in the unsteady mode, A_u for the area within it and A_s for the area within the steady mode (Table 10). On the ten levels with at least ten unsteady pairs, A_u averages 0.499 and never leaves 0.458 to 0.542: inside the unsteady mode the fault does not move the detectors' score far enough to rank, at any setpoint. It is invisible to these scores, not absent from the measurements: supervised models trained on labelled faults in the settled state rank it at the far edge (Section 4.5). This is the quantitative form of what the warm-started campaign observed at ×0.80 alone, that a six-hour window is a quarter of the cycle and the score a fault has to move is already swinging by more than a fault moves it (Section 4.3). Inside the steady mode the fault is visible, at 0.706 on average where $0.15 \leq p \leq 0.75$ against 0.776 on the pure plateau above ×0.830: a modest degradation near the transition, on five levels. The share p read from the pairs and the share read from the commissioning pool (Table 7) are two independent estimates of the same quantity and agree within sampling error. Because the pairs are split on the same statistic whose ranking is then measured, the split could manufacture part of this; it does not. Fewer than 5% of faulty episodes fall on the other side of the cut from their own twin, and excluding the pairs whose twin lies within a factor of 1.5 or 2 of the cut (40% and 76% of them) leaves A_u at 0.481 and 0.475 and A_s at 0.804 and 0.755.

Why informativeness falls with the square of the steady share. A pooled ROC area compares every faulty episode with every fault-free one, and those comparisons split by the mode each member of the comparison is in. Write p_F and p_N for the unsteady shares among faulty and fault-free episodes, and A_{xy} for the ROC area over the comparisons between a faulty episode in mode x and a fault-free one in mode y ($x, y \in \{s, u\}$); A_{us}, for instance, is the probability that an unsteady faulty episode outscores a steady fault-free one.

Lemma 1 (Exact decomposition)**.** *For any partition of the episodes into two modes,*

$$A \;=\; (1-p_F)(1-p_N)\,A_{ss} + (1-p_F)\,p_N\,A_{su} + p_F\,(1-p_N)\,A_{us} + p_F\,p_N\,A_{uu}, \tag{7}$$

and the identity holds exactly for the empirical Mann–Whitney areas and the empirical shares, at any sample size.

Proof. The Mann–Whitney statistic is a sum over the $n_F n_N$ faulty–fault-free comparisons; grouping the terms by the modes of the two members of each comparison gives the four weighted averages. □

Proposition 2 (Mixing law, with its error)**.** *Let $p_F = p_N = p$ and write $C = A_{us} + A_{su}$. Then*

$$A - \tfrac{1}{2} \;=\; (1-p)^2\big(A_{ss} - \tfrac{1}{2}\big) + p^2\big(A_{uu} - \tfrac{1}{2}\big) + p(1-p)\,\big(C-1\big), \tag{8}$$

and if $|A_{uu} - \frac{1}{2}| \le \delta$ and $|C-1| \le \varepsilon$,

$$\Big|A - \tfrac{1}{2} - (1-p)^2\big(A_{ss} - \tfrac{1}{2}\big)\Big| \;\le\; p^2\,\delta + p(1-p)\,\varepsilon. \tag{9}$$

Proof. Set $p_F = p_N = p$ in Equation (7), subtract $\frac{1}{2}$ using $(1-p)^2 + p^2 + 2p(1-p) = 1$, and bound the last two terms by the triangle inequality. □

The algebra is elementary and is not the claim; the claim is that both conditions hold in this plant, and both are measured. With the mode of a pair read off its twin, a faulty episode and its control carry one label, so $p_F = p_N$ by construction. A cross-mode comparison is then decided by the mode and not by the fault: $A_{us} \ge 0.995$ and $A_{su} \le 0.04$ at each of the sixteen mixed levels, so C lies between 1.000 and 1.040 (mean 1.008 on the fourteen levels of Table 10): the unsteady episode wins the comparison whether or not it carries the fault. That bounds ε by 0.04, and δ is the distance of $A_u = A_{uu}$ from one half, at most 0.042 on the ten well-populated levels. Under those conditions Proposition 2 reads, with $A_s = A_{ss}$ and an error of at most $p^2\delta + p(1-p)\varepsilon$,

$$A - \tfrac{1}{2} \;\approx\; (1-p)^2\,\big(A_s - \tfrac{1}{2}\big)\,: \tag{10}$$

the informativeness above chance at a level is the steady mode's, scaled by the *square* of the steady share, because only comparisons between two steady episodes carry information about the fault. With a single value of A_s, taken from the pure plateau and fitted to nothing in the transition, the law accounts for $R^2 = 0.967$ of the variation over the sixteen levels where the mode is mixed; with A_s measured at each level, 0.990.

The linear mixture $pA_u + (1-p)A_s$, which is what one would write first, is not the pooled area at all. Lemma 1 gives the difference exactly, $A - [\,pA_u + (1-p)A_s\,] = -p(1-p)\,(A_u + A_s - C)$, which is negative whenever the steady mode carries any information; measured with each level's own A_u and A_s, the linear form over-predicts at all sixteen mixed levels, by up to 0.056 and most in the middle of the region, where $p(1-p)$ is largest. The same quantity, $p(1-p)(A_u + A_s - C)$, is what scoring each episode against the fault-free distribution of its own mode would recover (the covariate-adjusted ROC area of Janes and Pepe [42], which a mode-keyed calibration estimates), and it is bounded by what the modes carry: it cannot restore information that is absent inside the unsteady mode, where A_u is one half. Across the fine axis the unsteady share is the strongest single correlate of informativeness, at Spearman −0.936, ahead of the paired effect size at +0.918.

Table 10. Informativeness inside each mode, on the fourteen fine levels where both modes are populated. p is the share of counterfactual pairs whose twin is in the unsteady mode; A_u and A_s are ROC areas computed within the unsteady and steady pairs; the last column is Equation (10) with the single plateau value $A_s = 0.776$. Below $p = 0.15$, A_u rests on fewer than ten pairs.

Setpoint	p	A_u	A_s	Pooled	Equation (10)
×0.815	0.89	0.458	0.896	0.472	0.504
×0.816	0.80	0.501	0.786	0.512	0.511
×0.817	0.83	0.515	0.840	0.521	0.508
×0.818	0.81	0.473	0.880	0.496	0.510
×0.819	0.79	0.482	0.727	0.500	0.512
×0.820	0.72	0.529	0.668	0.530	0.522
×0.821	0.56	0.500	0.714	0.544	0.553
×0.822	0.38	0.518	0.717	0.586	0.607
×0.823	0.33	0.467	0.723	0.595	0.623
×0.824	0.17	0.542	0.706	0.640	0.690
×0.825	0.10	0.660	0.749	0.700	0.724
×0.826	0.11	0.470	0.774	0.716	0.717
×0.827	0.07	0.636	0.784	0.745	0.738
×0.828	0.03	0.660	0.786	0.769	0.760

Equation (10) is also why a coarse grid sees a step. Where p moves from near one to near zero over a few tenths of a percent, $(1-p)^2$ moves faster still, and a grid whose spacing exceeds that span lands on one side or the other.

Where the loss goes. Measured from the plateau, the drop at each mixed level splits exactly into three parts: the share of episodes in the mode where the detectors cannot see the fault, $p\,(A_s^* - A_u)$; the degradation of the steady mode itself, $(1-p)\,(A_s^* - A_s)$; and the cross-mode term, $p(1-p)(A_u + A_s - 1)$, with $A_s^* = 0.776$ the plateau value. Where most episodes are unsteady ($p \geq 0.7$) the first carries 72 to 93% of the drop and the cross-mode term 12 to 20%. Where the modes are closest to even (p near 0.35), the first carries about half, the steady mode's degradation about a fifth and the cross-mode term about a quarter. Below $p = 0.15$ the drop is too small for its parts to be apportioned reliably.

One control loop saturates across the region, and it is the one that tracks the mode. The obvious alternative is that the plant runs out of control authority here, and two loops have to be told apart. The stripper steam valve `XMV(9)`, the loop Section 3.2 identifies as the first to run out of travel, is commanded below its lower limit at every fine level and in every one of the 1344 episodes, on the steady plateau above ×0.830 as much as on the unsteady side, so the valve sits on its stop throughout: its saturation is common to both modes and cannot be what separates them, which is the conclusion Section 4.3 reached from ×1.05. The purge valve `XMV(6)` is different. On the plateau it holds near 38% of its travel; across the transition its level mean falls to 0%, its lower stop, at ×0.817 and ×0.818, and follows $38.3\,(1-p)$ with $R^2 = 0.94$, where p is the unsteady share measured on a separate pool of episodes. Every other manipulated variable moves by less than six points of travel across the window, and the product split drifts smoothly from 58.5 to 57.8. A level mean cannot say whether every episode drifts down together or some sit at one position and the rest at another, so the episodes were recorded one by one, the first eight on exactly the seeds behind the level means, which they reproduce to the fourth decimal. The purge position is bimodal: of 336 fault-free episodes on the fine levels, 112 sit within five points of the stop and 207 between 30 and 45% of travel, and only 17 fall in between. Shares compared across levels cannot settle whether the two criteria pick the same episodes, since any two quantities that change across the same span correlate. Each of the 1344 commissioning episodes behind p was therefore re-simulated from its seed, its purge valve recorded, and the two classifications compared episode by episode. They agree on 97.2% of episodes, Cohen's $\kappa = 0.94$, and κ stays between 0.94 and 0.95 whether "on the stop" is taken as within 0.5% or 5% of travel,

for a quarter or for three quarters of the window. The disagreements are neither symmetric nor scattered: 423 of the 424 episodes the detectors call unsteady have the purge on its stop, while 36 episodes have it on its stop for most of the window yet score just below the cut, and the disagreements lie a median factor of 1.1 from the cut against 1.6 for the agreements, where the score cannot decide. The purge valve on its stop is, in other words, very nearly what the detector statistic calls the unsteady mode, and a slightly more inclusive marker of it. The unsteady mode is the state in which a second loop has lost its authority: a plant that settles into it has its purge valve on a stop.

The same decomposition, with each episode's own plant state. Reading the mode off the twin's score is convenient but couples the label to the statistic being ranked, and it gives the faulty episode its control's label rather than its own. The purge marker just described removes both objections: each faulty and each fault-free episode is labelled unsteady if its own purge valve sits on a stop for more than half of its own scoring window, read from the simulator's own-state record and involving no detector at all. With these labels the four terms of Lemma 1 are measured separately. On the mixed fine levels ×0.815 to ×0.825, A_{uu} lies between 0.44 and 0.53, A_{us} between 0.991 and 1.000 and A_{su} below 0.035, while A_{ss} is 0.71 to 0.91; the shares agree, $|p_F - p_N| \leq 0.043$ on every mixed level. The plant-side label of the twins agrees with the score-side one on 97.3% of episodes, Cohen's $\kappa = 0.938$. The fault does not move the plant between modes: the share of faulty episodes whose own mode differs between the six hours before the onset and the six hours after is matched by their twins, which carry no fault, to within 0.043 at every level; up to 17% of twins change mode between the same two windows, so in the bistable region the plant switches modes spontaneously, fault or no fault, and the fault adds nothing measurable to it.

That does not turn the result into an artefact of saturation. It says what the unsteady state physically is. The calibration key records the setpoint; whether the purge loop saturates at that setpoint is a property of the state the plant settles into, and that is the quantity informativeness follows, through Equation (10).

4.4.2. Removing the Override Removes the Mode, and the Loss with It

An association with a loop on its stop leaves the direction open: the override could hold the plant in the unsteady mode, or the oscillation could drive the purge onto its stop. We settled it by intervention. The plant was rebuilt with one change: the low-pressure branch of the purge override in controller 6, which shuts the purge when the separator pressure falls to 2300 kPa, was made unreachable, so the purge stays under its ordinary loop; nothing else in the plant, the controllers, the seeds, the faults, the twins, the commissioning pools or the frozen detectors and cut was changed. All twenty-one fine levels and seven dense levels were re-simulated, 5712 episodes, each with an exact counterpart in the original campaigns. Four outcomes were declared before the campaign ran, and Table 11 reports them.

Control. Where the original plant never engages the override, at ×0.90 and ×1.00, every unit statistic of every episode is bitwise identical in the two builds. At ×0.83 and ×0.84 the override engages once in the original plant, briefly, in the start-up transient about 1.4 h after the setpoint change, and never in the scoring window; there the detector-averaged ROC area moves by at most 0.015 and stays inside the original 90% interval at all eight such levels. The declared control had named ×0.83 and above as untouched; a probe run before the campaign found the start-up engagement, and the control was split into these two parts before any campaign episode was simulated. *Mode.* Without the override the unsteady mode does not occur. None of the 1792 commissioning episodes has its purge on a stop for most of its scoring window, against up to 89% of them with the override in place; by the frozen score cut, two of the 64 episodes at ×0.80 exceed it, for one detector and by one percent, and none at any other level. The fault-free coefficient of variation stays

between 0.018 and 0.026 at every level from ×0.80 to ×0.84, against a peak of 0.52 with the override, and the fault-free statistic still rises monotonically as the setpoint falls, without the step: displacement continues, the mixture does not. *Informativeness*. Across ×0.815 to ×0.825 the detector-averaged ROC area rises from 0.554 to 0.775, and at ×0.80 from 0.470 to 0.759 (0.703–0.812), the value of the steady plateau; at no level of either grid does it fall below 0.726. Because both builds are scored on the same seventy pairs, the effect itself has a paired interval: resampling pairs and evaluating both builds on each draw (5000 draws), the gain at ×0.80 is 0.289 (0.222–0.359, 95%). Its interval excludes zero at every level up to ×0.828, where the unsteady share of Table A9 is still 0.03 or more, and contains zero at every level from ×0.829 upward, where the share is 0.01 or less. Read across levels, the gain tracks the mode share. These are twenty-eight separate 95% intervals, not adjusted for multiplicity, and they are reported as estimates rather than as a family of tests. *Safety*. No episode trips in either build.

Table 11. The intervention: the same plant, seeds, faults, twins and detectors, with the low-pressure purge override disabled. *Unsteady share*: commissioning episodes above the frozen score cut (detector mean). *Purge on a stop*: commissioning episodes whose own purge valve sits at 0 or 100% for more than half of the scoring window. *Built*: the plant as distributed; *disabled*: the intervention. 64 commissioning episodes and 70 counterfactual pairs per level; no trips in either build. Table A13 gives all twenty-eight levels.

		Unsteady Share		Purge on a Stop		ROC Area, Detector Mean		
Setpoint	**Grid**	**Built**	**Disabled**	**Built**	**Disabled**	**Built**	**Disabled**	**90%, Disabled**
×0.800	dense	1.00	0.01	0.89	0.00	0.470	0.759	[0.703, 0.812]
×0.810	dense	0.89	0.00	0.86	0.00	0.501	0.801	[0.747, 0.855]
×0.815	fine	0.91	0.00	0.89	0.00	0.472	0.773	[0.718, 0.828]
×0.818	fine	0.86	0.00	0.88	0.00	0.496	0.764	[0.706, 0.816]
×0.820	fine	0.62	0.00	0.66	0.00	0.530	0.765	[0.709, 0.821]
×0.822	fine	0.52	0.00	0.55	0.00	0.586	0.803	[0.754, 0.852]
×0.824	fine	0.23	0.00	0.23	0.00	0.640	0.794	[0.738, 0.846]
×0.825	fine	0.24	0.00	0.30	0.00	0.700	0.797	[0.736, 0.850]
×0.827	fine	0.04	0.00	0.03	0.00	0.745	0.821	[0.772, 0.869]
×0.830	fine	0.00	0.00	0.00	0.00	0.792	0.807	[0.752, 0.859]
×0.835	fine	0.00	0.00	0.00	0.00	0.764	0.768	[0.715, 0.821]
×0.840	dense	0.00	0.00	0.00	0.00	0.798	0.800	[0.752, 0.848]
×0.900	dense	0.00	0.00	0.00	0.00	0.879	0.879	[0.837, 0.919]
×1.000	dense	0.00	0.00	0.00	0.00	0.965	0.965	[0.945, 0.983]

The direction is therefore the one the trace suggested: the override, engaged by low separator pressure, holds the plant in the unsteady mode, and removing it removes the mode and the loss together. Two lines of evidence point at the mode: inside the unmodified plant the steady-mode episodes at the same setpoint are still ranked (Section 4.4.1), and the intervention restores the pooled ranking. Neither isolates the mode as the causal mediator, because the intervention changes a control law as well as the state it holds the plant in. What it establishes is the effect of this one controller modification in this simulator: disabling the branch removes the override-held mode and the far-edge loss together. Two further limits belong with the result. It does not make the override a design error: it exists to protect the separator, and a plant without it is not one we recommend. And it does not remove the decline of informativeness with displacement: without the override the ROC area still falls from 0.965 at ×1.00 to 0.879 at ×0.90 and to about 0.78 at the far edge, the gradual decline the curve of Section 4.4.3 tracks; what disappears is the collapse to chance. For the distinction this paper is built on, the intervention is the cleanest evidence it contains. The two builds keep identical setpoints, eligibility keys, seeds, faults, counterfactual pairs and frozen detectors. Disabling the protective branch changes the control law and, with it, the distribution of operating states the plant settles into, and the ranking changes with

them; the operating point alone cannot account for the difference, since it is the same in both builds.

Three things this does not establish. *We do not claim a mathematical bifurcation, or a bifurcation mechanism.* The observed behaviour is reported as an empirical region of bistability, evidenced by a mode share that varies with the setpoint, a fault-free variance that peaks where the two modes are equally likely, and a period reproduced across independent runs within the unsteady mode; no equilibrium continuation, no eigenvalue crossing and no bifurcation diagram is computed anywhere in this work. *The region is not a physical limit of the plant.* It is a detector-performance transition under the declared ROC criterion, on this process and this detector family; the plant runs perfectly well below it. *And bistability was not among the alternatives this study set out to exclude.* Table 2 lists a transition towards the setpoint and an artefact of the injection phase, and the warm-started design excludes both: a plant held two days at the setpoint and injected at any phase is not in transit. It is not thereby in a single displaced state, which is what that section concluded. A plant held two days at the setpoint can settle into either of two modes, and the control that would have caught it is the one reported here. The transferable claim is the procedure, and the resolution discipline it needs, rather than the value ×0.825.

4.4.3. Predicting the Loss from Fault-Free Data

Corollary 1 rules out reading the loss off fault-free data directly. It does not rule out predicting it, and the dense axis is enough points to try. Two quantities are available to a plant before any fault is seen, both from the same commissioning pool that calibrates the threshold. The first is how far the operating point has moved the measurements: D_{shift}, the 2-Wasserstein distance between Gaussian fits of the standardised measurement distribution at the deployment level and over the commissioned range, with the level under test held out of the reference whenever it falls inside that range. The second is how unsteady the plant is once it gets there: the coefficient of variation of the fault-free unit statistic across the pool. We call this the *across-episode* CV, the spread of one six-hour unit statistic over the 64 independent commissioning episodes at a level, averaged over the three detectors; it is the quantity reported in Tables 7 and A8 and the only one that enters the fit. It is distinct from the *within-run* CV of Table 9, which measures how much the statistic fluctuates along a single long run; the two answer different questions and are not expected to agree.

Estimating a model's performance on unlabelled data under distribution shift is an established problem, and the construction used here belongs to it. Performance has been regressed on dataset-level feature statistics without labels [43], estimated from the confidence a classifier assigns to unlabelled target data [8], and benchmarked across those families, where distributional distances were found to be weaker predictors than confidence-based ones [10]. We claim neither the idea of predicting performance from unlabelled data nor the use of a distribution distance to do it. What differs is the target and the setting: a one-class monitor emits an uncalibrated score rather than a posterior, so the confidence-based estimators do not transfer unmodified; the quantity predicted is a ROC area, not a classification accuracy; the second regressor is a fault-free variability term motivated by an identified dynamical regime rather than fitted post hoc; and both regressors are measured during commissioning, before any fault has been seen, so neither uses a fault, a label, a twin or a threshold. The curve that maps them to a ROC area is a different matter: it is fitted on levels where faults are labelled, here in the simulator, and the Discussion states what that requires on a plant without fault data.

D_{shift} carries most of the variation on its own, at Spearman -0.93 ($p \approx 10^{-11}$) over the twenty-six levels. Fitted as a logistic curve floored at 0.5, it predicts a held-out level's ROC area with a mean absolute error of 0.026 against 0.106 for the level-average null. Leave-

one-level-out is a weak test here, because neighbouring levels are ×0.01 apart and nearly determine each other, so we also report the harder direction: fitting on the thirteen levels nearest the commissioned range and extrapolating to the thirteen most displaced gives 0.054 against 0.196 for the corresponding null. That is the direction a plant would use the curve in.

What the displacement reads inside the transition. Across the fine window D_{shift} falls from 118 at ×0.815 to 65 at ×0.835, its largest single step 10.4; the dense grid's step of 38.7 between ×0.82 and ×0.83, ten times its neighbours, was the sampling. Computed from the same fault-free commissioning pool as the mode share of Section 4.4.1, it is very nearly a linear function of that share, $D_{\text{shift}} = 66.4 + 55.9\,p$ with $R^2 = 0.997$, and it follows the share's own sampling fluctuations from one level to the next, rising with it between ×0.817 and ×0.818 and between ×0.824 and ×0.825. The same line reproduces the dense value at ×0.82, where that campaign's pool happened to be 72% unsteady (106.9 predicted, 106.2 published), and the fine value at ×0.820, where the fine pool was 62% unsteady (101.1 predicted, 101.8 measured). Inside the transition, then, the displacement, which needs no labels, is a commissioning-time measurement of how often the plant settles into the unsteady mode, the quantity on which, by Equation (10), pooled informativeness there depends.

The dense curve, asked about levels it has never seen. Fitted on the twenty-six dense levels only, the curve was asked about the twenty-one fine levels, simulated on seeds it never saw. It predicts them with a mean absolute error of 0.041 from displacement alone and 0.033 with the variability term, against 0.192 for the dense-level mean, no worse than the extrapolation column above. It places the ramp where it is, but at 58% of its height: it under-predicts the plateau above ×0.830 by about 0.04 and over-predicts at the unsteady end by 0.03 to 0.07. That is the expected failure of a logistic curve in D_{shift} fitted across the whole operating range, where the displacement mostly measures distance, applied inside a region where informativeness depends on the mode share quadratically. One caution belongs with the variability term. It was introduced for the high edge, where CV is elevated by the plant's own fluctuation; inside the transition CV is elevated for a different reason, the variance of a two-mode mixture, and peaks at ×0.823 where the modes are closest to even. The same regressor measures two different things at the two ends of the range, and its contribution should be read with that in mind.

The metric is not sufficient, and the way it fails is informative. Equal displacements on the two sides of the range do not cost equally: ×1.05 sits at $D_{\text{shift}} = 21.2$ with a ROC area of 0.863, while ×0.94 sits at 22.2 with 0.960. The displacement-only curve over-predicts the high edge by up to 0.067. Adding the fault-free variability removes it. The across-episode CV that enters the fit is elevated at exactly that edge, 0.424 at ×1.05 against 0.092 at nominal (Table A8), and the within-run CV of Table 9, measured on separate long runs, shows the same elevation, 0.269 against 0.104: the largest high-side residual falls to 0.010, the out-of-fold error to 0.022 and the extrapolation error to 0.041. The improvement is modest and the extra parameter is one more on twenty-six points, so the case for it rests on the out-of-fold and extrapolation columns and on the fact that the boundary diagnostic predicted the term before it was fitted, not on the in-sample gain.

What this supports is narrow. On this plant and this detector family, the informativeness a certificate sits on is predictable before deployment from fault-free commissioning data, once the curve has been fitted on levels where faults are labelled, to about 0.04 ROC area when extrapolating into the displaced half of the range. It is a calibration curve for one process, not a law; a second plant would need its own. What transfers is the procedure (measure the two quantities during commissioning, fit the curve where labels exist, read the envelope off it) and the finding that the quantity to watch is not the operating point but the displacement it produces and the steadiness of what it produces.

4.5. Capacity Does Not Restore the Ranking; Supervision Does Only Where the State Is Covered

One question sits behind both controls in this section, and it is the question a referee asks first: is the loss an artefact of a weak one-class model? Two variables answer it (how much capacity the detector carries, and whether it was allowed to see faults at all) and neither changes the outcome.

Figure 4 puts ten detector variants through the map: three principal-component statistics, a nearest-neighbour score in two variants, a shallow autoencoder, and LSTM, GRU, convolutional and Transformer temporal autoencoders.

Inside the band all ten are competent and close together, with mean ROC area between 0.934 and 0.978, obtained at between one and two orders of magnitude of difference in fitting time: under 1.7 s for every classical detector, 13.5 to 30.8 s for the deep ones. Outside the band all ten fall, to between 0.650 and 0.756 on average, and every one of them reaches chance on the fault mixture at its worst operating level, the worst-level ROC area ranging from 0.490 to 0.527. All ten score distance from a fitted in-control manifold, so the control establishes that capacity is not the binding constraint within that family rather than a law about detectors in general.

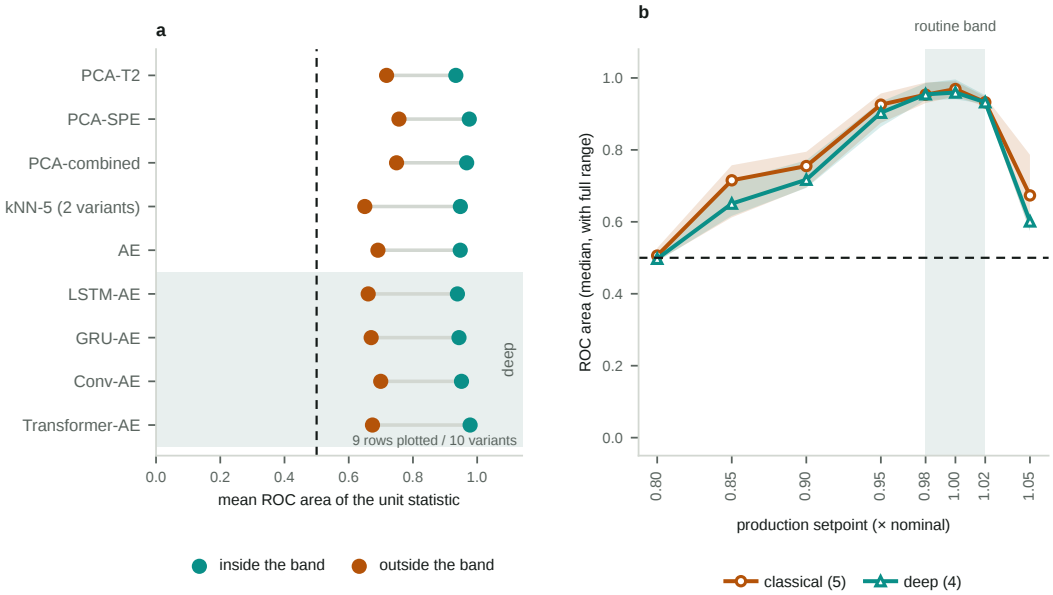


Figure 4. Nine of the ten one-class variants: *k*NN-5 (squared) is held back because it ranks identically to *k*NN-5 and would weight that family twice. Table A4 gives all ten. (**a**) Mean ROC area for each detector over the 3 levels inside the routine band and the 5 outside it; the four deep architectures are bracketed at the left. (**b**) Median ROC area against operating level, with the band spanning the full range across the variants of each family at that level, five classical and four deep, and the pale shading marking the routine band. The dashed lines mark chance ranking, an ROC area of 0.5. Each level rests on 70 counterfactual pairs. PCA-T2, Hotelling's T^2 of the principal component model; LSTM, long short-term memory; GRU, gated recurrent unit; Conv, convolutional.

The deep architectures were each fitted from three independent initialisations, so the comparison can be read against its own noise. Across the twelve fits, in-band ROC area occupies $[0.938, 0.978]$ and out-of-band ROC area occupies $[0.645, 0.707]$: the ranges do not overlap, and the best out-of-band run of any architecture and seed sits below the worst in-band one. The seed spread within an architecture (standard deviation between 0.003 and 0.008 out of band) is small enough that the shared pattern is not an artefact of one lucky initialisation.

Throughout, the re-calibrated false-alarm rate holds across all detector–level cells: capacity moves the bottom row of Figure 2 and leaves the top row alone.

All ten detectors of Section 3.4 share a training principle as well as a family, so capacity alone cannot test whether the boundary belongs to the one-class paradigm. Section 3.4 describes the four models that were given faults instead: an L_2-regularised logistic regression, a random forest, gradient-boosted trees and a multilayer perceptron, all fitted on labelled faults from the same band training episodes and all evaluated on the warm-started campaign against the same counterfactual twins.

Trained on the band, none of the four evaluated supervised classifiers rescues the far edge, and we state the scope plainly: this is a result about those four models under these training regimes, not about supervised learning. The random forest reaches 0.522 there (90% interval 0.481–0.563), which is chance, against 0.937 at the nominal rate; the perceptron reaches 0.531 (0.479–0.581), and gradient boosting, the strongest of the four in the band at a cross-validated 0.989, falls below chance to 0.421 (0.361–0.476). The logistic regression does better and still loses most of what it had: 0.620 (0.578–0.664) against 0.956. Between ×0.85 and ×0.95 the four stay close to their in-band values; the loss is concentrated at the far edge, and at ×1.05 they bracket the one-class average rather than beating it (0.634 to 0.861 against 0.847).

Labelled faults at the displaced setpoints change the far edge only when they come from the state the plant is in. Trained on labelled windows from all eight levels of the shift campaign, which reach each setpoint by a step, the four families rank well across the middle of the range (the logistic regression at 0.985 to 0.988 from ×0.85 to ×0.95, against 0.922 to 0.932 when trained on the band), so that training set does carry what it was given; at ×0.80 they reach 0.496 to 0.576, the logistic regression 0.533 (0.501–0.569). Trained instead on the calibration-role episodes of the warm-started campaign itself, at all eight levels, which is the geometry of the evaluation on clusters and seeds disjoint from it, they rank the far edge: 0.921 (0.885–0.956) for the logistic regression, 0.949 (0.922–0.975) for the random forest, 0.974 (0.951–0.996) for gradient boosting and 0.848 (0.797–0.894) for the perceptron, with paired win rates of 0.94 to 1.00, and 0.97 to 1.00 at every other level. Two things follow. The fault is not absent from the measurements in the unsteady mode: it is invisible to a score that measures distance from a fault-free manifold, because the mode's own fault-free variation swamps it (Section 4.4.1), and a discriminant fitted on labelled faults in that mode finds a direction the oscillation does not occupy. And what restores informativeness is coverage of the operating *state*, not of the operating point: the step-trained and the warm-started-trained models saw faults at the same setpoints, and only those that saw them in the state the plant occupies recover the ranking. This is the supervised counterpart of the coverage result of Section 4.6.3, and it carries the practical cost that result does not: labelled faults in the displaced state, which a plant rarely has before it reaches that state. Table A11 gives all twelve fits level by level.

The paired win rate separates two things the ROC area combines, and the separation is the most useful thing this control produces. At ×0.80 the supervised logistic still ranks a faulty episode above *its own* counterfactual twin in 80.0% of pairs (the fault is still visible when the comparison is against the same run without it), while its ROC area against the pooled fault-free population is the 0.620 above. The one-class detectors lose both, at 0.414 paired and 0.457 pooled. What band-trained supervision buys, then, is within-pair signal; what it does not buy is a usable pooled ranking, because a deployed alarm has no counterfactual and must compare against a population whose spread at that operating state is the 42% variation of Section 4.3. The warm-started all-levels models above show what does: labelled faults seen in that state. The displacement damages the ranking through the fault-free arm's variability at least as much as through the fault's own signature.

What the band-trained controls hold fixed is what the models were shown: all are fitted on the same band episodes and see the same operating range through the same

standardisation, so between them they vary capacity and training principle while holding the training distribution constant. Their failure therefore does not separate displacement from extrapolation in general, and the warm-started all-levels control above shows that the same models, given labelled faults in the displaced state, do much better; for the one-class arm that variable is moved by Section 4.6.3. What the band-trained controls settle is the narrower question they were built for: more capacity is not enough, and having seen faults and been optimised to rank them is not enough either, unless the faults were seen in the state being monitored. Table A4 gives each detector's three ROC areas, Figure 4b plots the descent level by level, and Tables A10 and A11 give the supervised classifiers at the same levels.

4.6. What Repairs the Loss, and What Does Not

Two questions close the sequence of interventions, and they have different answers. Repositioning a threshold cannot repair the loss, for a structural reason; coverage of the operating range in the training data widens the range over which the score stays useful.

4.6.1. Repositioning a Threshold Does Not Repair the Loss

The reason none of the calibration policies can do better on the quantity that matters is structural rather than empirical. Conditioning, reweighting and adaptation as used here all reposition a threshold on a *fixed* score, and the area under the ROC curve is invariant to any repositioning of a threshold, so every policy compared in this paper shares one informativeness curve: the curve of Figure 2b, which runs from 0.499 to 0.959 for the autoencoder. Interventions that change the score itself, such as an operating-condition-dependent statistic, local modelling or retraining, are not covered by that argument, and the coverage comparison below measures what one of them is worth.

Adaptive conformal inference belongs to that first class, and it is the strongest available scheme for holding a target rate without exchangeability, so what it does when asked to hold a rate it cannot honestly earn is worth reporting. The two implementations of (4) part company, and the difference is the finding. The *bounded* rule never recovers at any step size or level tested, plateauing at a post-change alarm rate of 0.207, 0.337 and 0.501 for $\gamma = 0.005$, 0.02 and 0.05 and staying there four months later; the plateau belongs to the clip, not the plant. The *unbounded* form does reach a final-twenty rate of exactly 0.100, the convergence result holds, but only by falling silent, spending 60% to 86% of the horizon at an infinite threshold and firing on the rest regardless of content, which under 20% fault prevalence costs detection, falling from 0.425 to 0.266 as γ rises. Eligibility conditioning reaches comparable marginal control (sustained recovery at every level, post-change rate 0.109) with a threshold inside the range of the data it is applied to. Neither recovers the ranking; the point is that a rate held by silence and a rate held by a threshold inside the data look identical in the one number a plant usually watches. Appendix C gives the plateau arithmetic, the recovery times, both prevalence variants and Figure A2.

4.6.2. FiLM Conditioning Did Not Help the Model We Evaluated

Table A5 answers the constructive question of Section 3.10. Trained on the band, RC-AE averages 0.701 outside it against 0.700 for the same backbone without conditioning: the difference is +0.002, and the worst level is 0.504 against 0.507. There is no reason to expect otherwise, since FiLM can interpolate between manifolds the model has seen and cannot supply one it has not, so handing the operating point to a model trained at a single band of operating points gives it nothing to interpolate with. Nor does conditioning earn its place once coverage exists: with the wide training set RC-AE reaches 0.887 against 0.905 for the unconditioned backbone, no better and somewhat worse.

4.6.3. Coverage Is the Lever, and It Is Not a Sample-Size Effect

Holding the training set at the same twelve episodes and spreading them across the levels instead of concentrating them in the band lifts the out-of-band mean from 0.701 to 0.884, and the worst level from 0.504 to 0.751. Enlarging the covered set to 96 episodes adds almost nothing further for the same model, to 0.887: the gain is coverage of the operating range, not the number of episodes. The wide set covers exactly the levels the map evaluates, so what this establishes is that showing a model its deployment operating levels helps, not that a continuously covered range generalises to levels between the ones seen.

Two limits bound the result. These three training conditions are not intended as a statistical comparison of training strategies; with one fit each they isolate the qualitative effect of extending the commissioning range under an otherwise identical modelling procedure, and that is the weight the recommendation carries. And even the best configuration does not recover the in-band level anywhere outside the band: its out-of-band mean is 0.905 against 0.951 to 0.954 in band, and its worst level is 0.871. Coverage widens the envelope over which a certificate carries information; it does not abolish it. The eligibility-conditioned false-alarm rate stays at target for every model and level throughout. Figure A4 plots the three training conditions against the operating level.

4.7. How Far the Observation Reaches

Three questions close the section: whether the explanation of an alarm survives the alarm, whether any of this survives outside simulation, and how much of the fragility reported above belongs to one design decision of our own.

4.7.1. The Explanation Inherits the Alarm's Envelope

If an alarm cannot be trusted at a given operating point, neither can its explanation. The reference and the attribution are not the same object: τ_j is a causal effect on the process and Δ_j a contribution to a model score, so τ is a causal process-effect reference and not a ground-truth attribution, and a variable with a large physical effect need not dominate any particular detector's score. Attribution scored against that reference is informative in band (the nearest-neighbour detector reaches a top-3 precision of 0.670 against a chance level of 0.058) and its ordering is not the ordering of detection quality, since the squared prediction error has one of the highest in-band ROC areas and the worst attribution at 0.380. Along the shift axis the two baselines part company, and the gap is the result. Against the twin baseline, top-3 precision falls from 0.608 at the nominal rate to 0.248 at ×0.80, in step with the ROC area of Figure 2b. Against the baseline a plant could actually use, the episode's own recent history, it falls further, from 0.593 to 0.097, because away from the band that history is itself moving and contaminates the comparison before the attribution begins. The explanation inherits the alarm's envelope. Appendix D gives the setpoint table, the per-detector and per-fault breakdown in Table A3, and Figure A3.

4.7.2. Calibration Ageing on a Real Testbed

The pattern reproduces on the testbed (Figure 5), and one result appears that simulation cannot produce. At high flow, at the same operating point as the reference session but on a different day, the frozen calibration already yields a false-alarm rate of 0.667 to 0.714. Session-to-session drift alone is enough to void the guarantee, without any change of operating point at all. This is the comparison Section 3.11 identified as the one the dataset supports cleanly, since both arms sit at the same flow level, and it is the strongest single argument in this paper against trusting a calibration to age well.

Across flow levels the frozen rate is 1.000 and eligibility conditioning returns it to 0.095 at high flow and 0.146 at low flow. That comparison confounds the operating point

with the campaign, so we report it as consistent with the simulation result rather than as independent evidence for it.

Informativeness varies by regime on real data as well: ROC area 0.952 to 0.960 at high flow, 0.900 to 0.940 at mid flow, and 0.668 to 0.677 at low flow, the last with intervals excluding 0.5 but well short of the in-band simulation values. The ranges are narrow for a reason worth stating: on this eight-sensor rig the three detectors' unit statistics agree at Spearman $\rho \approx 0.999$ and flag exactly the same windows, so they are three near-identical measurements rather than a spread. This comparison therefore carries little evidence about detector-specific differences, and what it does separate is the calibration policy.

The mid-flow regime supplies a cost of conditioning the simulator could not reveal, because in simulation one can always generate more episodes. Only five calibration units exist there, so the smallest certifiable level is $1/6 \approx 0.167$ and $\alpha = 0.10$ cannot be certified at all; the implementation returns an infinite threshold rather than a finite one that would misstate the guarantee. The finer the eligibility key, the more cells starve.

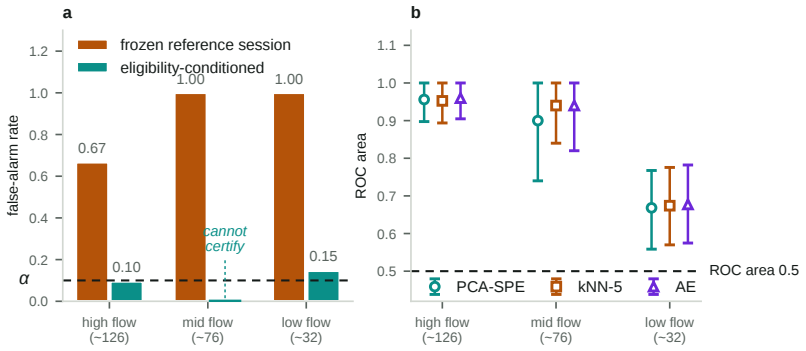


Figure 5. Additional evidence: calibration ageing on a public industrial anomaly dataset. This evaluation addresses calibration ageing rather than operating-point displacement, and does not replicate the boundary result. (**a**) False-alarm rate under a calibration taken from the anomaly-free reference session, by flow regime; the mid-flow bar is absent because $\alpha = 0.10$ cannot be certified there at all. (**b**) Informativeness by regime, with the bars giving the 90% bootstrap percentile interval over 1000 resamples. The sample sizes are small and uneven, and they carry the caveat: high flow has 21 calibration, 21 fault-free and 13 faulty units, mid flow 5, 10 and 5, and low flow 42, 41 and 30. The parenthesised numbers on the axis are the pump's volume flow rate as the dataset records it, which is the eligibility key here, not a sample size.

4.7.3. A Single-Setpoint Ablation

One ablation closes the section. Fitting the plant model at a single setpoint rather than across the routine band makes every off-nominal level break the frozen threshold, including ×0.98, where the false-alarm rate reaches 1.000 for each of the three headline detectors (Setting A, Section 4.1). The qualitative conclusions are unchanged but the apparent fragility is exaggerated, which is why the band design is the one reported above. The width of a study's in-control class is therefore worth reading alongside any brittleness result of this kind, including this one.

5. Discussion

The implication of this study is not that conformal monitoring fails, but that the statistical validity of an alarm rule and its ability to rank faults are separate properties of one alarm. A guarantee can remain satisfied after conditioning on an observable operating point while the alarm loses that ranking, because the guarantee addresses the false-alarm behaviour of the fault-free score distribution and does not by itself constrain fault separation.

A deployment change does not only move the calibration threshold; it can change what the score underneath it represents.

Because the same transition appears across heterogeneous one-class and supervised models, it cannot be attributed to one detector family, and the intervention of Section 4.4.2 places it in the plant: with the detectors, the seeds and the calibration unchanged and one override disabled, the ranking returns. It is not thereby a detector-independent law. All fourteen models are fitted on the same standardised measurements and may share sensitivity to the oscillation the plant enters, and the supervised models trained on labelled faults in the settled state rank them inside the unsteady mode where the one-class detectors do not (Section 4.5); what is lost in that state is the monitors' ranking, not the information in the measurements.

What the boundary analysis transfers is not its location. The numerical value is specific to the Tennessee Eastman process under the Russell–Chiang–Braatz control scheme and to the displacement path evaluated here; nothing in this work suggests that ×0.825 transfers. What transfers is the procedure: define informativeness independently of validity, locate where the ranking degrades against a declared floor on a grid fine enough to place the crossing between simulated levels rather than across an interpolated step, and examine whether the degradation coincides with a measurable change in the operating state. The fine sweep is the reason for the middle clause: on a one-percent grid the loss looked like a step at a boundary, and only levels simulated inside that step showed a ramp across a bistable region. The pre-deployment estimate transfers the same way: fit the curve while labels still exist; the curve itself, and the error it achieved here, do not.

A plant with no fault data at all can still use two of the three parts. The certificate needs none: every threshold in ECAC is set on fault-free commissioning runs. The displacement D_{shift} needs none either, since it is computed from fault-free runs alone, so a monitor can report, before any fault has occurred, that deployment has moved beyond the states it was commissioned on. What it cannot report without labels is how much ranking that displacement costs: the curve from displacement to ROC area is specific to a plant and has to be fitted where faults can be labelled, in a process model of that plant, from historical fault records, or from faults deliberately introduced during commissioning tests. Until it has been, the honest output at a displaced but calibrated operating point is the certificate together with a qualification: the false-alarm guarantee holds there, and the ability to detect is unverified. This is the counterpart of the *uncertified* status that Algorithm A1 reports for a key with no calibration units, where the guarantee itself is absent.

That a regime-blind threshold fails outside its calibration regime is unsurprising. What the campaign adds is that nothing on the monitor side announces the second failure, and Proposition 1 says why no diagnostic built on the fault-free calibration law can. Reporting a threshold-free informativeness measure alongside any conformal guarantee costs nothing and would prevent the confusion. Trustworthiness here is less a matter of model interpretability than of three things: the validity of a stated false-alarm rate, its reliability away from the commissioning conditions, and the integrity of a calibration as it ages. An assurance case therefore has to carry the operating envelope over which alarms were shown to be informative, not only the guarantee they were issued under.

Eligibility conditioning is a modest device (Mondrian conformal prediction [1,4] with the taxonomy supplied by the control system instead of learned) and its modesty is the point: coverage conditional on a declared, observable key is attainable where coverage conditional on an arbitrary covariate value is not. The SKAB check found the limit of the idea in the opposite direction: a key fine enough to be honest about the operating condition can be too fine to be certifiable, because each cell must carry its own calibration units. Key granularity is a budget, not a free parameter.

The same scope applies to explanations: an explanation cannot be more informative than the signal it is derived from, and explanations therefore inherit the operating envelope of the alarms they explain. That does not invalidate explainability methods; it places explanation quality downstream of the reliability of the monitored signal.

The practical implication is a commissioning consideration rather than a replacement for detector development: in this campaign, extending the operating range represented in the fault-free commissioning data helped more than the architectural changes evaluated here, which is enough to treat operating-range coverage as a commissioning decision and not enough to make range coverage a general rule that dominates model design. ECAC as we would deploy it, then: condition on an operating key the control system already logs; cover the anticipated levels rather than the operating point alone; and treat a trigger as an instruction to re-establish exchangeability rather than as an alarm, which places the monitor with the admission-control layers used to govern learned controllers at runtime [34]. Where exchangeability cannot be re-established, declare the certificate void rather than re-calibrate and continue.

5.1. Limitations

Several limitations define the scope of the conclusions, and the first three are the ones that bound every operating-point claim in the paper.

5.1.1. One Plant, One Control Structure, in Simulation

Every result on operating-point shift comes from the Tennessee Eastman process, under one control structure (the Russell–Chiang–Braatz closed-loop implementation), simulated. The bistable region, its location, the purge override that tracks its unsteady mode and the square law that connects the mode share to the ROC area are properties of that plant under that control structure; another plant, or the same plant under another control structure, may lose informativeness at a different displacement, through a different change of state, or not within its operating range at all. What is claimed to transfer is the procedure and the identity of Lemma 1, which holds for any two-mode partition, not the values. The intervention of Section 4.4.2 establishes that, in this plant, one protective override is what holds the plant in the mode where the ranking is lost; a plant whose protective logic differs may lose it through another loop, or not at all.

5.1.2. Confirmation Is Simulation-Based

The campaigns added in this revision repeat the measurement on seeds disjoint from every earlier campaign and reproduce it where the grids meet; that is independent simulation-based confirmation within the same plant model, not external validation. No real plant was walked across its operating range.

5.1.3. The Location Is a Property of the Declared Criterion

The crossing at ×0.825 is a detector-performance transition under a declared floor of 0.70 and a six-hour scoring window, not a physical limit. Section 4.4 measures what each convention contributes: floors from 0.55 to 0.75 move it by ×0.006, a four-hour window by ×0.002, and a two-hour window removes the crossing altogether. Its precision is that of the grid, between two simulated levels ×0.001 apart; the ×0.003 bootstrap interval of the original submission, computed across a single interpolated step, is withdrawn, and the statistical uncertainty of the campaign is in any case not the precision with which a real plant's transition would be known.

The real-data study does not replicate the operating-point result, since SKAB's operating levels are confounded with its experiment campaigns, so the only claim taken from

it is the single-level one (that a calibration does not transfer across sessions at the same flow), which is about calibration ageing rather than operating-point shift. Nothing here is a field study: there is no live deployment, no operator in the loop and no plant that ran this layer in service, so the paper's standing is that of an industrially motivated simulation with a public real-data check, and its recommendations are hypotheses for a plant trial rather than conclusions from one. Counterfactual twins exist only in simulation; on real data the attribution reference would have to come from process knowledge, with the weaknesses that entails. The operating range is bounded on both sides by control-loop saturation rather than by anything statistical, so results at ×0.80 and ×1.05 describe a plant with a valve on a stop as much as one at an unusual production rate. That saturation is not by itself what removes the ranking: the stripper steam valve is on its stop on the steady plateau above ×0.830 as well, where the ranking holds at 0.75 to 0.81, and the loop whose saturation tracks the loss is the purge override (Section 4.4.2). No equilibrium continuation or eigenvalue analysis is performed, so the bistable region is an empirical finding and not a characterised bifurcation. The supervised recovery of Section 4.5 needs labelled faults in the displaced, settled state, and how few would suffice was not measured. The eligibility key is one-dimensional, and multi-key conditioning would sharpen the calibration-budget problem the SKAB mid-flow regime already exhibits. The evaluated detector set represents important classes of monitoring model but does not cover all AI architectures. The comparison of commissioning strategies is limited by one fit per condition and is intended to show a qualitative effect rather than to compare training procedures statistically. Finally, the shift studied is a change of operating point; sensor degradation, fouling and catalyst ageing are different mechanisms and would deserve their own map. Future work on industrial data should test whether the same separation between validity and informativeness appears under other process dynamics, control structures, sensor configurations and deployment scenarios.

6. Conclusions

Calibration validity and fault-ranking informativeness are separable properties of a monitoring system. Conditioning a distribution-free alarm rule on an operating key the control system already knows buys the first at every evaluated level, in simulation and consistently with that on a real testbed, while regime-blind and pooled calibration fail badly and in different ways. It does not buy the second everywhere: at the far edge of the operating range the monitored statistic loses its relationship to the fault across ten one-class detector variants and four supervised classifiers trained on faults in the band, while the certificate over it holds exactly as it does at the nominal rate. That is not a failure of calibration. It is a limit on what a calibrated score preserves outside the operating conditions its in-control data represented.

Walking the range on a plant held at the setpoint for two days rather than moving towards it locates where the limit falls, and resolving the walk changes what the limit is. In one-percent steps the ranking appears to collapse in a single increment; in steps of ×0.001 it is a ramp, crossing the declared floor of 0.70 at ×0.825, across a region where the same plant settles into either a steady mode or an unsteady one with its purge valve on a stop. Inside the unsteady mode the detectors cannot rank the fault, and the pooled ranking above chance follows the square of the steady share, exactly as the decomposition of the ROC area requires. That crossing is a detector-performance transition under the declared criterion and not a physical operating limit; the plant runs below it perfectly well. Disabling the purge override, with every seed unchanged, removes the unsteady mode and restores the ranking, so informativeness here goes with the state the plant occupies, which the calibration key does not record. The intervention establishes the effect of one controller modification on mode and ranking together; it does not isolate the mode as the

unique mediator. The override is this plant's instance of that; another plant would supply its own, and nothing in the certificate would announce it.

None of it can be read off the alarm. A conformal certificate constrains the fault-free score law and nothing else, so no monitoring diagnostic relying only on that law identifies the loss without an additional assumption linking the two. It has to be predicted instead, and it can be: two quantities available during commissioning, mapped by a curve fitted where faults are labelled, estimate a level's ROC area to a mean absolute error of 0.041 extrapolated into the displaced half, averaged over held-out levels rather than for any single prediction.

Neither obvious repair works, and the reason is the same in both cases: repositioning a threshold cannot recover a ranking, and an architecture cannot invent an operating state it was never shown, which is why supervised models recover the far edge only when trained on labelled faults in the state the plant occupies there. Among the interventions that need no labelled faults, extending the operating range represented in the in-control data produced the largest improvement in this campaign, and even that does not close the envelope. The numerical boundary is campaign-specific; the transferable contribution is the procedure: map the envelope during commissioning while labels exist, and report it alongside the guarantee rather than the guarantee alone.

7. Reproducibility

The simulation and analysis code, the frozen configuration, the campaign manifests, the per-episode unit statistics and the figure sources are kept as one archive. A verifier runs 95 mechanical checks and refuses to write the freeze file if any of them fails. They cover the campaign shape, the seed and split bookkeeping, the trip accounting behind the conditional estimand, the exactness of each attribution decomposition, the independence of the displacement metric from anything it is used to predict, and the scope of the claims the text is allowed to make. A second script re-derives every number printed in this article from the released artefacts. A complete rebuild runs on two CPU cores and requires no graphics processing unit (GPU).

One limit of that offer is worth stating rather than leaving for a reader to discover. Everything the simulator and the classical detectors produce is seed-determined and reproduces to the last printed digit, campaign manifests included. The figures obtained from the four deep detectors and from the regime-conditioned model do not, because training accumulates floating-point operations in an order that depends on the arithmetic library and the core count. What this article reports of those models is, the worst-level column excepted, an average over operating levels, and averaging damps that: such an average should be read as reproducible to about 0.01 rather than exactly. A worst-level entry is not an average but an order statistic, so it inherits a single level's deviation instead of damping it, and should be read to about 0.03, as should any single level. The differences this paper's arguments turn on are 0.19 and 0.20 in ROC area, two orders of magnitude larger, so no comparison made here rests on that precision.

One correction belongs here because it concerns what "seed-determined" means. In preparing this revision the simulator was audited for isolation between episodes simulated by the same process, and the Fortran core was found to carry five quantities from one episode into the next: the state flag of the separator-pressure override in controller 6, which the main program never initialises, and the four temperatures that the energy balance solves by Newton iteration from their current value, which the initialisation routine never sets. An episode's trajectory therefore depended, usually in the last digits, on which episodes the same process had simulated before it. The released code resets all five to the values a fresh process starts with before every episode, and carries the tests that establish isolation: every audited episode, after random predecessors, is bitwise equal to the same episode run

first in a fresh interpreter and to a byte-level restore of the simulator's memory, and every fault–twin pair is identical before the onset. Every campaign and analysis reported here was regenerated with the corrected simulator on the original seeds. No episode changed between healthy and tripped, the largest change in any reported ROC area is 0.035, and no conclusion changed.

Author Contributions: Conceptualization, N.S. and Y.I.C.; methodology, N.S. and Y.I.C.; software, N.S.; validation, N.S. and Y.I.C.; formal analysis, N.S. and Y.I.C.; investigation, N.S. and Y.I.C.; resources, Y.I.C.; data curation, N.S.; writing—original draft preparation, N.S.; writing—review and editing, N.S. and Y.I.C.; visualization, N.S.; supervision, Y.I.C.; project administration, Y.I.C.; funding acquisition, Y.I.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Korea Evaluation Institute of Industrial Technology (KEIT) grant funded by the Ministry of Trade, Industry and Energy (MOTIE) in 2025 (Research and Development of Artificial Intelligence Standardization for the Healthcare Industry Based on Human Big Data and Artificial Intelligence, Project No. RS-2025-13882968), and by the Korean Agency for Technology and Standards (KATS) under the project 'Development of International Standard Technologies based on AI Model Lightweighting Technologies' (Project No. RS-2022-KT222340).

Data Availability Statement: The SKAB dataset is publicly available at https://doi.org/10.34740/KAGGLE/DSV/1693952. The artefacts supporting this study are the simulation and analysis code; the frozen configuration together with `FROZEN.json`, which records a SHA-256 digest of every generated artefact; the campaign manifests, carrying one row and one distinct seed per episode, including the counterfactual twins and their parent episodes; the per-episode unit statistics from which every reported figure and table is computed; and the two scripts that check them, one verifying the campaign against its configuration and one re-deriving every number printed in this article from the released artefacts. These are held as a single archive and are available from the corresponding author upon reasonable request, including to referees and editors during review.

Acknowledgments: During the preparation of this manuscript, the authors used an artificial intelligence tool as an assistant in editing the text, in improving the structure of the article, and in refining the English scientific expression. The whole of the research design, the experimental campaign, the analysis of the data, the interpretation of the results and the scientific conclusions were carried out and verified independently by the authors. The authors take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. The Calibration-Policy Comparison

Table 4 gives the mean over the five classical detectors of what each calibration policy does to the false-alarm rate in each regime. Figure A1 adds what a mean cannot show: the same nine cells with every detector plotted separately.

Intervals

Table A1 adds 95% intervals to every cell of Table 4 and to the paired differences between policies. They measure episode sampling, calibration included. A percentile bootstrap (5000 draws) redraws, within each regime and with replacement, the calibration episodes and the deployment episodes, re-derives every threshold from the redrawn calibration set, and scores all five detectors and all three policies on the same draw, so that every difference is paired; a deployment episode brings its fault-free unit and, if it is faulty, its faulty unit together. The training episodes and the fitted detectors are held fixed, so the interval is conditional on them; how much a cell depends on which detector is used is shown separately, by the open circles of Figure A1. Each regime contributes 64 calibration and 80 deployment episodes, 70 of them faulty (62, 79 and 69 at the high rate, after the trip exclusion). No contrast drawn in Section 4.1 is within these widths. Eligibility

conditioning lowers the off-nominal false-alarm rate against the regime-blind policy by 0.66 to 0.95, and every one of the four paired intervals excludes zero, the widest being −0.658 (−0.737 to −0.522) at the high rate with all-regime training. Against the pooled policy every off-nominal paired interval also excludes zero, with the sign Section 4.1 reports: pooling over-alarms at the low rate and under-alarms at the high rate. An earlier version of this table bootstrapped over the five detectors instead; that interval measures detector dependence, not sampling, and is not reported.

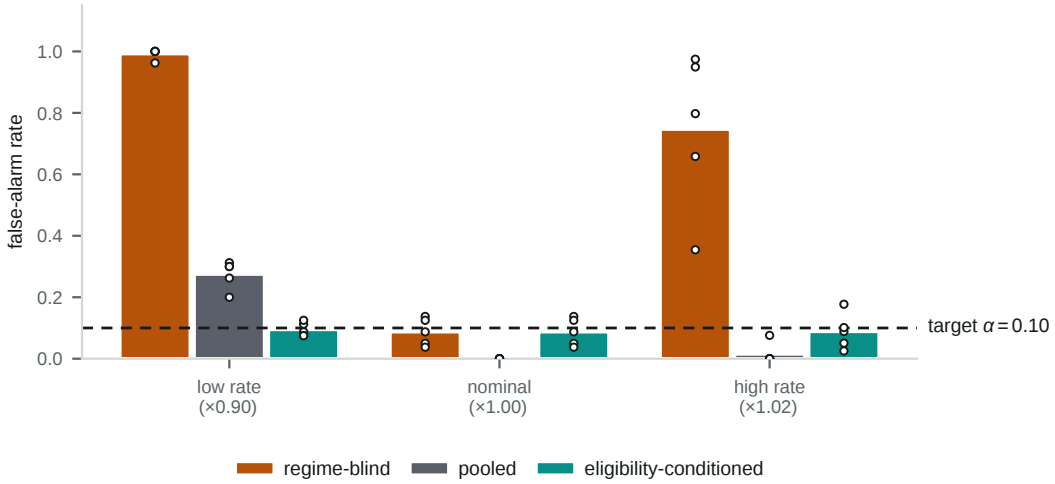


Figure A1. Three calibration policies at $\alpha = 0.10$ with the detector trained on all three regimes. Bars are the mean over the five classical detectors, which Table 4 gives to three decimals; each open circle is one of those detectors. At the high rate the five detectors scatter from 0.354 to 0.975, so the mean of 0.747 describes none of them.

Table A1. Table 4 with episode-level 95% intervals: mean over the five classical detectors, with a paired percentile bootstrap over calibration and deployment episodes in which every threshold is re-derived on each draw, and paired differences in false-alarm rate between policies on the same draws. An interval of [0.000, 0.000] or [1.000, 1.000] means that no draw produced an alarm (or a miss); it is the percentile bootstrap's degeneracy at the boundary, not certainty. For one detector at a fixed threshold, 0 alarms in 80 fault-free episodes has an exact two-sided 95% Clopper–Pearson upper bound of 0.045 (0.046 for 79), and 0 detections in 70 faulty episodes one of 0.051; 80 of 80 has a lower bound of 0.955. Main campaign, $\alpha = 0.10$; Low, Nominal and High are the three regimes.

Training	Calibration Policy	Low	Nominal	High
False-alarm rate				
Nominal only	regime-blind	1.000 [1.000, 1.000]	0.075 [0.028, 0.152]	1.000 [1.000, 1.000]
Nominal only	pooled	0.258 [0.162, 0.405]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
Nominal only	eligibility-cond.	0.055 [0.018, 0.155]	0.075 [0.028, 0.152]	0.132 [0.073, 0.213]
All regimes	regime-blind	0.992 [0.975, 1.000]	0.088 [0.040, 0.150]	0.747 [0.635, 0.830]
All regimes	pooled	0.275 [0.207, 0.370]	0.000 [0.000, 0.000]	0.015 [0.003, 0.030]
All regimes	eligibility-cond.	0.095 [0.052, 0.160]	0.088 [0.040, 0.150]	0.089 [0.053, 0.170]
Detection rate				
Nominal only	regime-blind	1.000 [1.000, 1.000]	1.000 [0.997, 1.000]	1.000 [1.000, 1.000]
Nominal only	pooled	0.329 [0.253, 0.423]	0.000 [0.000, 0.003]	0.003 [0.000, 0.009]
Nominal only	eligibility-cond.	0.286 [0.205, 0.375]	1.000 [0.997, 1.000]	0.594 [0.504, 0.681]
All regimes	regime-blind	1.000 [0.983, 1.000]	0.960 [0.909, 0.986]	0.939 [0.876, 0.978]
All regimes	pooled	0.606 [0.508, 0.706]	0.646 [0.538, 0.742]	0.623 [0.518, 0.723]
All regimes	eligibility-cond.	0.551 [0.451, 0.653]	0.960 [0.909, 0.986]	0.713 [0.612, 0.817]
Paired difference in false-alarm rate				
Nominal only	elig. − pooled	−0.202 [−0.305, −0.120]	+0.075 [+0.028, +0.152]	+0.132 [+0.073, +0.213]
Nominal only	elig. − regime-blind	−0.945 [−0.983, −0.845]	+0.000 [+0.000, +0.000]	−0.868 [−0.927, −0.787]
All regimes	elig. − pooled	−0.180 [−0.253, −0.125]	+0.088 [+0.040, +0.150]	+0.073 [+0.041, +0.149]
All regimes	elig. − regime-blind	−0.897 [−0.940, −0.828]	+0.000 [+0.000, +0.000]	−0.658 [−0.737, −0.522]

Appendix B. The Operating Envelope, Level by Level

Appendix B.1. Control Authority and Product Quality

Section 3.2 states that two physical properties single out the same operating interval. Table A2 gives them level by level. The stripper steam valve `XMV(9)` is the loop that runs out of travel first. Between ×0.80 and ×0.95 it sits between 0.1% and 2.4% of its travel, at ×1.05 on the opposite stop at 100%, and only within $[\times 0.98, \times 1.02]$ does it retain the 26% to 76% of range that leaves the stripper temperature cascade above it something to work with. Product G falls monotonically across the same axis, from 58.9% at the lowest rate to 52.9% at the highest, crossing its design value inside the same interval.

Appendix B.2. The Fault-Safe Envelope

Fault-free, the plant completes every episode at every level from ×0.80 to ×1.05: no disturbance-free run trips at any production rate reported here. Under fault the envelope is narrower and asymmetric. The whole low side, ×0.80 through ×1.00, survives all eight conditions in all seeds, while trips appear only above nominal: at ×1.02 on IDV(13) alone (2 of 18 seed clusters) and at ×1.05 on IDV(1) in all 18 clusters, IDV(13) in 11, IDV(8) in 10 and IDV(5) in 1. The fault-safe envelope is therefore a property of the plant *and* the fault.

Appendix B.3. What the Trip Conditioning Affects

The 49 excluded episodes are six in the main campaign, all on IDV(13); one in the band campaign, also on IDV(13); and 42 in the shift campaign. The exclusion is not uniform along the shift axis. No episode trips at any level at or below nominal (576 episodes at ×0.80 through ×0.95, none excluded), so the low end of the range, where the main text's central result is measured, is measured on a complete set. Trips occur only at ×1.02 (2 of 144) and ×1.05 (40 of 144, that is 27.8%), on IDV(1), IDV(5), IDV(8) and IDV(13). The ×1.05 column is the one place where the conditioning bites, and its numbers describe the faults the plant survived at that production rate.

Table A2. The operating range, characterised before any detector is applied. Fault-free episodes only; the valve position and product split are means over the post-transient part of the episode. XMV(n), the nth manipulated variable of the Tennessee Eastman model.

Setpoint	Stripper Steam Valve `XMV(9)`	Product G	Status
×0.80	2.3%	58.9%	saturated
×0.85	2.4%	57.3%	saturated
×0.90	0.1%	56.0%	saturated
×0.95	2.2%	54.7%	saturated
×0.98	26.3%	54.1%	all loops in range
×1.00	47.9%	53.7%	all loops in range
×1.02	75.5%	53.4%	all loops in range
×1.05	100.0%	52.9%	saturated

How far it can bite is bounded without assuming anything about the tripped runs. Every excluded episode is a faulty one; no fault-free twin trips. Against the same completed twins, the pooled ROC area with the tripped faulty episodes restored lies between the values obtained by ranking all of them below, or all of them above, every twin: $(nA+\{0,m\})/(n+m)$ for n completed and m tripped faulty episodes and an observed area A. At ×1.02 ($n = 69$, $m = 1$) the detector-averaged area 0.934 lies in $[0.920, 0.934]$, so the exclusion is immaterial there. At ×1.05 ($n = 46$, $m = 24$) the area 0.677 lies in $[0.445, 0.788]$: the reported value is conditional, but even its most favourable completion stays below the 0.934 at ×1.02 and the 0.973 at nominal, so the direction of the high-side decline does not depend on the exclusion, while its size does. A trip is itself an observable event, so in operation these episodes would

not pass unnoticed; the bound concerns the ranking only. The warm-started campaign, which carries the high-edge variability and the predictor of Section 4.4.3, has no trip at any level.

Appendix C. The Adaptive Conformal Comparison

Appendix C.1. Streams Rather than Episodes

Because the comparison concerns behaviour over time, it is run on continuous fault-free streams: twelve twenty-day runs at each of four operating levels, cut into consecutive six-hour windows. Half the runs at each level calibrate and half deploy, and calibration and deployment windows are drawn from the same construction throughout: a threshold fitted on windows taken just after start-up does not transfer to windows taken three weeks into a run. An operating change is modelled as a switch from a nominal run to a shifted one.

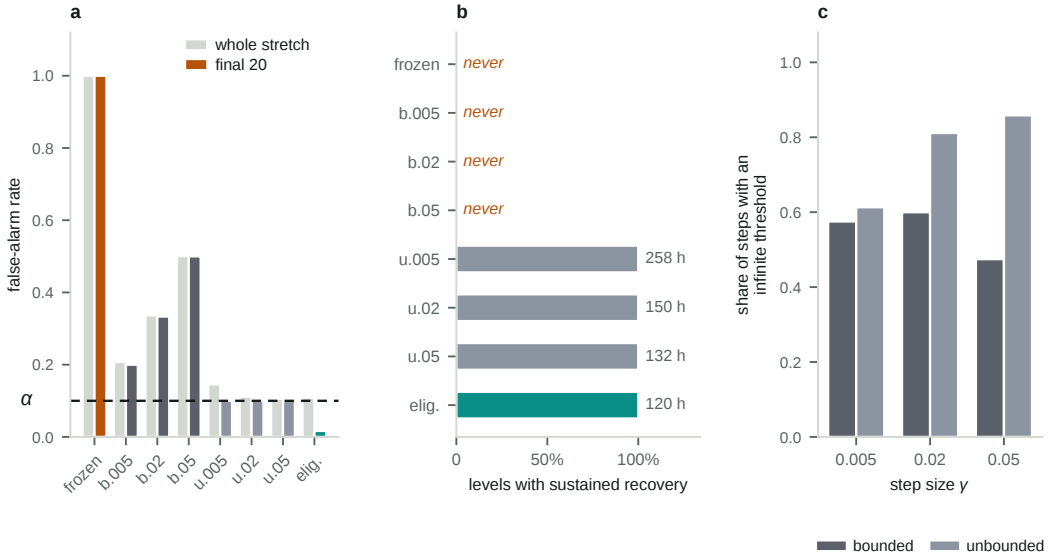


Figure A2. Calibration policies on a stream that changes operating level, over the four-month post-change horizon. Bars whose label begins *b* are the bounded implementation of Equation (4) and those beginning *u* the unbounded one, each at three step sizes. (**a**) Alarm rate over the whole stretch and over its final twenty windows; the bounded rule's two bars coincide, which is the plateau. (**b**) Share of levels reaching sustained recovery, with the median time; six hours per window. (**c**) Share of steps at which the adaptive threshold is infinite and no alarm is possible.

Appendix C.2. The Recovery Criterion and the Horizon

Recovery is judged as *sustained* recovery: the trailing twenty-window alarm rate must fall below 1.5α and stay below it for the next forty windows, with an excursion that cannot be confirmed within the remaining horizon counted as no recovery. A single crossing of the target by a shorter trailing window is not evidence that a rule has settled, which is why the criterion is stated this way; but it needs sixty windows of headroom, and a single post-change stretch is 79 windows, so it could only ever certify a recovery beginning in the first nineteen. The held-out runs at each level are therefore concatenated into one fault-free stretch at that setpoint (the seams break serial correlation but not the marginal distribution), giving the 474-window post-change horizon the main text reports.

Appendix C.3. Supplying the Error Indicator

Equation (4) needs err_t, which requires knowing whether an alarm was false. Two variants bracket what is available. The *oracle* variant updates only on genuinely fault-free windows and is not deployable. The *naive* variant counts every alarm as an error, which a

plant can run but which adapts to the faults themselves. The article reports both under 20% fault prevalence.

Appendix C.4. Injecting Faults into a Stream

The fault-prevalence variant needs faulty windows inside a stream of continuous-run statistics, and an episode statistic cannot simply be dropped in: the two are different constructions, and off the nominal setpoint the episode statistic also carries the setpoint step response. What transports is the fault's multiplicative effect on the score, measured against the episode's own counterfactual twin at the same setpoint and the same point of the transient. That ratio is applied to a stream window, which carries the stream's own scale.

Appendix C.5. Why the Bounded Rule Plateaus, in Full

With every post-change score above the whole calibration range, one alarm drives α_t below zero, the clip parks it there, and the recursion then climbs back at $\gamma\alpha$ per step, so the rule cycles with a period set by γ: larger γ gives a worse plateau, which is the opposite of the usual reading of that parameter. The final-twenty rates are 0.200, 0.333 and 0.500, the same numbers four months after the change as immediately following it.

Appendix C.6. Why No Finite Threshold Discriminates After the Change

Every fault-free window at ×0.90 and ×1.05 lies above the entire nominal calibration range, and 89% do at ×0.95. The unbounded rule's only remaining freedom is therefore *when* to be silent, which is what it uses.

Appendix C.7. The Oracle Variant

An *oracle* told which windows are genuinely fault-free, and therefore not deployable, runs at false-alarm rates of 0.262 to 0.119 with detection 0.545 to 0.365: better than the naive variant in level but no different in kind, since it too buys its rate with silence.

Appendix C.8. Recovery Times and the Prevalence Variant

The unbounded form recovers at every level and every step size, but slowly: sustained recovery arrives after 22 to 43 windows, five and a half to eleven days. Eligibility conditioning reaches sustained recovery at every level after 20 windows at ×0.95 and ×1.05 and 104 at ×0.90, with a post-change rate of 0.109 and a final-twenty rate of 0.017. Under 20% fault prevalence the *naive* rule reaches false-alarm rates of 0.178, 0.088 and 0.075 on fault-free windows as γ rises, which looks like success, while its detection rate falls from 0.425 to 0.266; the *oracle* variant runs at 0.262 to 0.119 with detection 0.545 to 0.365.

Appendix D. Attribution: Estimator Eligibility and the Detector Breakdown

Appendix D.1. Which Scores Admit a Decomposition

For the squared-error detectors the score is a sum of squared deviations, so the summand for variable j is its contribution and attribution is an exact decomposition. That holds for the squared prediction error, the autoencoder and the deep temporal models. It does not hold for two of the statistics used elsewhere in the main text: Hotelling's T^2 lives in score space and is not a sum over variables of the residual, and the conventional nearest-neighbour score is a mean Euclidean distance whose per-variable quantities sum to a mean *squared* distance instead. For the nearest-neighbour case the repair is free, since scoring by the mean squared distance makes the same quantities an exact decomposition, and both

variants are reported. T^2 and the combined statistic are excluded from the attribution study for this reason.

Appendix D.2. The In-Band Detector Breakdown

Against the deployable baseline the nearest-neighbour detector attributes best (top-3 precision 0.670, top-1 hit rate 0.833), ahead of the autoencoder (0.574, 0.699) and the squared prediction error (0.380, 0.464); against the twin baseline the same three reach 0.702, 0.600 and 0.405. This is not the ordering of detection quality.

Appendix D.3. The Causal Reference by Fault

The reference recovered from the twins is physically coherent, which is itself a check on the twin construction: IDV(1), the A/C feed ratio step, resolves to the A feed valve; IDV(4) and IDV(11), both acting on reactor cooling water inlet temperature, resolve to the reactor cooling water valve; IDV(5) resolves to the compressor recycle valve in 67% of in-band episodes; IDV(10) to stripper temperature; IDV(13) to separator pressure.

Table A3. Attribution against the causal reference, top-3 precision, averaged over the scored detectors. The deployable baseline differences against the episode's own pre-onset window; the twin baseline differences against the counterfactual twin. Chance is 3/52 = 0.058.

Setpoint	Deployable Baseline	Twin Baseline
×0.80	0.097	0.248
×0.85	0.122	0.382
×0.90	0.254	0.460
×0.95	0.489	0.606
×0.98	0.551	0.595
×1.00	0.593	0.608
×1.02	0.521	0.597
×1.05	0.298	0.433

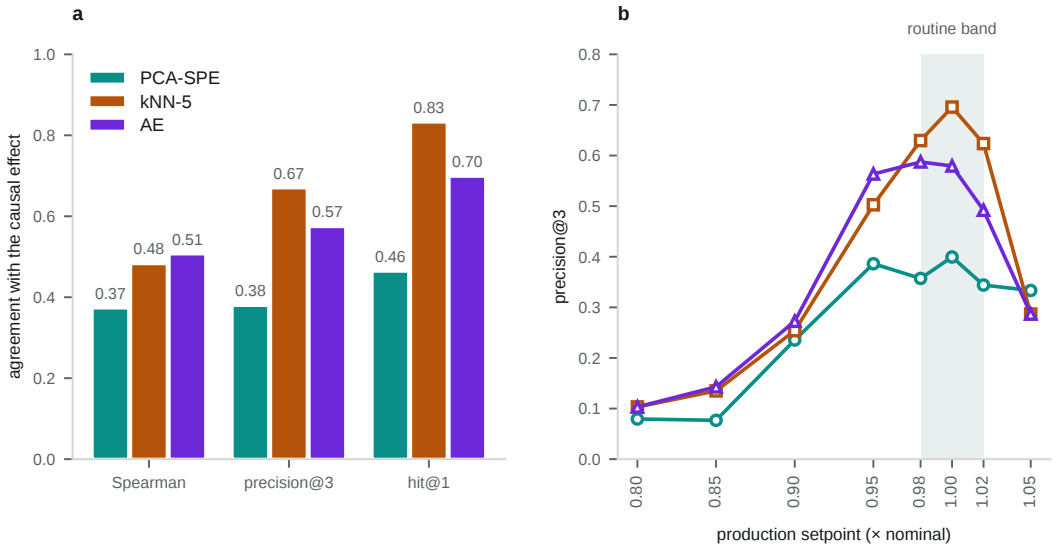


Figure A3. (**a**) Agreement between attribution and the causal effect measured on counterfactual twins, in band. (**b**) Top-3 precision along the shift axis; it follows the informativeness curve of Figure 2b.

Appendix E. The Detector Comparison, Level by Level

The table and figure below give the per-detector cells behind Section 4.5, which quotes their ranges, and the level-by-level view of the training conditions Sections 4.6.2 and 4.6.3 summarise.

Table A4. Capacity control. Mean ROC area inside and outside the routine band, and at each detector's worst operating level, all measured against counterfactual twins. Each of the ten detector variants reaches chance at its worst level, classical and deep alike.

Detector	In Band	Outside Band	Worst Level
PCA-T^2	0.934	0.718	0.506
PCA-SPE	0.975	0.756	0.510
PCA-combined	0.967	0.749	0.527
*k*NN-5	0.948	0.650	0.491
*k*NN-5 (squared)	0.948	0.650	0.491
Autoencoder	0.947	0.691	0.499
LSTM-AE	0.938	0.661	0.496
GRU-AE	0.943	0.670	0.498
Conv-AE	0.951	0.700	0.507
Transformer-AE	0.978	0.674	0.490

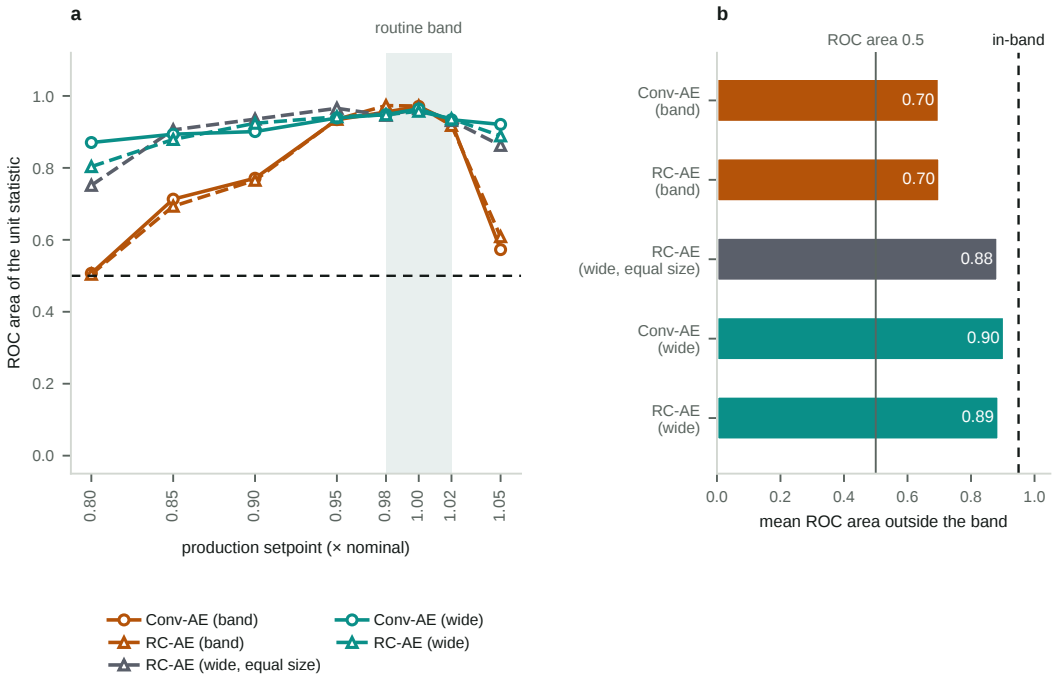


Figure A4. (**a**) ROC area against operating level for the conditioned model and its unconditioned backbone, under band training, equal-size covered training and full covered training. (**b**) Mean ROC area outside the band; the dashed line marks the in-band level that none of them reaches. RC-AE, regime-conditioned autoencoder.

Table A5. Conditioning the model on the operating point, against covering the operating range in the in-control data. Mean ROC area against counterfactual twins. The equal-size row holds the episode count at the band value and spreads those episodes across the levels.

Model and Training Data	In Band	Outside Band	Worst Level
Conv-AE, band	0.951	0.700	0.507
RC-AE, band	0.954	0.701	0.504
RC-AE, wide (equal size)	0.947	0.884	0.751
Conv-AE, wide	0.949	0.905	0.871
RC-AE, wide	0.947	0.887	0.803

Appendix F. The ECAC Procedure

The article describes each step of the procedure separately: the unit statistic in Section 3.5, the per-key quantile in Section 3.7 and the attribution in Section 3.9. The pseudocode below states them together, in the order a plant would run them.

Algorithm A1: Eligibility-conditioned alarm control (ECAC). The detector appears only through the scores s; no step inspects the model that produced them. q_{pool} is the quantile (2) of all calibration units, whatever their key; it is an operational fallback, and an alarm it raises carries no certificate for the key at hand

Input: fault-free calibration episodes with eligibility keys $\{(S_i, k_i)\}$; target level α; persistence m

Output: a threshold per key, an alarm decision, and an attribution over the variables

```
// offline, once per eligibility key
```
foreach *key* k **do**
  $n_k \leftarrow |\{i : k_i = k\}|$; $r_k \leftarrow \lceil (n_k + 1)(1 - \alpha) \rceil$
  if $r_k > n_k$ **then**
    $q_k \leftarrow \infty$ `// α not certifiable at this key; say so`
  else
    $q_k \leftarrow$ the r_k-th smallest of $\{S_i : k_i = k\}$
  end
end

```
// online, per deployed window
```
read the eligibility key k from the control system
$s \leftarrow$ detector scores; $S \leftarrow \max_t \min(s_{t-m+1}, \ldots, s_t)$
if k *was not calibrated* ***or*** $q_k = \infty$ **then**
  report the window as *uncertified at level* α
  `// optional operational fallback, never certified`
  if $S > q_{\text{pool}}$ **then** raise an alarm labelled uncertified
else
  if $S > q_k$ **then** raise an alarm and emit Δ_j over the variables
end

Appendix G. Per-Fault and Phase Detail

The two tables below carry detail the main text states the summary of: the fault-by-fault breakdown behind the pooled ROC area at ×0.80, and the fault-onset phase sweep on the limit cycle at that setpoint.

Table A6. Informativeness by fault condition at ×0.80, ROC area against counterfactual twins. The pooled figure is at chance for all three detectors, but the seven conditions are not alike: one is still ranked well and several are ranked below chance. IDV(n), the nth disturbance of the Tennessee Eastman model.

Condition		PCA-SPE	*k*NN-5	AE
IDV(1)	A/C feed ratio	0.82	0.80	0.80
IDV(4)	reactor cooling	0.40	0.45	0.49
IDV(5)	condenser cooling	0.60	0.60	0.59
IDV(8)	A/B/C composition	0.43	0.41	0.42

Table A6. *Cont.*

Condition	PCA-SPE	*k*NN-5	AE
IDV(10) C feed temperature	0.59	0.61	0.62
IDV(11) reactor cooling	0.45	0.41	0.44
IDV(13) kinetics drift	0.46	0.41	0.41
pooled over conditions	0.510	0.491	0.499
averaged over conditions	0.536	0.526	0.538

Table A7. Phase robustness of the warm-started ×0.80 regime: the fault onset varied across the limit cycle, with ×1.00 as a control; ranges over the three detectors. Each onset carries its own threshold, derived from the commissioning runs windowed at that onset, and the four onsets share their seeds so that only the phase differs.

Onset	False-Alarm Rate	Detection Rate	ROC Area	Paired Win
×0.80, *on the limit cycle*				
48 h	0.071–0.100	0.129–0.171	0.498–0.563	0.414–0.614
54 h	0.000–0.029	0.057–0.071	0.437–0.484	0.357–0.429
60 h	0.114–0.157	0.114–0.157	0.489–0.498	0.400–0.500
66 h	0.086–0.114	0.100–0.243	0.519–0.544	0.543–0.614
×1.00, *control*				
48 h	0.029–0.057	0.857–0.943	0.956–0.987	0.979–1.000
54 h	0.043–0.114	0.886–0.914	0.963–0.973	0.950–0.986
60 h	0.071–0.100	0.886–0.914	0.944–0.968	0.964–0.979
66 h	0.043–0.129	0.871–0.943	0.952–0.975	0.971–0.979

Appendix H. The Dense Sweep, Level by Level

Table A8 is the campaign behind Section 4.4: twenty-six warm-started levels, 64 fault-free commissioning episodes and 70 counterfactual pairs per level, three detectors. D_{shift} is the 2-Wasserstein distance between Gaussian fits of the standardised measurement distribution at that level and over the commissioned range, with the level held out of the reference whenever it falls inside that range; CV is the across-episode coefficient of variation of the fault-free unit statistic over the commissioning pool, the quantity that enters the fit. Both are computed on fault-free episodes alone (no fault, no twin, no threshold), which is what makes them available before deployment. The last column is the ROC area a logistic curve in D_{shift} predicts for that level after being fitted on the other twenty-five.

Table A8. The dense warm-started sweep. The step between ×0.83 and ×0.82 is one setpoint increment and is present in all three detectors separately.

Setpoint	D_{shift}	CV	PCA-SPE	kNN-5	AE	Mean	Predicted
0.80	144.2	0.204	0.462	0.476	0.472	0.470	0.518
0.81	124.4	0.268	0.490	0.508	0.506	0.501	0.539
0.82	106.2	0.366	0.543	0.522	0.523	0.529	0.580
0.83	67.5	0.055	0.807	0.785	0.781	0.791	0.718
0.84	63.3	0.024	0.828	0.782	0.782	0.798	0.745
0.85	59.1	0.025	0.837	0.779	0.777	0.798	0.772
0.86	54.9	0.028	0.857	0.759	0.768	0.795	0.797
0.87	50.7	0.040	0.870	0.778	0.799	0.816	0.818
0.88	46.5	0.034	0.857	0.765	0.797	0.806	0.840
0.89	42.2	0.039	0.894	0.728	0.771	0.798	0.862
0.90	38.0	0.049	0.950	0.818	0.868	0.879	0.874

Table A8. *Cont.*

Setpoint	D_{shift}	CV	PCA-SPE	kNN-5	AE	Mean	Predicted
0.91	34.0	0.063	0.964	0.807	0.871	0.881	0.891
0.92	29.8	0.072	0.941	0.828	0.889	0.886	0.905
0.93	25.9	0.104	0.947	0.852	0.920	0.906	0.917
0.94	22.2	0.097	0.986	0.927	0.968	0.960	0.925
0.95	18.7	0.086	0.984	0.926	0.954	0.955	0.935
0.96	15.2	0.102	0.995	0.950	0.976	0.974	0.942
0.97	11.8	0.100	0.980	0.965	0.965	0.970	0.949
0.98	9.8	0.091	0.985	0.949	0.961	0.965	0.953
0.99	6.6	0.090	0.968	0.947	0.939	0.951	0.960
1.00	5.0	0.092	0.983	0.946	0.965	0.965	0.962
1.01	6.5	0.096	0.971	0.965	0.954	0.963	0.959
1.02	9.8	0.104	0.970	0.956	0.965	0.964	0.953
1.03	12.0	0.086	0.958	0.929	0.950	0.946	0.950
1.04	15.9	0.195	0.949	0.877	0.926	0.917	0.943
1.05	21.2	0.428	0.905	0.804	0.879	0.863	0.933

Appendix I. The Fine Sweep, Level by Level

Table A9 carries the twenty-one fine levels behind Section 4.4: the displacement, which needs no labels, and the two fault-free quantities of the commissioning pool, then the ROC area of each detector against counterfactual twins, their mean and the paired win rate. Every level rests on 64 commissioning episodes and 70 counterfactual pairs, on seeds disjoint from every other campaign; no episode tripped.

Table A9. The fine sweep, ×0.815 to ×0.835 in steps of ×0.001. p is the share of commissioning episodes in the unsteady mode under the frozen rule; CV is the coefficient of variation of the fault-free unit statistic across the pool.

Setpoint	D_{shift}	p	CV	PCA-SPE	kNN-5	AE	Mean	Paired Win
×0.815	118.1	0.91	0.215	0.505	0.457	0.456	0.472	0.505
×0.816	114.0	0.82	0.291	0.517	0.509	0.509	0.512	0.486
×0.817	113.3	0.82	0.272	0.513	0.523	0.526	0.520	0.476
×0.818	114.6	0.86	0.259	0.517	0.485	0.485	0.496	0.529
×0.819	107.2	0.74	0.355	0.525	0.490	0.485	0.500	0.529
×0.820	101.8	0.62	0.415	0.545	0.523	0.522	0.530	0.486
×0.821	101.3	0.66	0.383	0.557	0.538	0.536	0.544	0.619
×0.822	95.0	0.52	0.458	0.594	0.586	0.578	0.586	0.629
×0.823	88.0	0.41	0.518	0.602	0.595	0.589	0.595	0.576
×0.824	77.5	0.23	0.465	0.643	0.639	0.639	0.640	0.671
×0.825	78.2	0.24	0.448	0.713	0.700	0.687	0.700	0.762
×0.826	70.8	0.06	0.312	0.716	0.716	0.718	0.716	0.724
×0.827	69.3	0.04	0.166	0.748	0.751	0.737	0.745	0.805
×0.828	69.0	0.03	0.163	0.799	0.755	0.753	0.769	0.776
×0.829	68.0	0.01	0.072	0.832	0.740	0.742	0.772	0.814
×0.830	67.6	0.00	0.025	0.833	0.778	0.767	0.792	0.790
×0.831	67.1	0.00	0.023	0.830	0.748	0.737	0.771	0.767
×0.832	66.8	0.00	0.026	0.806	0.724	0.733	0.754	0.724
×0.833	66.3	0.00	0.025	0.846	0.803	0.791	0.813	0.814
×0.834	65.9	0.00	0.024	0.807	0.760	0.763	0.777	0.776
×0.835	65.4	0.00	0.025	0.794	0.741	0.758	0.764	0.786

Appendix J. The Supervised Control, Level by Level

Table A10 accompanies Section 4.5. Both models are trained on the band campaign's training clusters, with the pre-onset window of each episode as the fault-free class and the post-onset window of each faulty episode as the faulty one, and are evaluated on the warm-started campaign against the same counterfactual twins as the one-class detectors. The paired win rate is the share of pairs in which the faulty episode outscores its own twin; the gap between it and the ROC area at ×0.80 is the point of the control.

Table A10. Supervised classifiers across the operating range, against counterfactual twins.

	Logistic		Random Forest	
Setpoint	**ROC Area**	**Paired Win**	**ROC Area**	**Paired Win**
0.80	0.620	0.800	0.522	0.564
0.85	0.932	0.986	0.935	0.957
0.90	0.930	0.929	0.914	0.957
0.95	0.922	0.971	0.946	0.943
0.98	0.921	0.971	0.919	0.900
1.00	0.956	0.986	0.937	0.957
1.02	0.942	0.971	0.939	0.957
1.05	0.861	0.843	0.781	0.836

Table A11. All four supervised families under three training sets, against counterfactual twins on the warm-started campaign: ROC area, 70 pairs per level. *Band*: the band training episodes, as in Table A10. *All levels, step*: labelled windows from the calibration-role episodes of the shift campaign at all eight levels, which reach each setpoint by a step. *All levels, warm*: labelled windows from the calibration-role episodes of the warm-started campaign itself, disjoint from the evaluated episodes. Hyperparameters chosen by five-fold cross-validated grid search within each training set. LR, logistic regression; RF, random forest; GBT, gradient-boosted trees; MLP, multilayer perceptron.

Training	**Model**	**×0.80**	**×0.85**	**×0.90**	**×0.95**	**×0.98**	**×1.00**	**×1.02**	**×1.05**
Band	LR	0.620	0.932	0.930	0.922	0.921	0.956	0.942	0.861
Band	RF	0.522	0.935	0.914	0.946	0.919	0.937	0.939	0.781
Band	GBT	0.421	0.888	0.909	0.938	0.958	0.968	0.978	0.845
Band	MLP	0.531	0.848	0.850	0.894	0.859	0.826	0.824	0.634
All levels, step	LR	0.533	0.985	0.987	0.988	0.959	0.970	0.975	0.942
All levels, step	RF	0.496	0.841	0.785	0.760	0.882	0.976	0.836	0.432
All levels, step	GBT	0.576	0.938	0.875	0.927	0.956	0.985	0.965	0.766
All levels, step	MLP	0.506	0.709	0.815	0.852	0.910	0.946	0.941	0.897
All levels, warm	LR	0.921	1.000	1.000	0.994	0.975	0.996	0.992	0.977
All levels, warm	RF	0.949	0.986	0.979	0.986	0.981	0.997	0.987	0.982
All levels, warm	GBT	0.974	1.000	1.000	0.996	0.988	0.999	0.986	0.989
All levels, warm	MLP	0.848	0.999	1.000	0.990	0.978	0.996	0.986	0.973

Appendix K. Sensitivity to the Scoring Conventions and to the Target Rate

Three conventions are fixed before any result is computed: a six-hour scoring window, a persistence of $m = 3$ samples, and a target false-alarm rate $\alpha = 0.10$. None of them is derived, so what each contributes is measured.

Appendix K.1. Window and Persistence

The warm-started campaign was re-scored at every combination of five window lengths (2 to 6 h) and five persistence lengths (1 to 8 samples), with the calibration pool, the pairing and the detectors unchanged (Table A12). The conclusion the paper turns on does not move. At ×0.80 the detector-averaged ROC area lies between 0.456 and 0.503 at all twenty-five combinations, while in the band it never falls below 0.899, and the realised false-alarm rate at $\alpha = 0.10$ averages between 0.064 and 0.089 over the levels. What the conventions do

change is the height of the curve between the edge and the band: a shorter window lowers the area at ×0.85 and at ×1.05, as a window that covers less of the plant's slow dynamics should, and a longer persistence lowers it slightly away from the band. On the fine axis the window also moves the located crossing, and Table 8 reports by how much.

Appendix K.2. Target Rate

The paper's measure of informativeness is a ROC area, which does not depend on any threshold, so α enters only the validity statements. Across the 141 detector–level cells of the dense and fine campaigns, the realised rate averages 0.043 at $\alpha = 0.05$, 0.090 at 0.10 and 0.193 at 0.20, each within sampling error of its exact marginal value for a calibration pool of 64 ($3/65 = 0.046$, $6/65 = 0.092$, $13/65 = 0.200$). At $\alpha = 0.01$ no finite threshold exists: the split-conformal rank $\lceil (n+1)(1-\alpha) \rceil = 65$ exceeds the pool, so the rule is uncertified at that level in every cell and reports so (Algorithm A1). A target of one percent needs at least 99 calibration episodes per key, which is the calibration-budget problem of Section 4.7.2 in its plainest form. Of the conventional targets, then, 0.05 and 0.10 are both certified by this design and behave alike, and 0.01 is the one it cannot certify; the paper reports 0.10 throughout because a realised rate near 0.09 can be distinguished from a broken one on 70 deployment pairs, while a rate near 0.05 is harder to tell from zero.

Table A12. The warm-started campaign re-scored at twenty-five window–persistence combinations: detector-averaged ROC area against counterfactual twins at ×0.80, ×0.85, in the band (×0.98 to ×1.02, mean) and at ×1.05, and the realised false-alarm rate (FAR) at $\alpha = 0.10$ over all eight levels (mean and largest cell). The declared pair is in bold.

Window (h)	m	×0.80	×0.85	Band	×1.05	FAR Mean	FAR Max
2	1	0.503	0.684	0.906	0.718	0.089	0.257
2	2	0.499	0.678	0.906	0.713	0.089	0.200
2	3	0.502	0.662	0.908	0.706	0.079	0.186
2	5	0.501	0.664	0.900	0.691	0.085	0.186
2	8	0.503	0.641	0.899	0.680	0.085	0.157
3	1	0.499	0.747	0.927	0.759	0.086	0.229
3	2	0.500	0.751	0.931	0.751	0.069	0.129
3	3	0.498	0.727	0.937	0.741	0.077	0.243
3	5	0.496	0.724	0.936	0.723	0.076	0.186
3	8	0.496	0.708	0.937	0.711	0.081	0.157
4	1	0.460	0.774	0.943	0.827	0.085	0.186
4	2	0.462	0.773	0.947	0.820	0.082	0.171
4	3	0.466	0.750	0.956	0.811	0.073	0.157
4	5	0.469	0.749	0.954	0.787	0.075	0.143
4	8	0.473	0.746	0.955	0.776	0.077	0.143
5	1	0.466	0.785	0.955	0.856	0.083	0.171
5	2	0.464	0.778	0.953	0.848	0.084	0.157
5	3	0.460	0.768	0.965	0.845	0.076	0.186
5	5	0.461	0.763	0.965	0.828	0.073	0.171
5	8	0.460	0.761	0.970	0.822	0.074	0.143
6	1	0.460	0.808	0.965	0.855	0.079	0.157
6	2	0.456	0.802	0.962	0.848	0.075	0.114
6	**3**	**0.457**	**0.791**	**0.972**	**0.847**	**0.064**	**0.114**
6	5	0.458	0.784	0.976	0.833	0.071	0.171
6	8	0.458	0.782	0.975	0.826	0.067	0.129

Appendix L. The Intervention, Level by Level

Table A13. All twenty-eight levels of the intervention campaign; columns as in Table 11. At ×0.90 and ×1.00 every unit statistic is bitwise identical in the two builds.

		Unsteady Share		Purge on a Stop		ROC Area, Detector Mean		
Setpoint	**Grid**	**Built**	**Disabled**	**Built**	**Disabled**	**Built**	**Disabled**	**90%, Disabled**
×0.800	dense	1.00	0.01	0.89	0.00	0.470	0.759	[0.703, 0.812]
×0.810	dense	0.89	0.00	0.86	0.00	0.501	0.801	[0.747, 0.855]
×0.820	dense	0.72	0.00	0.75	0.00	0.529	0.757	[0.696, 0.809]
×0.830	dense	0.01	0.00	0.00	0.00	0.791	0.806	[0.756, 0.858]
×0.840	dense	0.00	0.00	0.00	0.00	0.798	0.800	[0.752, 0.848]
×0.900	dense	0.00	0.00	0.00	0.00	0.879	0.879	[0.837, 0.919]
×1.000	dense	0.00	0.00	0.00	0.00	0.965	0.965	[0.945, 0.983]
×0.815	fine	0.91	0.00	0.89	0.00	0.472	0.773	[0.718, 0.828]
×0.816	fine	0.82	0.00	0.83	0.00	0.512	0.751	[0.693, 0.807]
×0.817	fine	0.82	0.00	0.86	0.00	0.520	0.771	[0.717, 0.830]
×0.818	fine	0.86	0.00	0.88	0.00	0.496	0.764	[0.706, 0.816]
×0.819	fine	0.74	0.00	0.73	0.00	0.500	0.794	[0.737, 0.848]
×0.820	fine	0.62	0.00	0.66	0.00	0.530	0.765	[0.709, 0.821]
×0.821	fine	0.66	0.00	0.70	0.00	0.544	0.726	[0.668, 0.781]
×0.822	fine	0.52	0.00	0.55	0.00	0.586	0.803	[0.754, 0.852]
×0.823	fine	0.41	0.00	0.42	0.00	0.595	0.790	[0.733, 0.847]
×0.824	fine	0.23	0.00	0.23	0.00	0.640	0.794	[0.738, 0.846]
×0.825	fine	0.24	0.00	0.30	0.00	0.700	0.797	[0.736, 0.850]
×0.826	fine	0.06	0.00	0.06	0.00	0.716	0.825	[0.772, 0.872]
×0.827	fine	0.04	0.00	0.03	0.00	0.745	0.821	[0.772, 0.869]
×0.828	fine	0.03	0.00	0.03	0.00	0.769	0.814	[0.761, 0.860]
×0.829	fine	0.01	0.00	0.00	0.00	0.772	0.778	[0.716, 0.839]
×0.830	fine	0.00	0.00	0.00	0.00	0.792	0.807	[0.752, 0.859]
×0.831	fine	0.00	0.00	0.00	0.00	0.771	0.771	[0.717, 0.827]
×0.832	fine	0.00	0.00	0.00	0.00	0.754	0.760	[0.699, 0.819]
×0.833	fine	0.00	0.00	0.00	0.00	0.813	0.814	[0.764, 0.860]
×0.834	fine	0.00	0.00	0.00	0.00	0.777	0.777	[0.718, 0.834]
×0.835	fine	0.00	0.00	0.00	0.00	0.764	0.768	[0.715, 0.821]

Appendix M. Intervals at 95%

Every ROC-area interval in the paper is a 90% pair-preserving percentile bootstrap. Recomputed at 95% from the same pairs (2000 resamples, all 165 detector–level cells of the shift, dense and fine campaigns), eleven intervals change a binary reading; Table A14 lists all of them. Five move from just excluding one half to containing it, for cells whose ROC area is within 0.04 of one half; six move from lying above, or below, the 0.70 floor to straddling it, at levels where the floor is applied to the detector-averaged point estimate and not to a single detector's interval. The dense step is unaffected: at 95% the ×0.83 and ×0.82 intervals are [0.748, 0.865] against [0.518, 0.572] for PCA-SPE, [0.712, 0.853] against [0.498, 0.549] for kNN-5 and [0.711, 0.850] against [0.499, 0.551] for the autoencoder.

Table A14. The eleven of 165 intervals whose binary reading differs between 90% and 95%.

Campaign	**Setpoint**	**Detector**	**ROC Area**	**90%**	**95%**	**What Changes**
dense	×0.800	AE	0.472	[0.440, 0.500]	[0.436, 0.505]	now contains $\frac{1}{2}$
dense	×0.820	kNN-5	0.522	[0.502, 0.544]	[0.498, 0.549]	now contains $\frac{1}{2}$
dense	×0.820	AE	0.523	[0.503, 0.546]	[0.499, 0.551]	now contains $\frac{1}{2}$
dense	×0.860	AE	0.768	[0.706, 0.828]	[0.693, 0.838]	above → straddles
dense	×0.890	AE	0.771	[0.712, 0.835]	[0.695, 0.845]	above → straddles
fine	×0.819	PCA-SPE	0.525	[0.502, 0.548]	[0.498, 0.553]	now contains $\frac{1}{2}$
fine	×0.821	AE	0.536	[0.505, 0.568]	[0.498, 0.575]	now contains $\frac{1}{2}$
fine	×0.824	PCA-SPE	0.643	[0.599, 0.693]	[0.591, 0.703]	below → straddles
fine	×0.830	AE	0.767	[0.707, 0.826]	[0.696, 0.837]	above → straddles
fine	×0.835	AE	0.758	[0.702, 0.815]	[0.691, 0.824]	above → straddles
shift	×0.850	PCA-SPE	0.752	[0.702, 0.803]	[0.690, 0.811]	above → straddles

References

1. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer: New York, NY, USA, 2005. [CrossRef]
2. Angelopoulos, A.N.; Bates, S. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591. [CrossRef]
3. Barber, R.F.; Candès, E.J.; Ramdas, A.; Tibshirani, R.J. The limits of distribution-free conditional predictive inference. *Inf. Inference J. IMA* **2021**, *10*, 455–482. [CrossRef]
4. Vovk, V. Conditional validity of inductive conformal predictors. *Proc. Asian Conf. Mach. Learn.* **2012**, *25*, 475–490.
5. Johannssen, A.; Qiu, P.; Yeganeh, A.; Chukhrova, N. Explainable AI for trustworthy intelligent process monitoring. *Comput. Ind. Eng.* **2025**, *209*, 111407. [CrossRef]
6. Chiang, L.H.; Russell, E.L.; Braatz, R.D. *Fault Detection and Diagnosis in Industrial Systems*; Springer: London, UK, 2001. [CrossRef]
7. Ma, J.; Zhang, J. Progress of Process Monitoring for the Multi-Mode Process: A Review. *Appl. Sci.* **2022**, *12*, 7207. [CrossRef]
8. Garg, S.; Balakrishnan, S.; Lipton, Z.C.; Neyshabur, B.; Sedghi, H. Leveraging Unlabeled Data to Predict Out-of-Distribution Performance. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
9. Yin, S.; Ding, S.X.; Haghani, A.; Hao, H.; Zhang, P. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *J. Process Control* **2012**, *22*, 1567–1581. [CrossRef]
10. Guillory, D.; Shankar, V.; Ebrahimi, S.; Darrell, T.; Schmidt, L. Predicting with Confidence on Unseen Distributions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 1134–1144.
11. Downs, J.J.; Vogel, E.F. A plant-wide industrial process control problem. *Comput. Chem. Eng.* **1993**, *17*, 245–255. [CrossRef]
12. Ricker, N.L. Decentralized control of the Tennessee Eastman challenge process. *J. Process Control* **1996**, *6*, 205–221. [CrossRef]
13. Xavier, G.M.; de Seixas, J.M. Fault detection and diagnosis in a chemical process using long short-term memory recurrent neural network. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8. [CrossRef]
14. Diallo, A.R.; Homri, L.; Dantan, J.Y. Reducing false alarms in fault detection: A comparative analysis between conformal prediction and classical methods applied to PCA and autoencoders. *J. Process Control* **2025**, *152*, 103495. [CrossRef]
15. Rieth, C.A.; Amsel, B.D.; Tran, R.; Cook, M.B. *Additional Tennessee Eastman Process Simulation Data for Anomaly Detection Evaluation*; Harvard Dataverse: Cambridge, MA, USA, 2017. [CrossRef]
16. Melo, A.; Câmara, M.M.; Pinto, J.C. Data-Driven Process Monitoring and Fault Diagnosis: A Comprehensive Survey. *Processes* **2024**, *12*, 251. [CrossRef]
17. Qin, S.J. Survey on data-driven industrial process monitoring and diagnosis. *Annu. Rev. Control* **2012**, *36*, 220–234. [CrossRef]
18. Park, P.; Di Marco, P.; Shin, H.; Bang, J. Fault Detection and Diagnosis Using Combined Autoencoder and Long Short-Term Memory Network. *Sensors* **2019**, *19*, 4612. [CrossRef]
19. Jang, K.; Hong, S.; Kim, M.; Na, J.; Moon, I. Adversarial Autoencoder Based Feature Learning for Fault Detection in Industrial Processes. *IEEE Trans. Ind. Inform.* **2022**, *18*, 827–834. [CrossRef]
20. Luna-Villagómez, E.; Mahalec, V. Exploring Kolmogorov–Arnold Networks for Unsupervised Anomaly Detection in Industrial Processes. *Processes* **2025**, *13*, 3672. [CrossRef]
21. Quiñones-Grueiro, M.; Prieto-Moreno, A.; Verde, C.; Llanes-Santiago, O. Data-driven monitoring of multimode continuous processes: A review. *Chemom. Intell. Lab. Syst.* **2019**, *189*, 56–71. [CrossRef]
22. Ge, Z.; Song, Z.; Gao, F. Review of Recent Research on Data-Based Process Monitoring. *Ind. Eng. Chem. Res.* **2013**, *52*, 3543–3562. [CrossRef]
23. Sun, W.; Zhou, Z.; Ma, F.; Wang, J.; Ji, C. Industrial Application of Data-Driven Process Monitoring with an Automatic Selection Strategy for Modeling Data. *Processes* **2023**, *11*, 402. [CrossRef]
24. Papadopoulos, H.; Proedrou, K.; Vovk, V.; Gammerman, A. Inductive confidence machines for regression. In Proceedings of the Machine Learning: ECML 2002, Helsinki, Finland, 19–23 August 2002; pp. 345–356. [CrossRef]
25. Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R.J.; Wasserman, L. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **2018**, *113*, 1094–1111. [CrossRef]
26. Laxhammar, R.; Falkman, G. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Ann. Math. Artif. Intell.* **2015**, *74*, 67–94. [CrossRef]
27. Burger, C. Distribution-free process monitoring with conformal prediction. *arXiv* **2025**, arXiv:2512.23602. [CrossRef]
28. Boström, H.; Johansson, U.; Löfström, T. Mondrian conformal predictive distributions. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications, Virtual, 8–10 September 2021*; PMLR: Cambridge, MA, USA, 2021; Volume 152, pp. 24–38.
29. Tibshirani, R.J.; Barber, R.F.; Candès, E.J.; Ramdas, A. Conformal prediction under covariate shift. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.

30. Gibbs, I.; Candès, E. Adaptive conformal inference under distribution shift. In *Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–14 December 2021*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 1660–1672.
31. Barber, R.F.; Candès, E.J.; Ramdas, A.; Tibshirani, R.J. Conformal prediction beyond exchangeability. *Ann. Stat.* **2023**, *51*, 816–845. [CrossRef]
32. Huang, Y.; Zhang, J.; Liu, R.; Zhao, S. Improving Accuracy and Interpretability of CNN-Based Fault Diagnosis through an Attention Mechanism. *Processes* **2023**, *11*, 3233. [CrossRef]
33. Saydirasulov, N.S.; Buriboev, A.S.; Isroilov, S.; Oh, R.; Buribayev, S.; Abduvaytov, A.; Umirov, J.; Badalov, J.I.; Axmedova, A.; Lee, C.; et al. A reproducible benchmark-validity audit and calibration study for cross-home fault diagnosis in smart-home sensor systems. *Sensors* **2026**, *26*, 5025. [CrossRef]
34. Saydirasulovich, N.S.; Davronbekov, D.A.; Mubashirovich, M.M.; Cho, Y.I. A closed-loop measurement study of runtime governance in AI-driven smart building climate control. *Sensors* **2026**, *26*, 4921. [CrossRef]
35. Ji, C.; Ma, F.; Rao, J.; Wang, J.; Sun, W. Bridging the Gap in Chemical Process Monitoring: Beyond Algorithm-Centric Research Toward Industrial Deployment. *Processes* **2025**, *13*, 3809. [CrossRef]
36. Law, A.M. *Simulation Modeling and Analysis*, 5th ed.; McGraw-Hill Education: New York, NY, USA, 2015.
37. Glasserman, P.; Yao, D.D. Some Guidelines and Guarantees for Common Random Numbers. *Manag. Sci.* **1992**, *38*, 884–908. [CrossRef]
38. Natella, R.; Cotroneo, D.; Madeira, H.S. Assessing Dependability with Software Fault Injection: A Survey. *ACM Comput. Surv.* **2016**, *48*, 44. [CrossRef]
39. Wright, R.D.; Ramsay, T.E. On the Effectiveness of Common Random Numbers. *Manag. Sci.* **1979**, *25*, 649–656. [CrossRef]
40. Vovk, V.; Nouretdinov, I.; Fedorova, V.; Petej, I.; Gammerman, A. Criteria of Efficiency for Conformal Prediction. In Proceedings of the 5th Symposium on Conformal and Probabilistic Prediction with Applications (COPA), Madrid, Spain, 20–22 April 2016.
41. Katser, I.D.; Kozitsin, V.O. Skoltech Anomaly Benchmark (SKAB). 2020. Available online: https://www.kaggle.com/dsv/1693952 (accessed on 25 October 2026).
42. Janes, H.; Pepe, M.S. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika* **2009**, *96*, 371–382. [CrossRef]
43. Deng, W.; Zheng, L. Are Labels Always Necessary for Classifier Accuracy Evaluation? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 15069–15078.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.