



Low-False-Alarm Evaluation Changes the Ranking of Multisensor Anomaly Detectors

Murilo Costa Salem^{1*}, Anderson Priebe Ferrugem¹

¹ Universidade Federal de Pelotas, Pelotas/RS – Brasil

* Corresponding Author: mcsalem@inf.ufpel.edu.br

Keywords

Multisensor anomaly detection
Low false-alarm evaluation
Multivariate time series
Conformal anomaly detection
Temporal modeling

Abstract

This study investigates multisensor anomaly detection under low false-alarm constraints in safety-critical systems, examining whether conventional aggregate metrics provide a reliable basis for model selection under strict false-alarm budgets. A detection framework combining temporal modeling, anomaly scoring, calibration, and temporal persistence is evaluated across forecasting, autoencoding, and conformal multisensor approaches on four public multivariate sensor datasets. Low-false-alarm evaluation changes detector rankings in seven model-dataset pairs: on the Mars Science Laboratory dataset, the Temporal Autoencoder moves from third to first place under recall at false-alarm rate up to 0.05, reaching 0.1062; on the Secure Water Treatment dataset, the Gated Recurrent Unit Forecaster reaches 0.6448; on the Water Distribution dataset, subsystem-aware grouping increases the same metric from 0.0241 to 0.1838, while correlation-based grouping degrades performance. Calibration improves false-alarm control but does not compensate for poor score separability, and cross-sensor modeling is beneficial only when groupings reflect the physical organization of the monitored system.

Article Information

Received 26 April 2026
Revised 03 August 2026
Accepted 07 October 2026
Available 09 October 2026

1. Introduction

1.1. Operational Motivation and Problem Statement

Anomaly detection in multisensor systems has become increasingly relevant in safety-critical applications, where monitoring failures, abnormal system states, or unexpected events must be identified with high reliability [1]. In these contexts, the operational cost of false alarms is often comparable to, or even greater than, the cost of missed detections, since excessive false positives may overload operators, reduce trust in automated monitoring, and compromise timely decision-making [2], [3]. As a result, the practical value of an anomaly detector is not determined only by its average discriminative performance, but also by its ability to operate under strict false-alarm constraints [2], [3].

Let $\hat{y}_t \in \{0, 1\}$ denote the alarm decision at time t and $y_t \in \{0, 1\}$ the ground-truth state. The operational objective can be expressed as minimizing

$$\mathcal{R} = c_{FP} \mathbb{P}(\hat{y}_t = 1 \mid y_t = 0) + c_{FN} \mathbb{P}(\hat{y}_t = 0 \mid y_t = 1), \quad (1)$$

where c_{FP} and c_{FN} denote the operational costs of false positives and false negatives, respectively.

This requirement is especially important in systems that integrate heterogeneous sensor channels and temporal evidence. Multisensor anomaly detection must account not only for temporal deviations within each channel, but also for dependencies among channels, calibration effects, and persistence over time [4], [5]. These factors make the design and evaluation of anomaly detectors more complex than standard classification or ranking problems [4]. In practice, a detector that performs well under aggregate metrics may still be unsuitable if its operating point cannot satisfy a low false-alarm budget, and post-hoc calibration alone does not eliminate limitations imposed by the underlying score quality [6].

This work investigated multisensor anomaly detection under low false-alarm constraints using public multivariate time-series datasets that represent different operational conditions. It considered forecasting, autoencoding, and conformal multisensor approaches, and examined whether evaluation under strict false-alarm budgets changed the relative assessment of detectors. It also analyzed whether post-hoc calibration was sufficient to improve operation, or whether the main limitation in challenging domains remained the

The peer review process was coordinated, and the final publication approved, by Humberto Baldessarini Pires.

separability of anomaly scores.

1.2. Limitations of Aggregate Evaluation Metrics

Multivariate anomaly detection is commonly reported with aggregate metrics such as precision-recall area, receiver operating characteristic area, or event-level F1-score [4]. Although these metrics are useful for broad comparisons, they do not necessarily reflect operational performance in safety-critical settings [7]. In particular, they may favor detectors that achieve good average ranking while remaining ineffective when the admissible false-alarm rate is very low. This mismatch becomes critical when the detector must operate under explicit false-alarm budgets rather than under a threshold chosen only to maximize a global summary metric [4], [8].

The experimental results of this study showed that this limitation was not only theoretical. When the evaluation emphasized recall at false-alarm rate up to 0.05, detector rankings changed in seven model-dataset pairs relative to rankings based on precision-recall area. The most representative case occurred on the Mars Science Laboratory dataset, where the Temporal Autoencoder moved from third place under precision-recall area to first place under low-false-alarm recall, reaching 0.1062. This result indicated that aggregate metrics alone could lead to different model choices from those supported by operationally constrained evaluation.

A related limitation concerns the interpretation of calibration. Thresholding and calibration strategies can modify the operating point of a detector, but they cannot create discriminative structure that is absent from the score itself. This distinction became evident in the calibration experiments, where fixed and split calibration produced nearly identical behavior in several settings, while rolling calibration improved false-alarm control in some cases but often reduced recall substantially. Therefore, a realistic evaluation protocol must distinguish between improvements in operational control and improvements in score separability [6].

1.3. Contributions

This paper made three main contributions. First, it presented a compact evaluation framework for multisensor anomaly detection under low false-alarm constraints, combining temporal modeling, anomaly scoring, calibration strategies, and temporal persistence rules in a unified operational setting. Second, it demonstrated experimentally, across four public datasets, that low-false-alarm evaluation changed detector ranking relative to conventional aggregate metrics, reinforcing the need for matched-false-alarm comparisons in safety-critical applications [4], [8]. Third, it provided a detailed case study on the Water Distribution dataset, showing that improved performance in the low-false-alarm regime depended more on backbone choice and subsystem-aware grouping than on calibration alone. In that dataset, subsystem-aware grouping increased recall at false-alarm rate up to 0.05 from 0.0241 to 0.1838, highlighting the importance of domain-structured multisensor modeling [4],

[5].

Taken together, these results supported a central claim: in safety-critical multisensor anomaly detection, calibration improved operational control, but effective low-false-alarm performance ultimately depended on the separability of anomaly scores and on the structural alignment between the model and the monitored system.

2. Related Work

2.1. Anomaly Detection and Evaluation Metrics

Anomaly detection in multivariate time series has been studied extensively, with approaches organized around forecasting-based, reconstruction-based, and consistency-based methods [4]. Forecasting-based detectors treat large predictive residuals as anomaly evidence; reconstruction-based detectors identify samples that cannot be faithfully reproduced under a learned normal representation; and consistency-based methods exploit inter-variable dependencies to detect violations of expected joint sensor behavior [5]. Despite this diversity, most published results are reported with aggregate metrics such as precision-recall area or point-adjusted F1-score, which summarize discriminative quality over all thresholds and do not reflect performance at a specific operating point. Mueller [8] discusses open evaluation challenges including metric sensitivity, point-adjustment inflation, and the difficulty of comparing detectors at different false-alarm rates. The present work addresses these gaps by adopting recall under an explicit false-alarm budget as the primary metric and by enforcing matched-budget comparisons before event-level performance is analyzed.

2.2. Calibration, Cross-sensor Modeling, and Safety-critical Operation

Threshold selection bridges score generation and alarm decision. Gloumeau et al. [6] show that post-hoc calibration improves operational false-alarm control without altering the underlying score, and distinguish between improvements in threshold placement and improvements in score separability—a distinction central to the conclusions of this study. On the modeling side, graph-based and attention-based methods have demonstrated that inter-variable structure improves detection when groupings reflect physical system organization [5], a finding confirmed and extended here by the Water Distribution case study. In safety-critical contexts, Erhan et al. [1] note that practical detector utility depends on behavior at the operating point rather than average discriminative quality, while Simonson et al. [2] and Wiczorek and Zirk [3] show that excessive false alarms degrade operator trust and response. The present work translates these motivations into a formal low-false-alarm evaluation protocol and demonstrates experimentally, across four heterogeneous datasets, that this protocol produces different model rankings than conventional aggregate metrics.

3. Method

3.1. Multisensor Anomaly Detection Framework

The study evaluated a multisensor anomaly detection framework designed for operation under strict false-alarm constraints. The framework combined temporal representation learning, anomaly score construction, calibration, and temporal persistence in a modular pipeline. This design allowed different detector families to be compared under the same operational protocol while preserving the characteristics of forecasting, reconstruction, and multisensor consistency modeling [4]–[6].

Let $\mathbf{x}_{1:T} \in \mathbb{R}^{T \times D}$ denote a multivariate time series with D sensor channels observed over T time steps. For each dataset, the input sequence was divided into sliding windows of fixed length, normalized using statistics computed on the training partition, and processed by a temporal backbone. Depending on the detector family, the backbone generated one or more anomaly score components. Forecasting-based models estimated future values from past observations and produced a predictive residual score. Autoencoding models reconstructed the input window and produced a reconstruction error score. The conformal multisensor detector combined predictive and reconstructive evidence with an additional cross-sensor consistency component derived from grouped channels [4]–[6].

The anomaly score at time step t was modeled as

$$s_t = \alpha s_t^{\text{pred}} + \beta s_t^{\text{recon}} + \gamma s_t^{\text{cross}}, \quad (2)$$

where s_t^{pred} denoted the prediction residual, s_t^{recon} denoted the reconstruction error, and s_t^{cross} denoted a cross-sensor consistency term. The scalar weights α , β , and γ controlled the contribution of each component. This formulation allowed the study to examine whether low-false-alarm performance benefited from predictive, reconstructive, or grouped multisensor information, and whether these contributions were domain dependent [4], [5].

For the conformal multisensor detector, grouped channels were used to encode subsystem structure or statistical relationships among sensors. In the most general form, one subset of channels was estimated from the remaining channels, and the discrepancy between the estimated and observed values contributed to the consistency score. This mechanism was intended to detect anomalies that were not only temporally unusual within one channel, but also inconsistent with the expected joint behavior of related channels [4], [5]. The framework also included a temporal persistence stage, in which raw anomaly decisions were smoothed with short persistence rules in order to reduce isolated spikes and better reflect event-level operation [4].

The same implementation protocol was applied across detector families. Forecasting, temporal autoencoding, and conformal multisensor models were trained on the normal operating portion of each dataset, and their scores were evaluated under the same downstream calibration and decision process whenever applicable. This unified design made it possible to compare model families as operational detectors rather than only as score generators [4], [6].

3.2. Low-false-alarm Evaluation Protocol

The evaluation protocol was designed to reflect the conditions of safety-critical monitoring, in which the detector must operate under explicit false-alarm budgets. Instead of relying only on aggregate ranking metrics, the protocol emphasized performance in the low-false-alarm regime. Let $s_t \in \mathbb{R}$ denote the anomaly score at time t , $y_t \in \{0, 1\}$ the corresponding ground-truth label, and τ a decision threshold. For a given threshold τ , the false-alarm rate was defined as

$$\text{FAR}(\tau) = \frac{\sum_{t=1}^T \mathbb{I}[s_t > \tau \wedge y_t = 0]}{\sum_{t=1}^T \mathbb{I}[y_t = 0]}, \quad (3)$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. The main metric was recall under a false-alarm budget ε , denoted as $\text{Recall@FAR} \leq \varepsilon$, and computed as

$$\text{Recall@FAR} \leq \varepsilon = \max_{\tau: \text{FAR}(\tau) \leq \varepsilon} \frac{\sum_{t=1}^T \mathbb{I}[s_t > \tau \wedge y_t = 1]}{\sum_{t=1}^T \mathbb{I}[y_t = 1]}. \quad (4)$$

In the experiments, the primary operating budget was set to $\varepsilon = 0.05$. Restricted precision-recall area in the same regime was also used to quantify score quality when the admissible false-alarm region was limited. In its threshold-based form, it was expressed as

$$\text{rPR-AUC}_\varepsilon = \int_{\tau: \text{FAR}(\tau) \leq \varepsilon} \text{Precision}(\tau) \left| \frac{d\text{Recall}(\tau)}{d\tau} \right| d\tau. \quad (5)$$

In addition, the study reported empirical false-alarm rate, event-level F1-score, and conventional precision-recall area to contrast operationally constrained and aggregate views of performance.

To separate ranking quality from threshold selection, the protocol evaluated both score behavior and final operating points. Score-based comparisons examined how well each detector ordered anomalous and normal samples under a low false-alarm budget. Operating-point comparisons then applied calibration and temporal persistence to determine whether the ranked scores could be converted into a practically useful detector. This distinction was essential because calibration could modify thresholds and empirical false-alarm behavior, but it could not recover discriminative structure absent from the anomaly scores.

Three calibration strategies were considered in the study. Fixed calibration used a single threshold estimated from a validation partition. Split calibration used a held-out calibration subset to define a quantile-based threshold. Given calibration scores $\{s_i^{\text{cal}}\}_{i=1}^n$ and target error level α , the split-calibrated threshold was defined as

$$\tau_\alpha = Q_{1-\alpha} \left(\{s_i^{\text{cal}}\}_{i=1}^n \right), \quad (6)$$

where $Q_{1-\alpha}(\cdot)$ denotes the empirical $(1 - \alpha)$ -quantile. The corresponding point-wise alarm decision was given by

$$\hat{y}_t = \mathbb{I}[s_t > \tau_\alpha]. \quad (7)$$

Rolling calibration updated the threshold over time using recent score statistics, with the goal of improving robustness to drift and nonstationarity. After calibration, temporal persistence rules were applied to reduce spurious isolated

detections and to produce event-oriented outputs. Using a k -of- n persistence rule, the final alarm was computed as

$$\hat{z}_t = \mathbb{I} \left[\sum_{j=t-n+1}^t \hat{y}_j \geq k \right]. \quad (8)$$

This step aligned the detector with operational use, in which alarms are typically interpreted as temporally sustained events rather than isolated points.

The protocol also included matched-false-alarm comparisons. In these experiments, detectors were compared at the same false-alarm budget before event-level performance was analyzed. For model m , the matched-budget operating threshold at budget ε was defined as

$$\tau_m^{(\varepsilon)} = \arg \max_{\tau: \text{FAR}_m(\tau) \leq \varepsilon} \text{Recall}_m(\tau). \quad (9)$$

This prevented a detector with a permissive threshold from appearing artificially better only because it generated more alarms. The same logic motivated the ranking analysis performed across datasets, where precision-recall area and $\text{Recall@FAR} \leq 0.05$ were compared directly. By combining low-false-alarm ranking, matched-budget comparison, calibration analysis, and event-level evaluation, the protocol provided a more realistic basis for assessing multisensor anomaly detection in safety-critical systems.

4. Experimental Setup

4.1. Datasets and Detector Families

The experimental evaluation was conducted on four public multivariate time-series datasets selected to represent distinct sensing and operational conditions. The Spacecraft Monitoring and Anomaly Processing dataset and the Mars Science Laboratory dataset represented telemetry monitoring scenarios with complex temporal dynamics and sparse anomalous events [9]. The Secure Water Treatment dataset and the Water Distribution dataset represented industrial control and water infrastructure monitoring scenarios, with larger numbers of channels and event structures that were challenging for low-false-alarm detection [10], [11]. This combination allowed the study to compare detector behavior across heterogeneous domains rather than within a single application setting.

Table I summarizes structural characteristics of the four datasets that are relevant to low-false-alarm evaluation. The Water Distribution dataset differs substantially from the others in anomaly density: approximately 20% of its duration corresponds to attack periods, compared with under 12% in the remaining datasets. This higher anomaly proportion, combined with a sliding window of length 128 and stride 1, generates a large fraction of mixed windows that contain both normal and anomalous timesteps. Score values in these transition windows are ambiguous, which raises the empirical false-alarm rate regardless of the backbone or score component used. This structural property explains why the empirical FAR in the Water Distribution dataset remained near 0.40 across all detector variants in Table II, while the other three datasets operated near zero.

Table I. Dataset structural characteristics relevant to low-false-alarm evaluation.

Dataset	Channels	Anomaly %	Window	Emp. FAR
SMAP	25	$\approx 5\%$	128	≈ 0.00
MSL	55	$\approx 10\%$	128	≈ 0.01
SWaT	51	$\approx 11\%$	128	≈ 0.01
WADI	127	$\approx 20\%$	128	≈ 0.40

All datasets were converted to a common processing format. Each time series was segmented into sliding windows, normalized using statistics from the training partition, and evaluated with the same downstream metrics. The experimental pipeline preserved the original train, validation, and test roles used in the study. When needed, validation partitions were used for threshold selection or calibration. In the Water Distribution dataset, the validation regime remained normal-only, which allowed false-alarm-oriented tuning without using anomalous test information to optimize recall.

Three detector families were considered. The first family was forecasting-based detection, implemented with temporal predictors that estimated future observations from historical windows and used predictive residuals as anomaly scores. The strongest forecasting baseline in the final seeded comparison was the Gated Recurrent Unit Forecaster. The second family was reconstruction-based detection, implemented with temporal autoencoding models that reconstructed the input window and used reconstruction error as anomaly evidence. The strongest reconstruction baseline in the final seeded comparison was the Temporal Autoencoder. The third family was the conformal multisensor detector - referred to hereafter as CMAD-LFA (Conformal Multisensor Anomaly Detector for Low False-Alarm operation) - which combined predictive, reconstructive, and grouped cross-sensor consistency components in a single score, followed by calibration and temporal persistence.

The conformal multisensor detector was also used for targeted ablation studies. In the Water Distribution dataset, additional experiments examined prediction-only, reconstruction-only, and mixed-score variants, as well as changes in backbone and grouping strategy. These experiments were used to determine whether the main bottleneck was thresholding or score separability. The resulting case study showed that, in this dataset, backbone and subsystem-aware grouping had a larger effect on low-false-alarm separability than score simplification alone.

To reduce the influence of random initialization, the main detector comparison was performed with three seeds. The seeded results were used in the main tables and ranking analyses, while targeted diagnostic studies were reported separately when they addressed specific hypotheses, such as calibration effects, grouping effects, or low-false-alarm trade-offs in the Water Distribution dataset.

Table II. Cross-dataset results in the low-false-alarm regime (mean $\pm$ std over 3 seeds). Best Recall@FAR ≤ 0.05 per dataset in bold.

Dataset	Model	FAR	Recall@FAR ≤ 0.01	Recall@FAR ≤ 0.05	rPR-AUC	PR-AUC
SMAP	GRU Forecaster	0.0000 $\pm$ 0.0000	0.0044 $\pm$ 0.0006	0.0317 $\pm$ 0.0002	0.0027 $\pm$ 0.0000	0.1063 $\pm$ 0.0037
SMAP	Temporal Autoencoder	0.0000 $\pm$ 0.0000	0.0345 $\pm$ 0.0005	0.0582 $\pm$ 0.0029	0.0203 $\pm$ 0.0011	0.1289 $\pm$ 0.0030
SMAP	CMAD-LFA	0.0000 $\pm$ 0.0000	0.0041 $\pm$ 0.0006	0.0328 $\pm$ 0.0002	0.0029 $\pm$ 0.0000	0.1056 $\pm$ 0.0014
MSL	GRU Forecaster	0.0117 $\pm$ 0.0039	0.0507 $\pm$ 0.0092	0.1049 $\pm$ 0.0030	0.0374 $\pm$ 0.0049	0.1655 $\pm$ 0.0066
MSL	Temporal Autoencoder	0.0136 $\pm$ 0.0005	0.0676 $\pm$ 0.0005	0.1062 $\pm$ 0.0000	0.0437 $\pm$ 0.0001	0.1575 $\pm$ 0.0016
MSL	CMAD-LFA	0.0026 $\pm$ 0.0001	0.0483 $\pm$ 0.0020	0.1029 $\pm$ 0.0020	0.0343 $\pm$ 0.0013	0.1653 $\pm$ 0.0016
SWaT	GRU Forecaster	0.0082 $\pm$ 0.0013	0.0307 $\pm$ 0.0030	0.6448 $\pm$ 0.0082	0.3298 $\pm$ 0.0188	0.3863 $\pm$ 0.0151
SWaT	Temporal Autoencoder	0.0116 $\pm$ 0.0020	0.0396 $\pm$ 0.0025	0.6444 $\pm$ 0.0125	0.2212 $\pm$ 0.0424	0.2810 $\pm$ 0.0339
SWaT	CMAD-LFA	0.0127 $\pm$ 0.0011	0.0147 $\pm$ 0.0004	0.6394 $\pm$ 0.0093	0.3181 $\pm$ 0.0088	0.3738 $\pm$ 0.0080
WADI	GRU Forecaster	0.4793 $\pm$ 0.0185	0.0246 $\pm$ 0.0278	0.0861 $\pm$ 0.0609	0.0812 $\pm$ 0.0606	0.9525 $\pm$ 0.0087
WADI	Temporal Autoencoder	0.4731 $\pm$ 0.0071	0.0018 $\pm$ 0.0025	0.0553 $\pm$ 0.0720	0.0522 $\pm$ 0.0696	0.9412 $\pm$ 0.0108
WADI	CMAD-LFA	0.3964 $\pm$ 0.0001	0.0105 $\pm$ 0.0001	0.0241 $\pm$ 0.0040	0.0206 $\pm$ 0.0038	0.9385 $\pm$ 0.0008

4.2. Metrics, Calibration, and Reproducibility Protocol

The study evaluated detectors with both aggregate and operational metrics. Precision-recall area was used as a conventional global metric to summarize ranking quality over all thresholds. Event-level F1-score and empirical false-alarm rate were used to quantify operational behavior after thresholding. However, the main low-false-alarm metrics were recall at false-alarm rate up to 0.05 and restricted precision-recall area in the same regime. These metrics emphasized performance under constrained false-alarm budgets and formed the basis of the ranking comparisons discussed later in the paper [4], [8].

To ensure fair comparison, the study also performed matched-false-alarm evaluation. In this protocol, detectors were compared under the same false-alarm budget before event-level performance was analyzed. This avoided the common distortion in which a more permissive detector achieved higher event-level metrics only because it triggered substantially more alarms. The matched-budget analysis was particularly important in the Mars Science Laboratory and Water Distribution datasets, where aggregate metrics and low-false-alarm metrics led to different rankings [4], [8].

Three calibration settings were examined. Fixed calibration used a threshold selected from a validation split. Split calibration used a held-out calibration subset to define a quantile-based operating threshold. Rolling calibration adapted the threshold over time using recent score statistics. These three settings were applied to evaluate whether operational control of false alarms improved when thresholding became more adaptive. Temporal persistence rules were then applied to convert point-wise anomaly scores into event-oriented alarms, reducing isolated detections and making comparisons more consistent with practical monitoring use [6].

The reproducibility protocol followed the same software and validation procedure across all experiments. The final code state passed linting, unit tests, and static type checking.

The seeded main comparison, the calibration analysis, the matched-false-alarm experiments, and the Water Distribution case study were all executed from scripted pipelines. The reduced-budget horizon and window sweep on the Water Distribution dataset was treated explicitly as exploratory analysis and was not used to redefine the main seeded comparison. This separation was maintained to preserve a clear distinction between core comparative results and targeted follow-up diagnosis.

5. Results and Discussion

5.1. Low-false-alarm Evaluation Changed Detector Ranking

The main seeded comparison showed that detector ranking depended strongly on the evaluation regime. Under low-false-alarm constraints, the best detector differed across datasets. On the Spacecraft Monitoring and Anomaly Processing dataset, the Temporal Autoencoder achieved the best recall@FAR ≤ 0.05 , reaching 0.0582. On the Mars Science Laboratory dataset, the same detector also achieved the best low-false-alarm result, reaching 0.1062. On the Secure Water Treatment dataset, the Gated Recurrent Unit Forecaster clearly outperformed the other seeded models in the low-false-alarm regime, reaching recall@FAR ≤ 0.05 of 0.6448. On the Water Distribution dataset, the seeded conformal multisensor detector remained the most conservative operating-point model among the main runs, with empirical FAR of 0.3964 and recall@FAR ≤ 0.05 of 0.0241.

These results became more informative when compared with aggregate ranking metrics. Across all model-dataset pairs, seven ranking inversions were observed between precision-recall area and recall@FAR ≤ 0.05 . The strongest inversion occurred on the Mars Science Laboratory dataset, where the Temporal Autoencoder moved from third place under precision-recall area to first place under low-false-alarm recall. This result showed that aggregate ranking quality did not necessarily identify the detector that was

most useful under strict false-alarm budgets. In other words, the detector that ranked anomalies well on average was not always the detector that retained useful sensitivity when false alarms had to remain rare.

This effect was also reinforced by the matched-false-alarm comparison. When detectors were compared at the same FAR budget, event-level conclusions changed relative to unconstrained comparisons. For example, under a FAR budget of 0.01, the best event-level F1-score was obtained by the conformal multisensor detector on the Water Distribution dataset, while on the Mars Science Laboratory dataset the Temporal Autoencoder exceeded the conformal multisensor detector under the same budget. These results indicated that event-level performance should not be interpreted independently of the false-alarm cost required to achieve it. Overall, the evidence supported the central claim of the paper: low-false-alarm evaluation changed detector ranking and provided a more realistic basis for model selection in safety-critical monitoring.

5.2. Calibration Improved Operational Control but not Score Separability

The calibration study showed that thresholding strategies affected detector operation, but their ability to improve final performance was limited by the quality of the anomaly scores. Fixed and split calibration produced nearly identical behavior in several settings, indicating that simple threshold selection and held-out quantile calibration often led to similar operating points when score distributions were already stable. Rolling calibration provided a more adaptive threshold and improved false-alarm control in some of the more difficult settings, especially on the Mars Science Laboratory and Water Distribution datasets. However, these improvements often came at the cost of a substantial drop in recall.

The Water Distribution dataset illustrated this trade-off clearly in Table III. With rolling calibration and a target of 0.01, the detector reduced empirical FAR to 0.0677, which was lower than the baseline operating point but still above the target, while recall dropped to 0.0459. This result showed that calibration could move the operating point toward a more conservative regime, but it could not create the separability needed to preserve high sensitivity at low false-alarm levels. The same interpretation was supported by the broader low-false-alarm analysis. In that dataset, threshold and smoothing tuning reduced false alarms substantially in some operating points, but the strongest gains in $\text{recall@FAR} \leq 0.05$ and restricted precision-recall area only appeared after the underlying score distribution was improved through model and grouping changes.

The calibration study therefore clarified an important distinction. Calibration improved operational control by adjusting the decision threshold to a desired alarm budget, but it did not fundamentally alter how well normal and anomalous samples were ranked. This distinction explained why some models remained weak even after careful threshold tuning, and why matched-false-alarm evaluation was necessary to separate threshold effects from score quality. For safety-critical anomaly detection, this meant that post-hoc calibra-

tion should be treated as an operational layer built on top of score generation, not as a substitute for discriminative temporal and multisensor modeling.

5.3. WADI Case Study: Backbone and Subsystem-aware Grouping Mattered Most

The Water Distribution dataset provided the most informative case study because it exposed the limitations of threshold tuning and the importance of structural modeling. According to Table IV, the baseline seeded conformal multisensor detector was highly conservative, with FAR of 0.3964, event-level F1-score of 0.2932, $\text{recall@FAR} \leq 0.05$ of 0.0241, and restricted precision-recall area of 0.0205. A first round of threshold and smoothing tuning showed that false alarms could be reduced substantially, but it also showed that sensitivity in the low-false-alarm regime remained weak. This result motivated a second round of experiments focused on score separability rather than calibration alone.

The ablation results showed that simplifying the score within the original temporal convolutional backbone did not solve the problem. Prediction-only and mixed prediction-reconstruction variants remained weak in the low-false-alarm regime, and the reconstruction-only temporal convolutional variant improved event-level F1-score only slightly while degrading restricted low-false-alarm ranking. In contrast, the backbone benchmark showed that a prediction-reconstruction Gated Recurrent Unit model improved low-false-alarm ranking substantially, reaching $\text{recall@FAR} \leq 0.05$ of 0.1037 and restricted precision-recall area of 0.0985. This result indicated that temporal representation had a large effect on score quality, even before calibration was considered.

The strongest result of the case study came from subsystem-aware grouping. The full conformal multisensor model with manual subsystem grouping achieved $\text{recall@FAR} \leq 0.05$ of 0.1838 and restricted precision-recall area of 0.1796, representing the best low-false-alarm ranking obtained on this dataset. Additional component analysis showed that this gain did not come primarily from the isolated cross-sensor term. Instead, the best grouped configuration improved the reconstruction component and the total score more than the cross term alone.

This result suggested that grouping was valuable not simply because it added one more score component, but because it aligned the model structure with the physical organization of the monitored system [4], [5]. The broader domain comparison reinforced this interpretation: adding the cross term produced large gains in the Water Distribution dataset, negligible gains in the Spacecraft Monitoring and Anomaly Processing and Mars Science Laboratory datasets, and mixed effects in the Secure Water Treatment dataset.

The cross-domain comparison in Table V also revealed an important failure mode. On the Spacecraft Monitoring and Anomaly Processing dataset, adding the cross-sensor component reduced Event-F1 by 0.3182—a collapse from 0.318 to effectively zero. This degradation resulted from the combination of two factors specific to that dataset. First, anomalous events in the Spacecraft Monitoring and Anomaly Processing dataset are sparse and short, affecting only a

Table III. Calibration ablation on the same checkpoints. Best empirical FAR control per dataset and target budget is reported.

Dataset	Method @ target FAR	Empirical FAR	Recall	Event-F1	Mean Time to Detection
SMAP	fixed @ 0.01	0.0001	0.0010	0.3093	381.8
SMAP	split @ 0.05	0.0001	0.0005	0.2540	255.8
MSL	rolling @ 0.01	0.0007	0.0220	0.1481	35.3
MSL	rolling @ 0.05	0.0004	0.0300	0.1633	18.5
SWaT	split @ 0.01	0.0143	0.6066	0.0033	1027.0
SWaT	rolling @ 0.05	0.0808	0.1438	0.0004	0.0 [†]
WADI	rolling @ 0.01	0.0677	0.0459	0.0497	374.7
WADI	rolling @ 0.05	0.1610	0.1091	0.0402	247.3

[†] A mean time to detection of 0.0 does not indicate instantaneous true detection. Rolling calibration on this dataset produced persistent alarms that preceded event onset; consequently, every event was “detected” at zero delay by a pre-existing false alarm. The Event-F1 score of 0.0004 confirms that precision was negligible at this operating point.

Table IV. WADI case study. Structural changes improved low-FAR score separability more than threshold tuning alone.

Variant	Backbone / Grouping	FAR	Recall	Event-F1	Recall@FAR ≤ 0.05	rPR-AUC
Baseline CMAD-LFA	TCN / manual	0.3964	0.5595	0.2932	0.0241	0.0205
Prediction only	TCN / manual	0.3970	0.5572	0.2030	0.0240	0.0205
Reconstruction only	TCN / manual	0.4398	0.5676	0.3083	0.0191	0.0158
Prediction + reconstruction	TCN / manual	0.3970	0.5589	0.2667	0.0090	0.0058
Prediction + reconstruction	GRU / manual	0.4453	0.5579	0.2240	0.1037	0.0985
Full model	TCN / subsystem-aware	0.3964	0.5595	0.3006	0.1838	0.1796

Table V. Utility of adding the cross-sensor component by domain. Deltas are computed relative to the corresponding pred+recon configuration.

Dataset	Grouping	Δ FAR	Δ Recall@FAR ≤ 0.05	Δ rPR-AUC	Δ Event-F1
SMAP	correlation	-0.0000	+0.0003	+0.0001	-0.3182
MSL	correlation	+0.0000	+0.0009	+0.0001	-0.0035
SWaT	manual	+0.0047	+0.0059	-0.0211	+0.0001
WADI	manual_subsystem	-0.0006	+0.1748	+0.1738	+0.0339

subset of channels at any given time. Second, the cross-sensor grouping in that dataset was derived from statistical correlation rather than physical subsystem structure. A correlation-based grouping captures joint behavior under normal operating conditions; during anomalies that violate channel independence, the cross-sensor term penalizes precisely the deviations that should be flagged. The result is that the cross-sensor score suppresses true detections rather than reinforcing them. This observation generalizes the finding from the Water Distribution case study: *domain-inappropriate grouping degrades performance more than omitting the cross-sensor component entirely*. The cross-sensor term is beneficial only when the grouping reflects the physical organization of the monitored system.

Taken together, these findings showed that cross-sensor modeling was not universally beneficial; its usefulness de-

pended on domain structure and grouping quality. In the most difficult dataset, backbone choice and subsystem-aware grouping mattered more than threshold tuning, confirming that separability was the main bottleneck.

5.4. Limitations and Practical Implications

The study had some limitations that should be considered when interpreting the results. The seeded main comparison was intentionally kept compact and did not exhaust all possible model and hyperparameter combinations. In the Water Distribution dataset, the additional horizon and window sweep was conducted as an exploratory analysis with reduced training budget and was therefore not used to redefine the main comparative ranking. In addition, the validation regime in that dataset remained normal-only, which allowed false-

alarm-oriented threshold selection without using anomalous test labels, but limited the ability to optimize sensitivity without risking information leakage. Finally, although the study analyzed the contribution of grouped multisensor modeling, the best result in the Water Distribution dataset depended on subsystem-aware grouping, which introduced domain knowledge not equally available in every application.

Despite these limitations, the practical implications of the results were clear. Safety-critical anomaly detection systems should be evaluated under explicit false-alarm budgets rather than only with aggregate ranking metrics [4], [8]. Calibration should be treated as an operational control mechanism, not as a substitute for score quality [6]. When multisensor dependencies are relevant, structural information about subsystems or sensor organization can significantly improve low-false-alarm performance, especially in challenging domains [4], [5]. These observations suggest that future work should focus on operationally constrained evaluation, domain-structured modeling, and robust score generation before relying on threshold adaptation alone.

6. Conclusion

This study investigated multisensor anomaly detection under low false-alarm constraints and showed that operationally constrained evaluation led to conclusions that differed from those obtained with aggregate metrics alone. Across four public multivariate sensor datasets, detector rankings changed when the analysis emphasized recall under limited false-alarm budgets rather than global precision-recall behavior. Seven ranking inversions were observed between precision-recall area and $\text{recall@FAR} \leq 0.05$, with the strongest case on the Mars Science Laboratory dataset, where the Temporal Autoencoder moved from third place under aggregate ranking to first place in the low-false-alarm regime. This result demonstrated that conventional metrics were not sufficient to support model selection in safety-critical settings.

The study also showed that calibration improved operational control of false alarms but did not compensate for weak score separability. Fixed and split calibration often behaved similarly, while rolling calibration reduced false alarms in some settings at the cost of lower recall. The Water Distribution case study clarified this limitation further. In that dataset, the main gains in low-false-alarm ranking were obtained only after changes in backbone architecture and subsystem-aware grouping, rather than through threshold tuning alone. The best grouped configuration increased $\text{recall@FAR} \leq 0.05$ from 0.0241 to 0.1838 and restricted precision-recall area from 0.0205 to 0.1796, indicating that structural alignment between the model and the monitored system was essential in difficult domains.

Taken together, these findings supported three conclusions. First, low-false-alarm evaluation provided a more realistic basis for detector comparison in safety-critical multisensor monitoring. Second, matched-false-alarm comparison was necessary to interpret event-level performance fairly. Third, the utility of cross-sensor modeling depended on domain

structure and grouping quality rather than being universally beneficial across all datasets.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

Author 1: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review and editing. **Author 2:** Writing – review and editing. All authors have read and agreed to the published version of the manuscript.

Data Availability

The data and source code that support the findings of this study are openly available in the Low False Alarm repository at <https://github.com/murilo-salem/LowFalseAlarm>.

Funding Information

This research received no external funding.

Acknowledgment

The authors would like to thank the Federal University of Pelotas (UFPel), the Center for Technological Development (CDTec), and the Artificial Intelligence Research Group (GAIA) for their infrastructure, technical support, and valuable contributions to this work.

References

- [1] L. Erhan, M. Ndubuaku, M. Di Mauro, W. Song, M. Chen, G. Fortino, O. Bagdasar, A. Liotta, "Smart anomaly detection in sensor systems: A multi-perspective review," *Information Fusion*, vol. 67, pp. 64–79, 2021.
- [2] R. J. Simonson, J. R. Keebler, E. L. Blickensderfer, R. Besuijen, "Impact of alarm management and automation on abnormal operations: A human-in-the-loop simulation study," *Applied Ergonomics*, vol. 100, p. 103670, 2022.
- [3] R. Wiczorek A. Zirk, "Using warning systems with adaptable thresholds: Choice of security level, compliance and performance in a simulated computer security task," *International Journal of Human-Computer Studies*, vol. 125, pp. 32–40, 2019.
- [4] F. Wang et al., "A survey of deep anomaly detection in multivariate time series," *Sensors*, vol. 25, no. 1, p. 190, 2025.
- [5] B. Xie, Zhang, "Multivariate time series anomaly detection by capturing coarse-grained intra- and inter-variate dependencies," 2025. [Online]. Available: <https://arxiv.org/abs/2501.16364>
- [6] S. Gloumeau et al., "Post-hoc calibrated anomaly detection," *arXiv preprint arXiv:2503.19577*, 2025. [Online]. Available: <https://arxiv.org/abs/2503.19577>
- [7] D. S. Castilho, "Active STPA: Integration of hazard analysis into a safety management system framework," Ph.D. dissertation, Massachusetts Institute of Technology, 2019. [Online]. Available: <https://dspace.mit.edu/handle/1721.1/124172>

- [8] A. Mueller, "Open challenges in time series anomaly detection: An industry perspective," 2025. [Online]. Available: <https://arxiv.org/abs/2502.05392>
- [9] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, T. Soderstrom, "Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018, pp. 387–395.
- [10] iTrust, Singapore University of Technology and Design, "iTrust labs: Dataset info," 2026, accessed 2026-04-05. [Online]. Available: https://itrust.sutd.edu.sg/itrust-labs_datasets
- [11] C. M. Ahmed, V. R. Palleti, A. P. Mathur, "WADI: A water distribution testbed for research in the design of secure cyber-physical systems," in *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks*, 2017.