

Author comment on wes-2026-145: response to Referee #1 and Referee #2

Manuscript: wes-2026-145. Discussion-version title: *Chronologically calibrated PCA for wind-turbine SCADA health monitoring: a leakage-controlled CARE v6 benchmark*. Proposed revised title: *Chronologically calibrated anomaly detectors for wind-turbine SCADA health monitoring: a CARE v6 benchmark* (see summary point 2).

Authors: Jui-Hung Liu, Jien-Chen Chen, Chun-Chieh Wang

We thank both referees for their careful and constructive reviews. The two reports agree on four points: the writing and terminology must be clarified; the candidate configurations and the methods must be described so that they can be reproduced; the paper must be positioned clearly, either as a controlled benchmark or as a methodological contribution, with the contributions stated accordingly; and the comparison set must either include state-of-the-art normal-behaviour-model (NBM) baselines or justify their absence. We have addressed all four. In addition, while preparing the revision we commissioned an independent audit of our evaluation code, which found that our scoring rule departed from the CARE reference implementation in a way that affected Farm A; we disclose the resulting post-hoc amendment first, because it changes the numbers quoted throughout this response. The main changes, described in detail in the point-by-point responses below, are:

0. **Scoring rule aligned with the reference implementation, disclosed as a post-hoc amendment, and verified against it.** Our evaluation code excluded every timestamp whose CARE status identifier was outside $\{0, 2\}$ in all three farms. For Farm A the status identifiers mark the 14 days before each fault (Gück and Roelofs, 2025), so this rule removed every window timestamp of 5 of 5 Farm A test anomaly cases and 6 of 7 Farm A training-validation anomaly cases: the “zero detections in Farm A” of the discussion version was an artefact of our rule, not a property of the detectors. On 4 October 2026, after the test results under the original rule had been seen, we recorded an amendment (new Sect. 3.3, Table 3) that adopts the behaviour of the reference package EnergyFaultDetector v0.7.1 in every respect in which the two differed (Farm A evaluable timestamps; criticality accumulated up to the annotated event end; earliness weights on all window timestamps with a quarter-window plateau; accuracy fallback at ≤ 0.5), redid the model selection on the training-validation turbines under the amended rule without changing the split, the blocks, the candidate grid, the α grid, the seeds or the selection criterion, and evaluated the test turbines once more. Every reported score was recomputed by the reference package on identical per-timestamp alerts and agreed to machine precision (13 490 case-level metric sets and 284 aggregate scores). Because the amendment was made after test results had been inspected, we call the revised numbers a post-hoc re-evaluation, and we report the complete original-rule results in the Supplement (Tables S1–S3). We also no longer describe the protocol as pre-registered; Sect. 3.3 gives its actual history.
1. **Two NBM baselines added under the same blocks, calibration and selection rule.** An autoencoder NBM (reconstruction error, multilayer-perceptron autoencoder) and a regression NBM (multi-output network predicting drivetrain, generator and transformer temperatures from operating variables) were evaluated with the same chronological blocks, the same split-conformal calibration, the same α grid and the same selection rule on training-validation turbines only. On the held-out test turbines they scored 0.544 (95 % turbine-level cluster-bootstrap interval 0.427–0.653) and 0.606 (0.496–0.689), compared with 0.569 (0.474–0.665) for the selected detector; the paired differences were 0.026 (−0.033 to 0.096) and −0.026 (−0.122 to 0.074), both including zero, while both baselines exceeded the fixed isolation-forest baseline (0.252) with intervals excluding zero. With 13 test turbines these intervals are wide, and we draw no conclusion of equivalence from them (Sect. 5.3). The recomputation from the exported per-case blocks reproduced the stored case-level metrics of the first two stages for all 11 590 case-configuration pairs (identical alert counts, event decisions and maximum criticality; pointwise metrics within 1.1×10^{-16}). The new baselines are described in Sect. 4.5 and Appendix B and reported in Sects. 5.2–5.5, Tables 4–5 and Figs. 2–3; the NBM baselines of the first revision were recomputed from double-precision exports and, for the regression NBM, with the channel lists recorded by the first full run, which changed their values slightly (Appendix D, Table D2).

2. **Contributions rewritten.** The paper is now explicitly positioned as a leakage-controlled benchmark study. The former Contribution 4 (“presentation of results”) has been removed; negative and null findings are reported as results. The contributions now are (i) the leakage-controlled protocol for CARE v6 with its documented four-stage history (Sect. 3.3, Appendix D), (ii) a comparison that varies the representation and the threshold rule, reported configuration by configuration (eleven of the twelve configurations compared with the CARE baseline exceeded it with intervals excluding zero, including two with fitting-block thresholds; the physics-guided family did not) together with the matched- α false-alarm effect of the block on which the PCA threshold is estimated, while the comparisons among calibrated representations remain inconclusive; the design does not separate the roles of calibration-data recency, fitting–calibration separation and representation, so the manuscript does not claim to have identified the source of the gain, (iii) an explicit uncertainty-quantification framework with the turbine as resampling unit, and (iv) a channel-frequency analysis of the alerts read case by case and reported as the authors’ interpretation (Sect. 1). Under the amended rule the selection criterion fixed in the first stage chose, from the pre-specified grid, a robust-deviation score (mean of the three largest absolute standardised channel deviations) with $\alpha = 0.020$ and Mondrian calibration rather than PCA; the best PCA configuration was within 0.006 on the training–validation turbines and the paired test difference includes zero (0.569 versus 0.589 for PCA at the same α and mode; interval -0.084 to 0.042). The manuscript reports the selected detector as the primary result and PCA as the reference representation, and Sect. 6.1 decomposes the change in the headline score into the part due to the rule ($+0.038$ on identical alerts) and the part due to re-selection ($+0.025$). Because the selected detector is no longer PCA, we propose to change the title to *Chronologically calibrated anomaly detectors for wind-turbine SCADA health monitoring: a CARE v6 benchmark*.
3. **Interpretability examined case by case rather than asserted.** A new Sect. 5.6 (Fig. 5, Table 6, Fig. S3) reports, for all eleven detected test events and all eight false alarms of the selected detector, the channel selection frequency, that is, the fraction of alert timestamps at which each channel was among the three largest standardised deviations. In our reading, for four detections a leading channel measures the annotated component or its immediate supply (hydraulic-oil temperature for a hydraulic-group event; pitch-battery charge and voltages for pitch-axis, hub-wiring and pitch-battery events); for three more the leading channels could plausibly be connected to it, which we state as a hypothesis; and for four, including both Farm B bearing events, the leading channels appear to reflect a standstill signature or belong to unrelated subsystems, so these detections may be correct in timing but not in mechanism. The false alarms concentrate in one subsystem per case. The classification is descriptive and has not been validated against maintenance records; the manuscript no longer claims that interpretability has been demonstrated.
4. **Terminology and language.** “Development” $\rightarrow$ “training–validation turbines”; “comparator” $\rightarrow$ “baseline”; “view” $\rightarrow$ “score family”; “weight configuration” $\rightarrow$ “fusion-weight setting”; “farm–turbine cluster bootstrap” $\rightarrow$ “turbine-level cluster bootstrap”; “physics-informed/physics view” $\rightarrow$ “physics-guided residual”. A terminology table (Appendix A, Table A1) maps every new term to the discussion-version term. The paragraphs singled out by Referee #2 (p. 6, lines 188–204) have been rewritten, Fig. 1 has been redrawn with a single caption, Table 1 now explains “fitting/calibration rows”, and the whole manuscript has been language-edited.
5. **Methods made reproducible.** A candidate-configuration table (Table 2) lists every family $\times$ score $\times$ pre-processing $\times$ threshold rule $\times$ hyperparameter grid; the fusion of score families is given as an explicit equation (Eq. 3) with the quantile mapping and the ten fixed weight vectors; the physics-guided residuals are specified with inputs, hyperparameters, minimum-sample rules and selection criteria (Sect. 4.4, Appendix B, Table B1).
6. **Design rationale and uncertainty quantification.** Section 2.2 now explains why split-conformal calibration was preferred over time-series conformal variants for unsupervised fault detection, and a new Sect. 4.8 and Sect. 6.3 state which sources of uncertainty the cluster bootstrap and the conformal rule each address, and which remain unquantified.

7. **Related work updated** with 2023–2026 SCADA NBM, benchmark, conformal-prediction and evaluation-methodology literature (Chesterman et al., 2023; Roelofs et al., 2024; Maldonado-Correa et al., 2024; Vervlimmeren et al., 2025; Zhu et al., 2025; Barber et al., 2023; Xu and Xie, 2021, 2023; Zaffran et al., 2022; Oliveira et al., 2024; Xu and Boström, 2026; Kapoor and Narayanan, 2023; Hewamalage et al., 2023; Bouthillier et al., 2021; Alcalá and Qin, 2009; Westerhuis et al., 2000).

A marked-up version showing all changes will be supplied with the revised manuscript. Section, table and figure numbers below refer to the revised manuscript; all scores quoted below are under the amended rule unless stated otherwise, and the corresponding original-rule values are in Supplement Tables S1–S2.

Response to Referee #1

General comments

The manuscript appears to be primarily a benchmark contribution rather than a methodological contribution. Such a contribution can be valuable, but the benchmark design, candidate configurations, method-selection rationale, and limitations need to be described much more clearly.

We agree with this characterisation and have adopted it. The revised Introduction states in its first sentence of the contribution paragraph that “this paper is a benchmark study, not a new detector”, and the four contributions are now formulated as protocol, controlled comparison, uncertainty framework and interpretability test (Sect. 1). The benchmark design is summarised in a redrawn Fig. 1, the candidate configurations in a new Table 2, the selection rule in Sect. 4.6, the design rationale in Sects. 2.2 and 4.8, and the limitations in Sect. 6.4.

Major comment 1 — writing quality and terminology

The manuscript requires substantial language editing. ... Terms and phrases, including “views”, and the extensive usage of dash-connected expression such as “farm-turbine”, should be defined clearly and used consistently.

The manuscript has been language-edited throughout. “View” has been replaced by “score family”, defined at first use in Sect. 4.4 (“one definition of a per-timestamp nonconformity score”). “Farm–turbine” has been removed; the resampling unit is now called the “turbine identity (farm and asset)” and the procedure the “turbine-level cluster bootstrap”. Other dash-connected expressions (“development-tuned”, “development-only”, “locked-test”) have been replaced by plain phrases (“validation-tuned”, “selected on training–validation turbines only”, “test turbines”). Appendix A (Table A1) lists every term with its definition and the discussion-version term it replaces, so that readers of the public discussion can map the two versions.

Major comment 2 — description of methods and evaluation candidates

It is currently difficult to determine which combinations of algorithms, views, thresholds, preprocessing steps, and calibration procedures constitute each evaluation candidate.

A candidate-configuration table (Table 2) now lists, for every family, the representation/score, the pre-processing, the threshold rule (fixed contamination, in-sample quantile, empirical calibration quantile, split-conformal rank, Mondrian variant), the hyperparameters in the grid and the number of candidates. The fusion grid (100 candidates), the six baseline families from the discussion version and the two new NBM families are all included. Section 4.6 states the selection rule (training–validation CARE score, ties broken by normal accuracy and event $F_{0.5}$, one configuration per family) and Fig. 1 shows where selection and test evaluation sit in the workflow. Appendix B (Table B1) gives inputs, fixed hyperparameters, selected hyperparameters and minimum-sample rules for each family.

Major comment 3 — rationale for focusing on PCA; interpretability not evaluated

If interpretability is a central advantage of the selected methods, the manuscript should demonstrate this advantage. The authors should also explain why the selected approaches provide an appropriate and sufficiently broad comparison set for complex wind-turbine systems.

We have done both. First, the comparison set has been broadened with two NBM baselines that represent the two dominant SCADA NBM designs, reconstruction (autoencoder) and regression (temperature NBM), evaluated under identical calibration (Sect. 4.5). Second, interpretability is now examined rather than asserted: Sect. 5.6 (new Fig. 5, Table 6 and Fig. S3) reports, for all eleven detected test events and all eight false alarms of the selected detector, the channel selection frequency at alert timestamps, compared with the same frequency over the calibration block (the per-channel PCA reconstruction-error terms at the same timestamps are in the reproducibility package). In our reading, for four detections a leading channel measures the annotated component or its immediate supply (hydraulic-oil temperature for the Farm A hydraulic-group event A-45; pitch-axis-3 battery charge for C-4; hub-side battery and supply voltages for the nacelle–hub wiring event C-55; axis-2 battery voltages for the pitch-battery event C-76); for three more (C-16, C-81, C-90) the leading channels could plausibly be connected to the annotated component, which we present as a hypothesis; and for four (B-27, B-77, C-9, C-31) the leading channels appear to reflect a standstill signature (reactive power, hub and outside temperature) or belong to unrelated subsystems, so these detections may be correct in timing but not in mechanism. For the false alarms the selected channels concentrate in one subsystem per case (24 V supply, hydraulic-oil level, a reactive-power set-point, or what we hypothesise to be seasonal thermal drift in Farm A). Section 5.6 states that this classification is a descriptive judgement by the authors on eleven events and has not been validated against maintenance records or by independent raters; within that limit the channel frequencies offer a triage aid and can flag detections that may be correct for the wrong reason, they do not improve detection, and analyst uptake was not tested. The rationale for PCA as the reference representation (deterministic, seconds per case, per-channel error terms that can be inspected) is given in Sects. 2.1 and 6.1, with the caveat that under the amended rule the selection criterion chose the robust-deviation score, and the motivation of “avoiding unnecessarily complex models” has been removed from the Introduction, which now acknowledges that there are good reasons for richer architectures (Sect. 2.4).

Major comment 4 — scope and contribution of the benchmark

The authors should explain more explicitly what additional knowledge the extension provides beyond previous benchmark versions. The contribution would be strengthened by either including well-motivated additional comparators, or by clearly justifying the restricted scope.

We chose the first option and added two well-motivated baselines (above). The additional knowledge relative to the CARE paper is now stated explicitly in Sects. 1, 2.1 and 6.1: the CARE paper compared a fixed isolation forest with a per-turbine, Optuna-tuned autoencoder without holding out turbine identities during tuning and reported a CARE score of 0.66 for the autoencoder; our protocol holds out 13 turbines from every selection decision, forms chronological calibration blocks within each case, and applies the same calibration procedure to every representation. The resulting findings are reported configuration by configuration: eleven of the twelve reported configurations compared with the fixed baseline exceeded it with paired intervals excluding zero, among them the in-sample PCA threshold (bootstrap mean difference 0.170, observed 0.168) and the validation-tuned isolation forest (0.298, observed 0.299), whose thresholds come from the fitting block, while the physics-guided family did not (bootstrap mean difference 0.203, observed 0.258; interval -0.306 to 0.422); and at matched α the PCA threshold estimated on the calibration block produced 3 and 4 false-positive events against 12 and 15 for the fitting-block threshold. The manuscript does not attribute the improvement to a single element, because the design does not separate the roles of calibration-data recency, fitting–calibration separation and representation. The paired differences among the calibrated representations have wide intervals that include zero, except for one small, unadjusted comparison with the conformalised isolation forest that is confounded with α ; we do not read this as equivalence or as evidence that the representation is unimportant (Sects. 5.3, 6.1). This is knowledge the previous benchmark version could not provide. It is consistent with the transfer-learning study of Roelofs et al. (2024), in

which autoencoders transferred to a new turbine and adjusted only through a re-calibrated threshold already detected the anomalies preceding faults; we note, to avoid overstating the agreement, that in that study fine-tuning the decoder improved performance further, so the representation is not irrelevant in general. The manuscript contains no new methodological component and now says so.

Technical comments

Pages 6–7, lines 158–204 — evaluation candidates (table or flowchart). Added as Table 2 (candidate configurations) and reflected in the redrawn Fig. 1, panels (d)–(f).

Page 6, lines 193–199 — combination of views; is the physics view a 50/50 combination; how are the weights selected? Section 4.4 now gives the procedure explicitly. Each score family is first mapped to the probability scale of its own calibration block (self-quantile rank/($n + 1$) for calibration scores; empirical quantile u_v for prediction scores), and the fused score is the weight-normalised average over the families available at that timestamp (Eq. 3). The physics-guided family is itself the mean of its thermal and power-curve components where both are available. The ten weight vectors are listed; they were fixed a priori, were candidates in the selection, and were not tuned on data. The thermal score is the mean of the two largest standardised residuals across up to ten temperature targets.

Page 6, lines 195–198 — why ridge and histogram gradient boosting; in what sense physics-informed; inputs, hyperparameters, training, selection criteria. We have renamed the family “physics-guided residual” and state in Sect. 4.4 that it “uses domain knowledge only to choose the structure of empirical regressions ... and contains no first-principles model”. Ridge regression was used for the one-step thermal model because the lagged target makes the model a linear first-order thermal response whose coefficients remain interpretable; gradient boosting was used for the power curve because the wind–power relation is strongly nonlinear and the model is small (100 iterations, 15 leaves). Inputs, hyperparameters, minimum-sample rules and the selection criteria are tabulated in Appendix B (Table B1).

Page 12, line 347 — statement about physics-informed effectiveness too broad. The statement has been narrowed in Sect. 5.5 (“This result concerns only the specific residual construction tested here ... it says nothing about physics-informed methods in general”) and in Sect. 6.2 (“the physics-guided residual family tested here did not improve on generic multivariate scores, which says nothing about physics-informed monitoring in general”). Under the amended rule the family detected 2 of 18 test events and scored 0.510 rather than zero, which the text now reports.

Response to Referee #2

General assessment

We are grateful for the assessment that the dataset, the CARE score, the conformal thresholding and the bootstrap uncertainty quantification are positive elements, and we have tried to turn the four shortcomings identified into concrete revisions.

Text quality

“Comparator” (p. 7, line 213). Replaced throughout by “baseline”; the sentence now reads “The fixed isolation forest reproduced the CARE paper’s simple baseline ... it is the pre-specified primary baseline” (Sect. 4.5).

“Development-tuned”, “development-only selection”, “development” (p. 4, line 146). “Development” has been removed from the manuscript. The 23 turbines are now the “training–validation turbines”, used to fit per-case models and to select one configuration per family, and the 13 held-out turbines are the “test turbines” (Sect. 3.2; Table A1). We kept the compound “training–validation” rather than separate “training” and “validation” sets

because in CARE each case file already carries a “train/prediction” designation, and within each case we further distinguish fitting, calibration and prediction blocks; the compound avoids a second meaning of “training”.

P. 6, line 188 (“The term conformal is used here for the split-conformal rank correction ...”). Rewritten in Sect. 4.3: “The word *conformal* is used here only for the finite-sample rank rule of Eq. (2). Under exchangeable calibration and test scores this rule bounds the per-timestamp false-alarm probability by α ; SCADA scores are autocorrelated and drift, so that bound is not claimed.”

P. 6, line 190 (“Operational test metrics therefore remain the evidentiary endpoint”). The referee’s interpretation is exactly what was meant, and the sentence now says so: “The consequence for this study is that the theoretical guarantee cannot replace empirical validation: the operational metrics of Sect. 4.7, computed on later prediction blocks of held-out turbines, are the quantities on which conclusions rest” (Sect. 4.3).

P. 6, line 200 (“Ten weight configurations”). The ten settings are indeed weight vectors for combining the three score families, not ten pipelines; the term is now “fusion-weight setting”, defined with Eq. (3) and listed explicitly in Sect. 4.4 and Table 2.

P. 6, line 192 (whole paragraph). The former Sect. 4.4 has been rewritten as the new Sect. 4.4 with the explicit quantile mapping, Eq. (3), the weight list and the renamed physics-guided family.

Figure 1 (two captions; “Frozen before evaluation ...” unexplained). Figure 1 has been redrawn with six annotated panels (a)–(f), a single caption that explains each panel, and a footer that lists the protocol elements recorded in the hashed protocol file and refers to Table 1 where they are tabulated. The word “frozen” has been replaced by a precise statement of what was fixed when: the split rule, block sizes, the fusion grid, α grid, selection rule and bootstrap seed were written to the protocol file at the start of the first full run (9 August 2026); the baseline families were added in two later stages (same day; 3 October 2026) under the same selection rule after the first results existed; and the scoring rules were amended post hoc on 4 October 2026 (Sect. 3.3, Appendix D). Test-turbine outcomes were summarised once per stage after selection.

Table 1 (caption on a different page; “Maximum fit/calibration rows”). The caption now reads “Maximum fitting / calibration rows are the caps on the sizes of the chronological fitting and calibration blocks defined in Sect. 4.1”, and the blocks are defined in Sect. 4.1 and Table A1. Caption and table are kept together in the revised file (keep-with-next formatting); we will check the page layout again in the typeset proof.

Novelty

Motivation (p. 1, line 37) and reproducible benchmarking (p. 2, line 40) are not contributions. We agree. The Introduction now says that the false-alarm objective “is shared by most condition-monitoring studies, and we treat it as the evaluation target rather than as a contribution”, and it credits CARE for reproducibility while identifying the two leakage paths that the dataset leaves open to its users (tuning on evaluated turbines; thresholds estimated on training observations).

Complex models (p. 4, line 119). The sentence has been rewritten to acknowledge that “there are good reasons for this: turbine behaviour is nonlinear and regime-dependent, and richer models can capture it”, before stating that complexity alone does not establish operational benefit (Sect. 2.4).

Comparison against state-of-the-art NBMs. Two NBM baselines have been added and are compared under the same blocks, calibration and selection rule (summary point 1 above; Sects. 4.5, 5.2–5.4, Tables 4–5, Figs. 2–3). We note two deliberate limits, stated in Sects. 4.5 and 6.4: the baselines are compact per-case models with fixed hyperparameters and a single seed, because that is the setting in which the PCA detector is evaluated, and they are therefore not tuned to the extent of the CARE paper’s per-turbine, Optuna-tuned autoencoder (CARE score 0.66 without identity hold-out). Within that setting: the autoencoder NBM (0.544) and the regression NBM (0.606) have paired differences from the selected detector (0.569) and from PCA at the same α (0.589) whose intervals include zero, both beat the fixed baseline (0.252) with intervals excluding zero, and both share the Farm A difficulty (1 of 5 and 0 of 5 Farm A events detected; three Farm A events were missed by every one of the 142

candidate configurations). With 13 test turbines these comparisons are exploratory and unadjusted, and they neither rank the representations nor establish equivalence.

Contribution 1 (temporal splitting is standard). Agreed; the revised Contribution 1 says that “neither element is new in isolation, but their joint application to CARE, with the fusion grid and its selection rule fixed before any test-turbine result was inspected, has not been reported”, and it cites the leakage literature (Kapoor and Narayanan, 2023; Hewamalage et al., 2023) for why this matters. We have also removed the word “pre-registered”: Sect. 3.3 now gives the protocol’s actual history (hashed protocol file at the start of the first full run; documented later extensions; the post-hoc scoring amendment of 4 October 2026).

Contribution 2 (“chronologically calibrated” undefined; PCA not novel; conformal aspect undeveloped). “Chronologically calibrated” is now defined in Sect. 4.1 (“We call a threshold *chronologically calibrated* when it is estimated from a calibration block that follows the fitting block in time and is disjoint from it”) and in Table A1. The revised Contribution 2 claims no novelty for PCA; it claims a comparison in which representation and threshold rule are varied so that their effects can be examined, which the new baselines extend; the manuscript now states explicitly that the design is not fully crossed and that the comparisons among calibrated representations are inconclusive (Sect. 5.3). The role of the conformal rule is developed in Sects. 2.2, 4.3, 4.8 and 6.3 (below).

Contribution 3 (uncertainty quantification potentially the most novel; sources not distinguished). This has become the core of a new Sect. 4.8 and Sect. 6.3. Two sources are now treated separately. *Within-case threshold uncertainty*, the finite calibration sample behind each quantile, is what the split-conformal rank rule addresses under exchangeability; its practical effect is tested directly for the global PCA threshold and found to be nil at the calibration size and α grid used (identical alerts to the empirical quantile in all 95 cases and at all five α values, which Sect. 5.3 shows follows from $n = 6000$ and the grid); the statement is limited to global thresholds and does not extend to the Mondrian calibration of the selected detector, nor does it say anything about the effect of serial dependence or drift, which Sect. 6.3 now lists as untested. *Sampling uncertainty of the benchmark summary*, i.e. which turbines, with their events and normal periods, were observed, is what the paired turbine-level cluster bootstrap addresses; it preserves within-turbine dependence between cases, recomputes the CARE score with its special cases for each replicate, and is probed by farm-stratified resampling and leave-one-turbine-out influence. Section 6.3 then states what remains unquantified: uncertainty in the CARE event annotations, the single realisation of the identity split, and the training randomness of the neural baselines.

Contribution 4 (presentation of results is not a contribution; negative results unsurprising). Removed. Negative and null findings are now reported as results (Sects. 5.3 and 5.5) and interpreted in Sect. 6.2. We agree that the limited sensitivity of a conformally thresholded PCA is not surprising in itself; what we consider informative is that the compact autoencoder and regression NBMs under the same calibration landed in the same range as PCA and the selected robust-deviation detector, with paired differences whose intervals include zero, and that the Farm A difficulty is shared by every representation once the scoring rule evaluates Farm A at all.

Related work

Section 2 has been updated with recent SCADA NBM work (Chesterman et al., 2023; Roelofs et al., 2024; Maldonado-Correa et al., 2024; Vervlimmeren et al., 2025; Zhu et al., 2025), conformal prediction for dependent data (Barber et al., 2023; Xu and Xie, 2021, 2023; Zaffran et al., 2022; Oliveira et al., 2024; Xu and Boström, 2026), evaluation methodology (Kapoor and Narayanan, 2023; Hewamalage et al., 2023; Bouthillier et al., 2021) and contribution-plot diagnostics (Alcala and Qin, 2009; Westerhuis et al., 2000). The CARE paper’s own autoencoder baseline is now described (Sect. 2.1).

Discussion of methodology and justification of methodological choices

Why conformal prediction with PCA; why not time-series conformal variants or other frameworks. Section 2.2 now explains that EnbPI, adaptive conformal inference and weighted conformal prediction are designed for forecasting, where the realised value becomes available after each step and the interval can be adapted, whereas in unsupervised early fault detection the ground truth is not revealed online, so adaptation would have to be driven

by the detector’s own alerts. Split-conformal calibration was therefore used as the simplest label-free, pre-specifiable threshold rule, supported by recent results on its approximate validity under mixing (Oliveira et al., 2024) and by its use for leakage-free anomaly thresholding (Xu and Boström, 2026). Whether the finite-sample correction changes any alert was then checked for the global PCA threshold (Sect. 5.3). Section 6.3 now states that whether serial dependence and drift limit the present thresholds in practice was not tested and remains an open question for this benchmark.

Why clustered bootstrap; which uncertainty sources. Section 4.8 states that the turbine is the sampling unit because several turbines contribute more than one case, so case-level resampling would pseudo-replicate; that the paired design is needed because the quantity of interest is a difference between methods on the same cases; and that percentile intervals of the recomputed CARE score, including its discontinuities, are reported. The alternatives considered (case-level bootstrap, which ignores within-turbine dependence, and permutation tests, which do not give an interval for the difference) are mentioned, and the sources the bootstrap does not capture are listed (Sect. 6.3).

Recommendations

All five recommendations have been acted on: additional experiments with NBM baselines (Sects. 4.5, 5.2–5.4); a strengthened uncertainty-quantification treatment distinguishing bootstrap and conformal sources (Sects. 4.8, 6.3); a thorough language and terminology revision (throughout; Appendix A); an expanded and updated related-work section (Sect. 2); and a fuller explanation of the methodology (Sects. 4.4–4.8, Table 2, Appendix B).

Additional changes made by the authors. (a) The scoring amendment of summary point 0: new Sect. 3.3 and Table 3 document the protocol history, the amendment and the time boundary of normal cases; Sect. 4.7 now describes the CARE evaluation under the amended rules and states the verification against the reference package; Sect. 2.3 corrects our description of the CARE score (the reference implementation uses a quarter-window earliness plateau and does not mask Farm A timestamps). (b) The Supplement contains the complete original-rule results (Tables S1–S3), the implementation differences (Table S4), the verification record (Table S5), the α sensitivity under the amended rule (Fig. S2) and the false-alarm contribution analysis (Fig. S3). (c) The CARE paper’s autoencoder baseline is now described as a per-turbine model (Sects. 2.1, 4.5, 6.4). (d) In the discussion version we summarised Roelofs et al. (2024) inaccurately as showing that threshold re-calibration matters more than adapting the network; the revised Sects. 2.1 and 6.1 describe that study’s findings as published (threshold-only transfer already detected the anomalies, decoder fine-tuning improved performance further, multi-source training gave no gain). (e) The NBM baselines were re-run from double-precision exports of the per-case blocks, and the regression NBM now uses the channel lists recorded by the first full run instead of lists re-derived from the exported block; Supplement Table S1 and Appendix D report both sets of values and do not attribute the difference to precision alone. (f) A new Appendix D gives the four-stage version history of the analysis and a change log of every difference between the first revision and this version that affects a reported number, including the regression NBM’s change of channel lists and the change of export precision. (g) The acknowledgements now thank the referees, the maintainers of the reference implementation and the independent auditor. (h) We propose the revised title *Chronologically calibrated anomaly detectors for wind-turbine SCADA health monitoring: a CARE v6 benchmark*, because the configuration selected under the amended rule is not PCA. The reproducibility package has been extended with the per-case block exporter, the amended-rule scoring and cross-validation code, the reproduction check and the contribution analysis (Appendix C). Numbering: Fig. 5 (contribution analysis) is new; Tables 2 (candidate configurations), 3 (scoring amendment), 6 (contribution correspondence), A1 (terminology) and B1 (hyperparameters) are new; the former Tables 1–3 are now Tables 1, 4 and 5. Equation (3) (score fusion) is new and the CARE score is now Eq. (4).