https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Chronologically calibrated PCA for wind-turbine SCADA health monitoring: a leakage-controlled CARE v6 benchmark

Jui-Hung Liu[1], Jien-Chen Chen[2], and Chun-Chieh Wang[2]

[1] Department of Intelligent Automation Engineering, National Chin-Yi University of Technology, No. 57, Sec. 2, Zhongshan Road, Taiping District, Taichung 411030, Taiwan
[2] Mechanical and Mechatronics Systems Research Laboratories, Industrial Technology Research Institute, No. 195, Sec. 4, Chung Hsing Road, Chutung, Hsinchu 310401, Taiwan

Correspondence: Jui-Hung Liu (dofliu@gm.ncut.edu.tw)

Abstract

Reliable wind-turbine prognostics and health management requires anomaly detectors that limit false-alarm burden without concealing missed events. This study evaluates chronological calibration of principal component analysis (PCA) reconstruction error on CARE to Compare v6, which contains 5 242 948 real ten-minute supervisory control and data acquisition (SCADA) records from 36 turbines in three wind farms. Turbine identities were partitioned before evaluation into 23 development turbines (58 cases) and 13 locked-test turbines (37 cases). Within each case, PCA was fitted on an earlier normal block, and a disjoint chronological block calibrated the alert threshold. Development-only selection chose global conformal PCA with $\alpha = 0.005$. On the locked test set it achieved a CARE-style score of 0.506 (farm–turbine cluster-bootstrap 95 % confidence interval 0.378–0.632), compared with 0.252 for a fixed-contamination isolation forest comparator; the paired difference was 0.249 (95 % confidence interval 0.026–0.466). Mean normal-case accuracy increased from 0.252 to 0.865, and false-positive events decreased from 17 to 4. However, only 5 of 18 anomaly events were detected, including none of the five events in Farm A. A development-tuned isolation forest scored 0.542 and a conformalized isolation forest scored 0.502; neither paired comparison distinguished the selected PCA method. Ordinary empirical calibration generated exactly the same locked-test alerts as the finite-sample conformal rank rule. The evidence supports chronological, identity-disjoint calibration as a transparent reference for SCADA health monitoring, but it does not support universal superiority of PCA or of its conformal rank correction. We report the negative and equivalence findings in full so that the benchmark can be reused and contested.

1 Introduction

Wind-turbine prognostics and health management (PHM) increasingly depends on supervisory control and data acquisition (SCADA) streams because they are available fleet-wide and describe environmental, electrical, thermal, and rotational behaviour at regular intervals. SCADA-based health monitoring is therefore attractive for early fault detection, particularly when dedicated high-frequency condition-monitoring systems are unavailable or cover only selected drivetrain components (Bangalore and Tjernberg, 2015; Cambron et al., 2018; Gück et al., 2024; Schlechtingen et al., 2013b; Tautz-Weinert and Watson, 2017). The operational objective is not merely to flag unusual samples: a useful detector must identify health-relevant deviations while limiting persistent false alarms that consume analyst attention and erode trust.

https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Comparisons in this field have historically been difficult. Many studies rely on private data, selected case studies, or random sample splits that allow the same turbine and highly autocorrelated neighbouring observations to appear on both sides of validation (Bangalore and Tjernberg, 2015; Cambron et al., 2018; Schlechtingen et al., 2013b; Sun et al., 2016; Zaher et al., 2009). CARE to Compare was introduced to address this limitation with anonymized real-world SCADA data from multiple farms, detailed event windows, and an evaluation score that explicitly combines coverage, accuracy, reliability, and earliness (Gück et al., 2024). The version used here, CARE v6, contains 95 cases from 36 turbines across three wind farms. Unlike a remaining-useful-life dataset, it provides event windows and fault context for early fault and anomaly detection; it does not provide ground-truth time-to-fault trajectories suitable for supervised remaining-useful-life regression.

Normal-behaviour models are well suited to this setting. They learn multivariate behaviour from periods designated as normal and treat deviations during subsequent operation as candidate anomalies (Bangalore and Tjernberg, 2015; Cambron et al., 2018; Schlechtingen and Santos, 2011; Schlechtingen et al., 2013b). Principal component analysis (PCA) provides a transparent multivariate baseline: after robust preprocessing, its reconstruction error summarizes departure from a lower-dimensional normal subspace. The unresolved step is threshold selection. Fixed contamination rates can raise excessive alarms under heterogeneous operating conditions, whereas empirical calibration can directly control how extreme a new reconstruction error must be relative to an independent normal block.

Conformal methods provide a finite-sample rank correction for calibrated nonconformity scores under exchangeability (Angelopoulos and Bates, 2023; Gibbs and Candès, 2021; Lei et al., 2018; Vovk et al., 2005). In time series, exchangeability is not automatically valid, and this study does not claim timestamp-level coverage. Instead, split-conformal quantiles are used as a transparent thresholding mechanism after a chronological fit and calibration separation, and operational performance is measured on future prediction blocks. Candidate methods and their hyperparameters are selected only on development turbines; the final test turbines remain locked until a single method is selected.

The contributions are fourfold. First, we provide a traceable CARE v6 evaluation with archive hashing, chronological within-case separation, and turbine-identity-disjoint development and test sets. Second, we compare chronologically calibrated PCA with fixed, development-tuned, and conformalized isolation forest baselines and with in-sample and independently calibrated PCA thresholds, using CARE criticality aggregation rather than pointwise accuracy alone. Third, we quantify uncertainty with paired farm–turbine, farm-stratified, and leave-one-turbine-out analyses, thereby avoiding pseudo-replication across multiple cases from the same turbine. Fourth, we report negative and equivalence evidence: physics and multi-view residual candidates did not improve the selected detector, and finite-sample conformal ranking did not change locked-test alerts relative to ordinary empirical calibration. The resulting contribution is an auditable benchmark for chronological calibration and for the false-alarm–sensitivity trade-off, rather than a new state-of-the-art detector.

Section 2 reviews related work, Sect. 3 documents the data and the frozen evidence-governance rules, Sect. 4 describes the methods, Sect. 5 reports results, Sect. 6 discusses what the evidence does and does not support, and Sect. 7 concludes.

https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

2 Background and related work

2.1 SCADA normal-behaviour modelling

SCADA-based anomaly detection commonly models expected power, temperature, or other condition variables as functions of operating and environmental variables (Bangalore and Tjernberg, 2015; Cambron et al., 2018; Schlechtingen et al., 2013a, b; Sun et al., 2016). Neural networks and regression models can capture nonlinear dependence, whereas reconstruction models learn a multivariate representation of normal operation. Reviews consistently identify data quality, operating-regime variability, thresholding, and the scarcity of public fault labels as limiting factors (Bangalore and Tjernberg, 2015; Cambron et al., 2018; Schlechtingen et al., 2013b; Tautz-Weinert and Watson, 2017). Generalized and cross-turbine models offer scale, but turbine-specific normal-behaviour models often remain necessary because sensor definitions, ranges, and control policies differ across farms.

PCA remains a useful benchmark because it is deterministic, interpretable, and computationally modest. Its reconstruction error can reveal multivariate departures that are not obvious in any single channel. Isolation forest is another common unsupervised baseline that recursively isolates unusual samples (Liu et al., 2008). The CARE paper reported that a fixed-contamination isolation forest detected many anomalies but recognized normal behaviour poorly, demonstrating why a metric that penalizes false alarms is necessary (Gück et al., 2024).

2.2 Threshold calibration and early-event scoring

A threshold converts a continuous anomaly score into an operational alert. Quantile thresholds estimated on the fitting data are optimistic because the same samples determine both the representation and the cutoff. Split calibration separates those roles. For n calibration scores, the finite-sample split-conformal threshold uses the ceiling rank of $(n + 1)(1 - \alpha)$, clipped to the available sample count (Angelopoulos and Bates, 2023; Gibbs and Candès, 2021; Lei et al., 2018; Vovk et al., 2005). This rank correction is distribution-free for exchangeable calibration and test scores, but serial dependence and distribution shift in SCADA require empirical validation rather than an unconditional coverage claim.

Laguna et al. (2026) recently combined turbine clustering, regression models, and temporal conformal intervals for converter monitoring on more than 40 turbines. Their fleet, component targets, and alert definitions differ from CARE v6, so numerical performance cannot be compared directly; the present study instead contributes evaluation governance and calibration evidence under the public CARE event protocol rather than a claim of outperforming that framework.

Pointwise metrics do not fully capture the operational burden of alarms. CARE therefore combines pointwise F0.5 on anomaly cases, accuracy on normal cases, eventwise F0.5 after a criticality accumulator, and a weighted earliness score (Gück et al., 2024; Lavin and Ahmad, 2015). A criticality threshold of 72 corresponds to 12 h of consecutive ten-minute alerts, with isolated gaps gradually decreasing criticality. The score assigns zero if no anomaly event is detected and falls back to normal accuracy if normal accuracy is below 0.5. Otherwise, the four components are combined with double weight on normal accuracy (Gück et al., 2024).

2.3 Evidence gap addressed here

https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Recent work continues to propose increasingly complex autoencoders, temporal networks, and physics-guided features for wind-turbine monitoring (Feng et al., 2023; Laguna et al., 2026; Liu et al., 2020, 2023; Sun et al., 2016). Complexity alone does not establish operational benefit: a detector may improve pointwise recall while producing an unacceptable event-level false-alarm burden, or it may appear effective because development and evaluation observations share turbine identity. This study addresses that gap by freezing method selection before inspecting locked-test outcomes and by treating the turbine as the inferential cluster.

3 Data and evidence governance

3.1 CARE v6 archive

The analysis used CARE to Compare v6 from Zenodo (https://doi.org/10.5281/zenodo.15846963). The local archive SHA-256 checksum was ca61379e98956d891041ad45c885109bd8a14199fde0688d0184a11c2d4194f1. The unpacked corpus contained 5 242 948 rows, 95 cases, 36 turbines, and three wind farms. In v6, 45 cases were labelled anomalous and 50 represented normal behaviour. Each file contains a training and prediction designation, a timestamp, turbine status, anonymized sensor channels, and case-level event metadata. Only ten-minute average channels with sufficient finite coverage and non-zero variance were eligible for PCA.

CARE was designed for early fault detection rather than remaining-useful-life regression (Gück et al., 2024). Accordingly, this study does not construct synthetic remaining-life labels, censored remaining-useful-life targets, or prediction intervals for time to fault. Event start and end annotations are used only for evaluation and for development-set model selection. The per-case fit and calibration steps use normal-status training observations and do not use the case event label.

3.2 Turbine-disjoint development and test split

The split rule was frozen before the full run. For each farm, turbines were sorted by SHA-256(farm | asset identity), and the lowest ceiling of one third were assigned to the final test set. This deterministic identity-only rule does not depend on event labels or model scores. The held-out test identities were A: {13, 21}; B: {12, 7, 2}; and C: {38, 34, 5, 53, 50, 56, 55, 35}. All cases belonging to a turbine inherited its split. Development comprised 23 turbines and 58 cases (27 anomalous, 31 normal); the test set comprised 13 turbines and 37 cases (18 anomalous, 19 normal). Figure 1 summarizes the resulting workflow; the case-level composition of the frozen split is shown in Fig. S1 of the Supplement.

https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

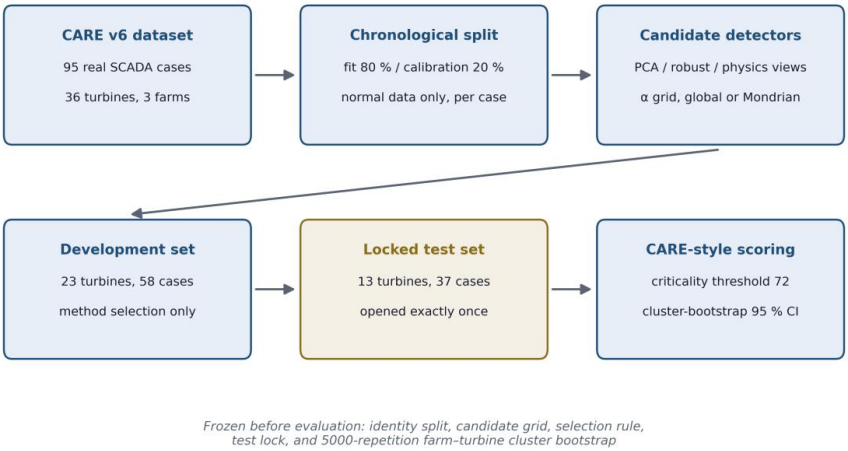

Figure 1. Leakage-controlled evaluation workflow (Sect. 3.2 and Sect. 4).

The protocol distinguishes hyperparameter generalization from case personalization. A candidate method may fit a normal-behaviour model to the normal training segment supplied for a locked-test case, because such historical normal data would be available when commissioning a turbine-specific detector. However, no event label, prediction-segment outcome, or test-turbine aggregate metric is used to choose the method, the view weights, α, or the calibration mode.

4 Methods

4.1 Chronological preprocessing

Rows designated as training and having CARE normal status identifiers {0, 2} formed the normal reference sequence. The first 80 % of that sequence defined the fit and calibration boundary. To preserve temporal order and constrain computation, the most recent contiguous block of at most 12 000 rows before the boundary was used for fitting, and the following disjoint contiguous block of at most 6000 rows was used for calibration. Prediction rows were never used to fit imputers, scalers, or PCA loadings.

Features were restricted to ten-minute average channels present in each farm. Channels with less than 50 % finite coverage in the fit block, or with negligible variance, were excluded. Remaining missing values were median-imputed from the fit block and standardized using fit-block means and standard deviations. Because anonymized sensor identifiers are farm-specific, semantic descriptions were parsed separately by farm for the exploratory thermal and power-curve views; identifiers were not assumed to be aligned across farms.

4.2 PCA nonconformity score

PCA was fitted to the standardized fit matrix (Hotelling, 1933). The number of components was the smallest number retaining at least 99 % of fit variance. For a standardized observation x, the nonconformity score was the mean squared reconstruction error:

https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

$$s_{\mathrm{PCA}}(x) = p^{-1}\left\|x - P^{\mathrm{T}}Px\right\|_2^2, \qquad (1)$$

where p is the number of retained input channels and P contains the retained loading vectors. The same imputer, scaler, and PCA transformation were applied to the independent calibration and prediction blocks.

4.3 Split-conformal threshold

Let $s_1, \ldots, s_n$ be the calibration reconstruction errors sorted in ascending order. For a nominal tail probability α, the global threshold was

$$q_\alpha = s_{(k)}, \quad k = \min\{n, \max[1, \lceil (n+1)(1-\alpha) \rceil]\}, \qquad (2)$$

A prediction timestamp was flagged when $s_{PCA}(x) > q_a$. Candidate α values were {0.005, 0.010, 0.020, 0.030, 0.050}. A secondary Mondrian candidate divided calibration observations into a 3 × 3 grid using farm- and case-specific tertiles of wind speed and active power; cells with fewer than 100 calibration observations fell back to the global threshold. These regime thresholds were evaluated as candidates but were not selected.

The term conformal is used here for the split-conformal rank correction, not for a claimed guarantee under arbitrary time dependence. Chronological separation reduces direct reuse, but SCADA autocorrelation and non-stationarity violate simple exchangeability. Operational test metrics therefore remain the evidentiary endpoint.

4.4 Pre-specified alternative views

To test whether domain-informed residuals improved over generic multivariate reconstruction, two additional views were pre-specified. The robust view was the mean of the three largest absolute standardized feature deviations. The physics view combined two data-driven residual families: (a) one-step ridge regressions for up to ten temperature channels using ambient temperature, active power, wind speed, rotational speed, and the lagged target temperature; and (b) a histogram gradient-boosting normal power-curve model using wind speed, rotational speed, and ambient temperature. Calibration scores from each available view were mapped to empirical quantiles before weighted fusion.

Ten weight configurations were considered: PCA only, robust only, physics only, three pairwise 50/50 combinations, equal thirds, and three biased mixtures, with missing views renormalized among the available weights. These candidates were empirical residual models motivated by drivetrain thermal inertia and the wind–power relation, not first-principles models; their inclusion made the original physics-informed hypothesis falsifiable.

4.5 Development-only selection and comparators

The complete candidate grid comprised ten view-weight configurations, five α values, and global or Mondrian calibration. Candidates were ranked on the development CARE-style score; ties were broken by mean normal accuracy and then by event F0.5. This procedure selected PCA only, α = 0.005, and global calibration. The locked test set was then summarized once for that selected configuration. Table 1 lists every element of the protocol that was frozen before the run.

Table 1. Frozen experimental configuration.

https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

Protocol element	Frozen value
Random seed	42
Fit and calibration chronology	80/20 within the normal training sequence
Maximum fit / calibration rows	12 000 / 6000
PCA variance retained	≥ 99 %
Candidate α	0.005, 0.010, 0.020, 0.030, 0.050
Calibration modes	Global; 3 × 3 wind–power Mondrian
CARE criticality threshold	72
Bootstrap	5000 farm–turbine cluster resamples

The pre-specified primary comparator reproduced the CARE paper's simple isolation forest logic: 100 trees, contamination 0.09, and random seed 42, fitted after a 99 %-variance PCA transform of the same normal fit block (Gück et al., 2024; Liu et al., 2008). To test comparator fairness, four additional baseline families were selected independently on development turbines: isolation forest contamination in {0.005, 0.010, 0.020, 0.030, 0.050, 0.090}; isolation forest scores calibrated on the disjoint block; PCA thresholds estimated in-sample from the fitting block; and ordinary empirical PCA calibration without the split-conformal ceiling-rank correction. Each family used the same α grid where applicable. These additions separate representation choice, independent calibration, and the finite-sample rank correction. Comparisons with α = 0.010 PCA, robust, physics, equal fusion, and Mondrian PCA remained secondary ablations. No test result was used to choose a method or a baseline setting.

4.6 CARE-style evaluation and statistical analysis

Evaluation followed the CARE definitions (Gück et al., 2024). Only timestamps with normal operational status were eligible. Coverage was the mean pointwise F0.5 across anomaly cases; accuracy was the mean correct-normal proportion across normal cases; reliability was event F0.5 after criticality aggregation; and earliness used the CARE piecewise event-position weights. Criticality increased by one for an eligible alert and decreased by one, to a minimum of zero, for an eligible non-alert. An event was predicted anomalous if the maximum criticality reached 72. The aggregate score was

$$\mathrm{CARE} = \frac{\overline{\mathrm{F}_{0.5}} + \overline{\mathrm{WS}} + \mathrm{F}_{0.5,\mathrm{event}} + 2\overline{\mathrm{Acc}}}{5}, \qquad (3)$$

with the CARE special cases retained: the score was zero if no anomaly event was detected, and it equalled normal accuracy when normal accuracy was below 0.5. Because the implementation was independently reconstructed and was not executed through an official v6 leaderboard service, results are termed CARE-style scores throughout.

Uncertainty was quantified with 5000 paired cluster-bootstrap resamples (seed 42). The resampling unit was the unique farm–turbine identity; all cases of a sampled turbine were carried together, and aggregate CARE-style scores were recalculated for each replicate. The pre-specified primary comparison was judged positive only when the 95 % percentile interval for the selected-minus-fixed-isolation-forest score excluded zero. Secondary analyses used farm-stratified cluster resampling and leave-one-turbine-out influence estimates. Comparisons added to assess baseline sensitivity are descriptive and were not adjusted for multiplicity.

https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

5 Results

5.1 Run integrity and selected configuration

All 95 requested cases completed successfully with no failed files. The full run used the verified CARE v6 archive and finished in 15.8 min on Windows 11 with an Intel Core i9-13980HX central processing unit and 31.6 GB of memory; graphics-processing-unit acceleration was not used. The analysis was implemented with NumPy, pandas, and scikit-learn (McKinney, 2010; Pedregosa et al., 2011; van der Walt et al., 2011). Ten unit tests covering scoring, empirical quantiles, missing-view renormalization, Mondrian fallback, and label-independent splitting passed before the full run. Development-only ranking selected conformal PCA with $\alpha = 0.005$. Its development CARE-style score was 0.544, with 13 of 27 anomaly events detected, 11 false-positive events among 31 normal cases, event F0.5 of 0.528, and mean normal accuracy of 0.775.

5.2 Locked-test performance

On the 37 locked-test cases, conformal PCA achieved a CARE-style score of 0.506; its farm–turbine cluster-bootstrap 95 % interval was 0.378–0.632. Isolation forest with PCA scored 0.252. The paired difference was 0.249 (95 % confidence interval 0.026–0.466), excluding zero for this single pre-specified comparator. In 98.68 % of paired resamples, the selected score exceeded the comparator. Figure 2 shows the locked-test scores with cluster-bootstrap intervals, and Table 2 lists all comparators.

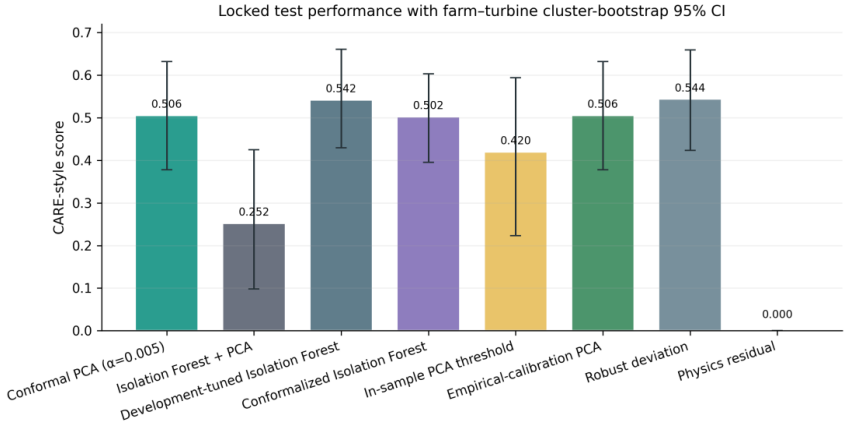

Figure 2. Locked-test CARE-style scores with 95 % farm–turbine cluster-bootstrap intervals (Sect. 5.2).

The improvement was operational rather than sensitivity-dominant. Conformal PCA detected 5 of 18 anomaly events, compared with 13 for isolation forest. However, false-positive events fell from 17 to 4, and mean normal-case accuracy rose from 0.252 to 0.865. Event F0.5 was similar (0.463 versus 0.471). The higher CARE-style score therefore reflects far better recognition of normal behaviour, which the CARE metric weights explicitly. Figure 3 makes this trade-off explicit.

https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

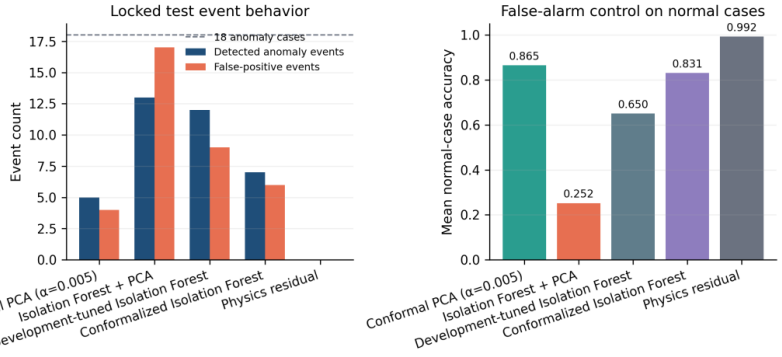

Figure 3. Locked-test trade-off between detected anomaly events and false-alarm burden (Sect. 5.2).

Table 2. Locked-test comparison. Secondary comparisons were not adjusted for multiplicity and were not used to select the reported method.

Method	CARE-style	95 % cluster CI	Detected	False-positive events	Normal accuracy
Conformal PCA (α = 0.005)	0.506	0.378–0.632	5/18	4	0.865
Isolation forest + PCA	0.252	0.098–0.425	13/18	17	0.252
Development-tuned isolation forest	0.542	0.430–0.661	12/18	9	0.650
Conformalized isolation forest	0.502	0.395–0.603	7/18	6	0.831
In-sample PCA threshold	0.420	0.223–0.594	13/18	12	0.420
Empirical-calibration PCA	0.506	0.378–0.632	5/18	4	0.865
Conformal PCA (α = 0.010)	0.511	0.387–0.633	6/18	5	0.838
Robust deviation	0.544	0.423–0.659	7/18	4	0.896
Physics residual	0.000	0.000–0.000	0/18	0	0.992
Equal fusion	0.447	0.000–0.519	1/18	2	0.963
Regime conformal PCA	0.517	0.400–0.622	8/18	8	0.814

Additional fairness baselines were selected on development turbines without using locked-test outcomes. Their locked-test scores were 0.542 for tuned isolation forest, 0.502 for conformalized isolation forest, 0.420 for an in-sample PCA threshold, and 0.506 for ordinary empirical PCA calibration. The selected-minus-tuned-isolation-forest paired interval was −0.188 to 0.080, and the selected-minus-conformalized-isolation-forest interval was −0.092 to 0.083; neither distinguished the methods. Ordinary empirical PCA calibration generated exactly the same locked-test score and alerts as the finite-sample conformal rule (paired difference and interval both zero). These results attribute the practical gain mainly to independent chronological calibration rather than to a unique finite-sample correction.

https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

The selected-minus-fixed-isolation-forest difference remained 0.247 under farm-stratified cluster resampling (95 % confidence interval 0.070–0.412). Across 13 leave-one-turbine-out analyses, the difference ranged from 0.171 to 0.314.

5.3 Cross-farm heterogeneity

Performance differed markedly by wind farm. In Farm A, conformal PCA raised no event alarm: 0 of 5 anomaly cases were detected and 0 of 3 normal cases became false-positive events, yielding a CARE-style score of zero under the CARE special rule. In Farm B, it detected 1 of 2 anomalies without a false-positive event and scored 0.608. In Farm C, it detected 4 of 11 anomalies with 4 false-positive events and scored 0.517. Normal accuracy was 0.999, 0.987, and 0.806 for Farms A, B, and C, respectively. These differences show that a respectable overall score can coexist with complete failure in one domain. Figure 4 and Table 3 give the farm-resolved breakdown.

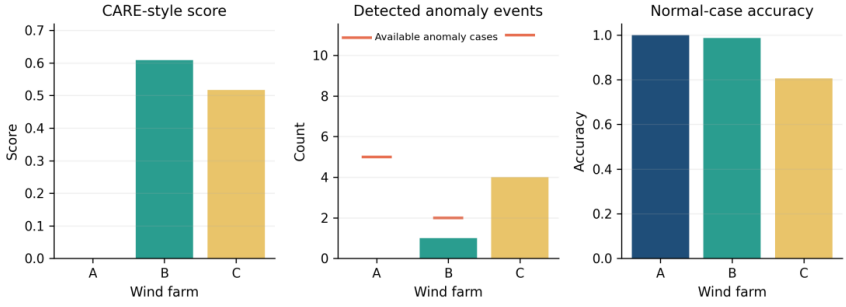

Figure 4. Farm-wise locked-test results for the selected configuration (Sect. 5.3).

Table 3. Farm-wise locked-test results for the selected conformal PCA configuration. Lead time is conditional on an eligible alert inside the annotated event window and is not a remaining-useful-life estimate.

Farm	Cases	Detected anomalies	False-positive events	Normal accuracy	CARE-style	Median lead time (d)
A	8	0/5	0	0.999	0.000	–
B	5	1/2	0	0.987	0.608	58.46
C	24	4/11	4	0.806	0.517	14.99

5.4 Sensitivity and ablation evidence

Development α sensitivity was shallow between 0.005 and 0.030, with CARE-style scores from 0.534 to 0.544, but the operational composition changed. As α increased, more anomaly events were detected and more normal events were falsely alarmed; at α = 0.050 the development score fell to 0.493. The frozen selection rule chose α = 0.005 because it had the highest development score (0.544), not because it performed best on the test set. The development-only sensitivity curve is shown in Fig. S2 of the Supplement.

On the locked test set, α = 0.010 scored 0.511 and robust deviation scored 0.544, both numerically above the selected 0.506. Their paired differences from the selected method had 95 % intervals that included zero: −0.029 to 0.008 and −0.137 to 0.041, respectively. Regime conformal PCA scored 0.517, with a selected-minus-regime interval of −0.074 to 0.038. These secondary results do not

https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

support superiority of the selected α or of global calibration on the test set; they support only the integrity of the development-only choice.

The physics-only candidate supplied the strongest negative finding. It detected 2 development anomaly events and scored 0.482, but it detected zero locked-test anomaly events and therefore received a CARE-style score of zero. The selected-minus-physics paired difference was 0.501 (95 % confidence interval 0.378–0.632). Equal fusion detected one locked-test anomaly event and scored 0.447. Consequently, neither the physics view nor its inclusion in an equal fusion improved the selected detector on real data. No physics-informed effectiveness claim is retained.

5.5 Conditional early-warning timing

Among anomaly cases with at least one eligible alert inside the annotated event window, the selected method's median first-alert lead time was 18.17 d. This statistic is conditional on alerts and does not imply that all detected cases reached the CARE event criticality threshold. It is based on only five locked-test events with eligible alerts and must not be interpreted as a remaining-useful-life prediction. Farm B's single alerted anomaly had 58.46 d of lead time; Farm C's alerted events had a median of 14.99 d; Farm A had no eligible alert and therefore no lead-time estimate.

6 Discussion

6.1 What the supported improvement means

The primary result is narrower than a claim of generally superior fault detection. The paired bootstrap supports a higher overall CARE-style score than the fixed-contamination isolation forest comparator, and the score components show why: conservative chronological calibration dramatically reduced normal-period alerts, which matters because analysts must investigate alarms and unreliable detectors are often disabled in practice. The result also reproduces the CARE paper's observation that a fixed isolation forest can detect many abnormal samples yet perform poorly on normal data (Gück et al., 2024).

The fair-baseline analysis further narrows the interpretation. Development-tuned isolation forest scored 0.542, numerically above the selected PCA score of 0.506, while conformalized isolation forest scored 0.502; both paired intervals included zero. Ordinary empirical calibration and finite-sample conformal ranking produced identical locked-test alerts. Consequently, the evidence does not isolate a unique advantage of the PCA representation or of the ceiling-rank correction. It supports the broader design principle of separating representation fitting, chronological calibration, and locked evaluation.

PCA nevertheless provides a stable, inspectable multivariate normal subspace, and the disjoint calibration block converts reconstruction error into a case-specific tail threshold without fault labels. Its computational modesty and transparent score decomposition make it a useful low-false-alarm reference implementation, provided that sensitivity and cross-farm limitations are reported alongside the aggregate score.

6.2 What is not supported

Four broader claims are not supported by this evidence. First, the selected detector is not superior to all fairly tuned alternatives. Second, it is not a high-sensitivity universal early fault detector: it

https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

detected only 5 of 18 locked-test anomaly events and none in Farm A. Third, the physics-residual view did not generalize in this dataset and therefore does not support a physics-informed effectiveness claim. Fourth, conditional first-alert timing is not remaining useful life. CARE supplies annotated event windows, not continuously verified degradation-to-failure targets, and a first alert cannot be converted into a calibrated time-to-fault trajectory without additional ground truth.

The physics ablation is scientifically informative. One-step thermal residuals may be too conservative, may absorb degradation into normal variability, or may depend on farm-specific sensor semantics, and the power-curve residual may be insensitive to component-specific failures. More elaborate physics models require new development data and a newly untouched evaluation cohort before an effectiveness claim is warranted. We report this outcome in full because pre-specified negative results are part of what makes a benchmark trustworthy.

6.3 Cross-farm deployment implications

Farm heterogeneity is the principal deployment barrier. A threshold that almost perfectly suppresses normal alerts in Farm A also suppresses all anomaly events there. Farm C shows a more balanced pattern but lower normal accuracy. This suggests that global per-case quantiles alone do not resolve differences in sensor sets, control logic, fault composition, or event labels. The tested 3 × 3 wind–power Mondrian calibration did not establish a reliable improvement, although its numerical test sensitivity was higher. Future work should explore hierarchical calibration with farm- and component-level partial pooling, evaluated under a new pre-registered split.

Operational deployment also requires prospective validation: the present evidence is retrospective and software-only, and it does not demonstrate online latency, alarm routing, analyst response, cost reduction, or integration with IEC 61400-25 information models.

6.4 Limitations

The held-out set contains 13 turbines and 37 cases; cluster intervals are therefore wide. Several turbines contribute multiple cases, and the cluster bootstrap addresses within-turbine dependence only approximately. The deterministic identity split is transparent but represents one realization rather than repeated cross-farm external validation. CARE event labels have uncertain onset boundaries, as acknowledged by the dataset authors (Gück et al., 2024).

Each test case uses its own earlier normal segment to fit a personalized detector. The study therefore evaluates commissioning on a turbine that already has historical normal data, not zero-shot transfer to a turbine without history. Calibration and test scores remain serially dependent, so no exact timestamp-level conformal false-alarm guarantee is claimed. The CARE-style implementation was reconstructed and validated by unit tests but was not submitted to an official v6 scoring service. Finally, the comparator set is intentionally compact; no claim is made against all recent deep-learning models.

7 Conclusions

This study evaluated chronological calibration for wind-turbine SCADA health monitoring under a frozen turbine-disjoint protocol. Development-only selection chose global conformal PCA. On 13

https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

held-out turbines it scored 0.506, versus 0.252 for the fixed-contamination isolation forest comparator, with a paired difference of 0.249 (95 % confidence interval 0.026–0.466); the difference arose mainly from a lower false-alarm burden and higher normal-case accuracy.

Fairly tuned alternatives narrowed the method-level conclusion: tuned isolation forest scored 0.542, conformalized isolation forest scored 0.502, and neither paired comparison distinguished the selected PCA method. Ordinary empirical PCA calibration exactly matched the finite-sample conformal rule on the locked test set. Together with 5 of 18 anomaly-event detections, zero detections in Farm A, and a zero-scoring physics candidate, the evidence supports a transparent leakage-controlled benchmark for chronological calibration rather than a universally superior detector.

We therefore offer this work as a reusable evaluation protocol. The frozen split rule, the candidate grid, the selection criterion, and the complete set of positive, null, and negative outcomes are published together so that other groups can reproduce the comparison, challenge the interpretation, or reuse the identity-disjoint partition for their own detectors. Future studies require a newly frozen external cohort and should continue to report both false alarms and missed events.

Appendix A: Reproducibility artefacts

The reproducibility package contains the frozen protocol with its implementation SHA-256 checksum, the analysis source code, ten unit tests, machine-readable locked-test summary tables, the run manifest recording completion of all 95 cases, and environment metadata. The frozen protocol was hashed before any full-run output file existed and explicitly prohibits remaining-useful-life, time-to-failure, and timestamp-level coverage claims; the locked-test turbine identities were recorded before model execution. The Supplement contains the case-level composition of the frozen split (Fig. S1) and the development-only α sensitivity curve (Fig. S2).

Code availability

The exact analysis version used in this study is supplied as a checksum-verified supplement containing source code, unit tests, the frozen protocol, machine-readable result tables, and environment metadata. Raw CARE files are not redistributed. The archive will be deposited in a FAIR-aligned repository with a persistent identifier upon acceptance, and the identifier will be inserted here.

Data availability

CARE to Compare v6 is publicly available from Zenodo at https://doi.org/10.5281/zenodo.15846963 (Gück et al., 2025). The analysed local archive had SHA-256 checksum ca61379e98956d891041ad45c885109bd8a14199fde0688d0184a11c2d4194f1. Derived case-level metrics, the frozen protocol, and the analysis code are included in the accompanying reproducibility package; raw CARE files are not redistributed.

Author contributions

JHL conceptualized the study, designed the methodology, developed the software, performed the formal analysis and validation, curated the data, produced the visualizations, and prepared the original draft. JCC contributed to data curation, provided resources, and participated in validation. CCW

https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

provided resources and participated in validation. All authors reviewed and edited the manuscript and approved the submitted version.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

The authors thank the CARE to Compare dataset creators and the participating wind-farm operators for making anonymized real-world SCADA benchmark data available. During preparation of this work, the corresponding author used a large-language-model assistant for language refinement, consistency checking, code assistance, and document formatting. All scientific content, analysis decisions, and interpretations are the authors' own, and the authors take full responsibility for the content of the publication.

References

Angelopoulos, A. N. and Bates, S.: Conformal prediction: a gentle introduction, Found. Trends Mach. Learn., 16, 494–591, https://doi.org/10.1561/2200000101, 2023.

Bangalore, P. and Tjernberg, L. B.: An artificial neural network approach for early fault detection of gearbox bearings, IEEE T. Smart Grid, 6, 980–987, https://doi.org/10.1109/TSG.2014.2386305, 2015.

Cambron, P., Masson, C., Tahan, A., and Pelletier, F.: Control chart monitoring of wind turbine generators using the statistical inertia of a wind farm average, Renew. Energ., 116, 88–98, https://doi.org/10.1016/j.renene.2016.09.029, 2018.

Feng, C., Liu, C., Jiang, D., Kong, D., and Zhang, W.: Multivariate anomaly detection and early warning framework for wind turbine condition monitoring using SCADA data, J. Energy Eng., 149, 04023004, https://doi.org/10.1061/JLEED9.EYENG-4843

Gibbs, I. and Candès, E.: Adaptive conformal inference under distribution shift, in: Advances in Neural Information Processing Systems, vol. 34, 1660–1672, https://doi.org/10.48550/arXiv.2106.00170, 2021.

Gück, C., Roelofs, C. M. A., and Faulstich, S.: CARE to Compare: a real-world benchmark dataset for early fault detection in wind turbine data, Data, 9, 138, https://doi.org/10.3390/data9120138, 2024.

Gück, C., Roelofs, C. M. A., and Faulstich, S.: CARE to Compare v6, Zenodo [data set], https://doi.org/10.5281/zenodo.15846963, 2025.

Hotelling, H.: Analysis of a complex of statistical variables into principal components, J. Educ. Psychol., 24, 417–441, https://doi.org/10.1037/h0071325, 1933.

Laguna, J., Janal, Y., Reder, M., and Melero, J. J.: Temporal conformal prediction for wind turbine anomaly detection: a robust distribution-free framework with conditional clustering, J. Phys. Conf. Ser., 3224, 062054, https://doi.org/10.1088/1742-6596/3224/6/062054, 2026.

Lavin, A. and Ahmad, S.: Evaluating real-time anomaly detection algorithms – the Numenta anomaly benchmark, in: 2015 IEEE 14th International Conference on Machine Learning and Applications, 38–44, https://doi.org/10.1109/ICMLA.2015.141, 2015.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L.: Distribution-free predictive inference for regression, J. Am. Stat. Assoc., 113, 1094–1111, https://doi.org/10.1080/01621459.2017.1307116, 2018.

Liu, F. T., Ting, K. M., and Zhou, Z.-H.: Isolation forest, in: 2008 Eighth IEEE International Conference on Data Mining, 413–422, https://doi.org/10.1109/ICDM.2008.17, 2008.

Liu, J., Yang, G., Li, X., Wang, Q., He, Y., and Yang, X.: Wind turbine anomaly detection based on SCADA: a deep autoencoder enhanced by fault instances, ISA T., 139, 586–605, https://doi.org/10.1016/j.isatra.2023.03.045, 2023.

Liu, X., Lu, S., Ren, Y., and Wu, Z.: Wind turbine anomaly detection based on SCADA data mining, Electronics, 9, 751, https://doi.org/10.3390/electronics9050751, 2020.

https://doi.org/10.5194/wes-2026-145
Preprint. Discussion started: 31 August 2026

McKinney, W.: Data structures for statistical computing in Python, in: Proceedings of the 9th Python in Science Conference, 56–61, https://doi.org/10.25080/Majora-92bf1922-00a, 2010.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: machine learning in Python, J. Mach. Learn. Res., 12, 2825–2830, 2011.

Schlechtingen, M. and Santos, I. F.: Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection, Mech. Syst. Signal Pr., 25, 1849–1875, https://doi.org/10.1016/j.ymssp.2010.12.007, 2011.

Schlechtingen, M., Santos, I. F., and Achiche, S.: Using data-mining approaches for wind turbine power curve monitoring: a comparative study, IEEE T. Sustain. Energ., 4, 671–679, https://doi.org/10.1109/TSTE.2013.2241797, 2013a.

Schlechtingen, M., Santos, I. F., and Achiche, S.: Wind turbine condition monitoring based on SCADA data using normal behavior models, part 1: system description, Appl. Soft Comput., 13, 259–270, https://doi.org/10.1016/j.asoc.2012.08.033, 2013b.

Sun, P., Li, J., Wang, C., and Lei, X.: A generalized model for wind turbine anomaly identification based on SCADA data, Appl. Energ., 168, 550–567, https://doi.org/10.1016/j.apenergy.2016.01.133, 2016.

Tautz-Weinert, J. and Watson, S. J.: Using SCADA data for wind turbine condition monitoring – a review, IET Renew. Power Gen., 11, 382–394, https://doi.org/10.1049/iet-rpg.2016.0248, 2017.

van der Walt, S., Colbert, S. C., and Varoquaux, G.: The NumPy array: a structure for efficient numerical computation, Comput. Sci. Eng., 13, 22–30, https://doi.org/10.1109/MCSE.2011.37, 2011.

Vovk, V., Gammerman, A., and Shafer, G.: Algorithmic learning in a random world, Springer, https://doi.org/10.1007/b106715, 2005.

Zaher, A., McArthur, S. D. J., Infield, D. G., and Patel, Y.: Online wind turbine fault detection through automated SCADA data analysis, Wind Energy, 12, 574–593, https://doi.org/10.1002/we.319, 2009.