

Dear Authors,

I have carefully evaluated the manuscript and provide my assessment below.

General Assessment

The manuscript presents a methodology for anomaly and failure detection in wind turbines. The proposed approach combines a Principal Component Analysis (PCA) model with conformal prediction-based confidence intervals and further incorporates clustered bootstrap techniques for uncertainty quantification. The methodology is validated using the open-access CARE to Compare dataset, and performance is evaluated using the CARE score.

The paper contains several elements. First, it relies on an open-access dataset, which promotes reproducibility and facilitates fair comparison with future studies. Second, it employs the CARE score as a performance metric, which is an interesting choice for evaluating anomaly detection methods in this domain. Third, it utilizes conformal prediction for threshold determination. Finally, it incorporates bootstrap-based uncertainty quantification.

However, despite these positive aspects (which are in my opinion borderline sufficient for a publication), the manuscript has several significant shortcomings that currently limit its suitability for publication.

Text Quality

A major concern is the overall quality of the writing. The manuscript is often difficult to follow, and several sections are insufficiently clear, making it challenging to understand both the methodology and the underlying rationale. The paper would benefit substantially from a thorough revision focused on clarity, consistency, and scientific writing quality.

One issue is the frequent use of non-standard or unconventional terminology. Examples include, but are not limited to:

- "Comparator", used to describe a model employed for comparison purposes. For example, p. 7, line 213: *"The pre-specified primary comparator reproduced ..."*. The intended meaning is not immediately clear and should be reformulated.
- "Development-tuned" and "development-only selection". Throughout the manuscript, the term *development* appears to be used where *training* and *validation* datasets would be more appropriate. For example, p. 4, line 146: *"Development comprised 23 turbines and 58 cases ..."*. A clearer formulation would be: *"The training and validation datasets comprised 23 turbines and 58 cases..."*.

In addition, many sentences are difficult to interpret or lack sufficient clarity. Examples include:

- P. 6, line 188: *"The term conformal is used here for the split-conformal rank correction, not for a claimed guarantee under arbitrary time dependence."* The meaning and relevance of this statement are not sufficiently explained.
- P. 6, line 190: *"Operational test metrics therefore remain the evidentiary endpoint."* It is unclear what is meant by *operational test metrics*. From the surrounding discussion, I assume the authors intend to state that violations of the exchangeability assumption limit the applicability of the theoretical guarantees provided by conformal prediction and therefore necessitate empirical validation. If this interpretation is correct, it should be stated explicitly.
- P. 6, line 200: *"Ten weight configurations were considered ..."* As far as I can determine, the authors are comparing ten different pipeline configurations. The term *weight configurations* appears inappropriate in this context and should be clarified.
- P. 6, line 192: The entire paragraph lacks clarity and should be substantially revised.

The figures and tables are generally acceptable, but some issues remain:

- Figure 1: The figure is difficult to interpret because the annotations and explanatory text are insufficient. Furthermore, the figure appears to contain two captions. In particular, the text *"Frozen before evaluation: identity split, candidate grid, selection rule, ..."* is not adequately explained. Its purpose and meaning should be clarified.
- Table 1: The caption appears on a different page than the table itself. Furthermore, several terms used in the table are not self-explanatory. For example, *"Maximum fit/calibration rows"* is unclear. The authors should explain precisely what is meant by *fit rows* and *calibration rows*.

Overall, the writing quality significantly hinders comprehension of the methodology and the results. Improving the clarity of the manuscript should be a high priority.

Novelty

A second major concern relates to the novelty of the work.

The motivation stated on p. 1, line 37 is valid:

"The operational objective is not merely to flag unusual samples; a useful detector must identify health-relevant deviations while limiting persistent false alarms."

However, this objective is shared by most anomaly detection and condition monitoring studies. While important, it does not in itself constitute a novel contribution.

Similarly, on p. 2, line 40, the authors state that comparisons in this field have historically been difficult. I agree with this observation, and the use of an open-access dataset such as CARE is commendable. Nevertheless, this alone does not provide substantial novelty, as reproducible benchmarking was already a key objective of the original *Care to Compare* work by Gück et al.

The discussion on p. 4, line 119 argues that increasingly complex models do not necessarily deliver operational benefits. While this may be true, there are valid reasons why contemporary research employs more sophisticated architectures. Wind turbine behavior is highly complex, and recent approaches attempt to capture this complexity through advanced modeling techniques.

The proposed methodology relies on PCA, a well-established statistical technique that has been widely used in condition monitoring research for many years. A more compelling contribution would have been to compare the proposed framework against several state-of-the-art normal behavior models. Such a comparison could provide valuable insights into whether modern approaches offer tangible benefits over simpler alternatives. The absence of such benchmarking reduces the novelty and broader impact of the work.

The manuscript identifies four contributions (p. 2, line 66 onwards). My assessment of these contributions is as follows:

Contribution 1

The use of CARE v6 is valuable from a reproducibility perspective. However, the temporal data-splitting strategy is a standard approach in time-series analysis and does not represent a novel methodological contribution.

Contribution 2

This contribution introduces a *chronologically calibrated PCA* model and compares it against a conformalized Isolation Forest baseline. However, the term *chronologically calibrated* is not clearly defined.

PCA itself is not novel in this context, as statistical and conventional machine learning techniques have been extensively applied in condition monitoring for many years. The primary innovation appears to be the application of conformal prediction. However, conformal prediction is typically employed to calibrate prediction intervals and quantify uncertainty, and this aspect is not sufficiently developed in the manuscript.

Contribution 3

The uncertainty quantification framework based on clustered bootstrap techniques is potentially the most novel aspect of the work. Uncertainty quantification remains relatively underexplored in condition monitoring research.

However, the treatment of uncertainty is currently insufficient. The manuscript does not clearly distinguish which sources of uncertainty are captured by the bootstrap procedure, nor does it adequately discuss the role of conformal prediction within the overall uncertainty quantification framework.

Contribution 4

The presentation of experimental results is listed as a contribution. This is unusual, as contributions typically describe methodological advances or new scientific insights rather than the existence of results themselves.

The authors further emphasize the importance of reporting negative results. I agree that negative results are valuable and should not constitute grounds for rejection. However, given the existing literature, the relatively limited performance of PCA combined with conformal thresholding may not be particularly surprising. This again raises questions regarding the novelty and scientific contribution of the work.

Related Work

A third concern, is the related work section.

The review of SCADA-based normal behavior modeling literature is not sufficiently up to date, and a number of recent contributions appear to be missing. The literature review should be expanded to better position the proposed work within the current state of the art and to provide a more comprehensive overview of recent developments in the field.

Discussion of Methodology and Justification of Methodological Choices

A fourth concern relates to the clarity and justification of the proposed methodology. While the poor writing quality contributes to this issue, several key methodological aspects are also insufficiently explained. As a result, it is often unclear why certain design choices were made.

For example, the authors do not adequately justify why conformal prediction was selected in combination with PCA for anomaly detection. It remains unclear why this approach was preferred over alternative conformal prediction methods that are specifically designed for time-series data, or indeed over entirely different anomaly detection frameworks.

Similarly, the rationale for using clustered bootstrap for uncertainty quantification is not sufficiently discussed. The manuscript should explain what sources of uncertainty this method is intended to capture and why it was chosen over alternative approaches.

Overall, the paper would benefit from a clearer discussion of the methodological decisions and the reasoning behind them. Such a discussion would help readers better understand and evaluate the proposed framework.

Conclusion

In its current form, the manuscript suffers from three primary weaknesses:

1. Insufficient writing quality and clarity.
2. Limited methodological novelty.
3. An incomplete and outdated review of related work.
4. The methodology is not sufficiently explained.

Taken together, these issues make it difficult to recommend the manuscript for publication in *Wind Energy Science* in its present form.

To improve the manuscript, I recommend the following:

- Conduct additional experiments using more advanced state-of-the-art models and compare their performance against the proposed approach.
- Substantially strengthen the uncertainty quantification section, including a clearer discussion of the uncertainty sources captured by both bootstrap and conformal prediction techniques.
- Thoroughly revise the manuscript to improve readability, clarity, and scientific writing quality.
- Expand and update the related work section to better reflect the current literature.
- Explain the methodology better.

I hope these comments are constructive and helpful in improving the manuscript.

Sincerely,

Reviewer