Author's response

RC4 (Reviewer 4):

We thank the Reviewer for taking the time to provide detailed feedback on our revised manuscript. Following are the responses to the reviewer feedback. The structure is as follows: 1) **Feedback** from the reviewer, 2) **Response** to the feedback, and 3) **Changes to the manuscript:** based on the latexdiff file.

Feedback:
The reported confusion matrix reveals a notable imbalance: 30 false positives and 0 false negatives. While minimizing false negatives is critical in fault detection—particularly in wind turbine drivetrains where undetected faults can escalate into severe mechanical failures—a high number of false positives also poses practical challenges, such as unnecessary maintenance actions and associated costs.

The current model uses a two standard deviation threshold for classifying a measurement as faulty. This leads to the following questions:

- Is this threshold deliberately chosen to prioritize fault sensitivity over specificity?

- Was the trade-off between false positives and false negatives empirically evaluated?

- Would a more conservative threshold (e.g., three standard deviations) significantly reduce false positives without compromising early fault detection?

A brief discussion of the rationale behind the threshold selection, including whether any sensitivity analysis was performed to explore different thresholds, would strengthen the methodological justification and help readers understand the model's design priorities. Even if two sigma remains the chosen setting, explaining the decision will improve transparency and reinforce the paper's practical relevance.

Response:

The proposed method employs the two-sigma rule for fault detection. A measurement is flagged as faulty when the deviation between the actual and predicted indicator exceeds two standard deviations from the mean. Additionally, the method incorporates a two-level alarm approach:

1. Deviations between two and four standard deviations are classified as warnings,
2. Deviations exceeding four standard deviations are classified as alarms.

We acknowledge that this was inaccurately described in the originally submitted manuscript, and it has now been corrected in the revised version.

The following are our detailed responses to each of the questions and suggestions raised by the reviewer:

Is this threshold deliberately chosen to prioritise fault sensitivity over specificity?

The threshold was chosen to prioritise fault sensitivity over specificity. The primary objective of this study is early fault detection, which depends on the ability to identify subtle changes in indicator trends that may signal the initiation of a fault. These early changes are typically confirmed when the fault trends continue to grow over time. The proposed approach facilitates predictive maintenance by enabling human experts to visually assess indicator trends before initiating physical inspections. This intermediate step helps reduce false alarms by allowing experts to evaluate the evolution of fault indicators over time. In critical systems like wind turbines, such early detection is particularly valuable, as it can significantly reduce unplanned downtime and associated maintenance costs.

Was the trade-off between false positives and false negatives empirically evaluated?

An empirical evaluation of the trade-off between false positives and false negatives has not yet been conducted in this study, as the focus was primarily on early fault detection. The proposed method was validated using real wind farm data, which introduces several challenges for such analysis. A major limitation is the uncertainty in the exact timing of fault initiation. Although fault cases were confirmed through manual inspection, the lack of precise information makes it difficult to accurately label data for a robust classification of false positives and false negatives.

Additionally, the complex structure of wind turbine gearboxes means that fault-related frequencies may also appear in sensors monitoring neighbouring components. This further complicates the identification of true vs. false alarms and makes threshold selection particularly sensitive.

Given these constraints, fault cases verified by manual inspection were treated as ground truth, and all other cases were treated as healthy. However, to systematically evaluate the false positive/negative trade-off, a more controlled and reliably annotated dataset would be appropriate.

Would a more conservative threshold (e.g., three standard deviations) significantly reduce false positives without compromising early fault detection?

A higher threshold can help reduce false positives; however, it comes at the cost of decreased sensitivity to subtle changes in indicator trends, which are critical for early fault detection. In our case, using a higher threshold risks missing early signs of degradation. Moreover, due to the complexity of the wind turbine gearbox, fault frequencies originating from one component may also be captured by neighbouring sensors. As a result, even a conservative threshold may not be able to eliminate false alarms without compromising the model's ability for early fault detection.

Changes to the manuscript:

Page: 9 Lines: 228-230 and 235-245 in Normal Behaviour Models subsection.

Page: 19 Lines: 385-390 in Discussion section.