Response to Report #2

This work proposes a PCA-LSTM anomaly detection approach for a lab-scale wind turbine failure detection. While the methodology is sound and demonstrates clear value, several technical details require clarification for improved reproducibility.

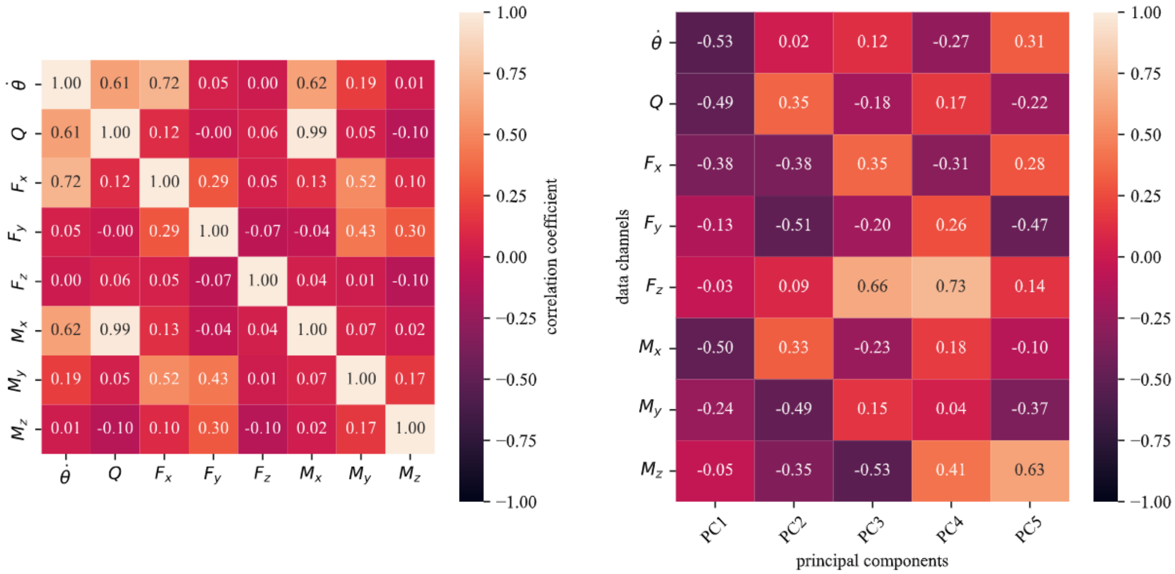
1) The discussion of vibration-based condition monitoring in introduction seems misplaced since the approach used angular velocity, torque, and force measurements rather than vibration signals. Please clarify the relevance or consider focusing the literature review on multi-sensor anomaly detection methods more aligned with your methodology.

We appreciate the reviewer's comment on this. Our intention was not to emphasize vibration-based condition monitoring techniques but rather to highlight that effective fault detection can be achieved using measurements already available in the system, such as angular velocity or torque measurements, without the need for additional sensors. We have revised the paragraph in the introduction to clarify this point and shifted the focus from vibration-based sensors and clarified our emphasis on multi-sensor anomaly detection methods that leverage existing control system data. We have added another citation regarding mutli-sensor fusion for anomaly detection that aligns more directly with the methodology used in this study.

~~Vibration-based condition monitoring~~ In many condition monitoring applications, anomaly detection is performed using dedicated sensors. For instance, vibration-based techniques, often evaluated using the root mean square (RMS) of velocity or acceleration signals, are widely used for drivetrain fault detection ~~, particularly to determine whether signal amplitudes exceed the thresholds defined by~~ and are assessed against standards such as ISO 10816-21 (ISO, 1996). ~~For example,~~ However, deploying additional instrumentation is not always feasible or cost-effective. Nejad et al. (2018) demonstrated that angular velocity measurements already available ~~in~~ within existing control systems can be repurposed for fault detection, ~~avoiding~~ thereby eliminating the need for ~~costly additional instrumentation. Their approach was motivated by the challenge of identifying faults in complex systems with many interconnected components, where vibration signals may originate from various internal sources at different frequencies. In such cases, incorporating multiple sensor channels is often necessary to obtain a more complete understanding of system behavior. However, this added complexity can increase computational cost and the risk of misinterpreting irrelevant or noisy signal components.~~ supplementary sensors. Similarly, Dameshghi and Refan (2019) proposed a diagnostic approach for gearbox faults based on SCADA information multi-sensor fusion, avoiding the need for additional data collection systems. These approaches illustrate the potential of multi-sensor anomaly detection methods that leverage existing system measurements.

2) Please show absolute values in both correlation and covariance matrices in figure 4 for clearer interpretation.

Figure 4 has been updated to show the absolute values



(a) Correlation matrix between channels of interest. (b) Covariance loadings matrix.

3&4) Please maintain consistent terminology throughout the paper. Authors refer to RNN in lines 113-115 but use LSTM elsewhere. Please consider using LSTM consistently to avoid confusion. Why choose MAE over MSE for reconstruction error? MAE can be less sensitive to outliers, but MSE might better capture the magnitude of deviations. Please justify this choice.

Thank you for pointing this important distinction. We chose the mean absolute error (MAE) over the mean squared error (MSE) as our reconstruction error to reduce the model's sensitivity to transient spikes or noise that may not necessarily correspond to actual anomalies. This allows the model to only detect sustained system variations from nominal behavior. With that said, we will consider incorporating a comparative analysis in possible future extensions of this work to further asses the impact of the choice of reconstruction error metric. In the meantime, we have added a justification to the use of MAE in the manuscript as follows:

The PCs were then used to train the ~~RNN model(s)~~ LSTM models that will later be used for prediction. As new data is acquired, it is transformed/projected onto the same PCs that were used in training the models. For the purpose of anomaly detection, the mean absolute error (MAE) is computed between measurements and predictions from the ~~RNN~~ LSTM models, and the error derivative is calculated, to estimate rapid fluctuations in the quality of the predictions. The choice of MAE as the reconstruction error metric was made to reduce model sensitivity to transient spikes or noise, which may not correspond to true anomalies. An anomaly alert is reported to the operator when certain anomalous conditions are met. In this research, we investigate conditions when both the error and its derivative were crossing certain thresholds. This procedure is illustrated in Figure 5 and is explained in section 2.6.

5) Please specify the input sequence length for the LSTM model.

The input sequence length is specified in the manuscript as a look-back to prediction ratio of n/m=10, corresponding to a look-back window (input sequence length) of 10 timesteps for a 1-step prediction horizon. To improve clarity, we have now explicitly stated the input sequence length in the text.

Models $\mathcal{M}_1$ and $\mathcal{M}_3$ were configured with identical training hyperparameters, except for the number of training epochs. Both models utilize a prediction horizon of a single timestep and a look-back to prediction ratio of $n/m = 10$, corresponding to an input sequence length of 10 timesteps. The network architecture consists of a single hidden layer with 100 neurons,

6) Given the standardization approach in equation 1 where data is normalized to mean zero and standard deviation of one, and considering that synthetic anomalies are created by amplifying monitored signals, a fundamental question arises: if both training healthy data and anomalous test data are normalized to the same scale, how can the resulting deviations be detected by the model? Please clarify in this regard.

We appreciate the reviewer's thoughtful question. To clarify, the standardization procedure described in Equation 1 is applied using the mean and standard deviation computed from the healthy training dataset **only**. The resulting scaler is then stored and applied to transform all subsequent input data (could be healthy or anomalous), prior to PCA projection (the PCA transform, not to be confused with scaler transform) and prediction. This ensures that deviation from the distribution of the healthy training data are preserved and detectable. We have updated the manuscript to explicitly state that the standardization model (scaler) is fitted once on the training data and reused for transforming new data.

Data are then standardized to ensure all channels (features) in the training dataset have a mean value, μ, of 0 and a standard deviation, σ, of 1:

$$x_i = \frac{x_i - \mu_i}{\sigma_i}, i = 1, ..., N, \tag{1}$$

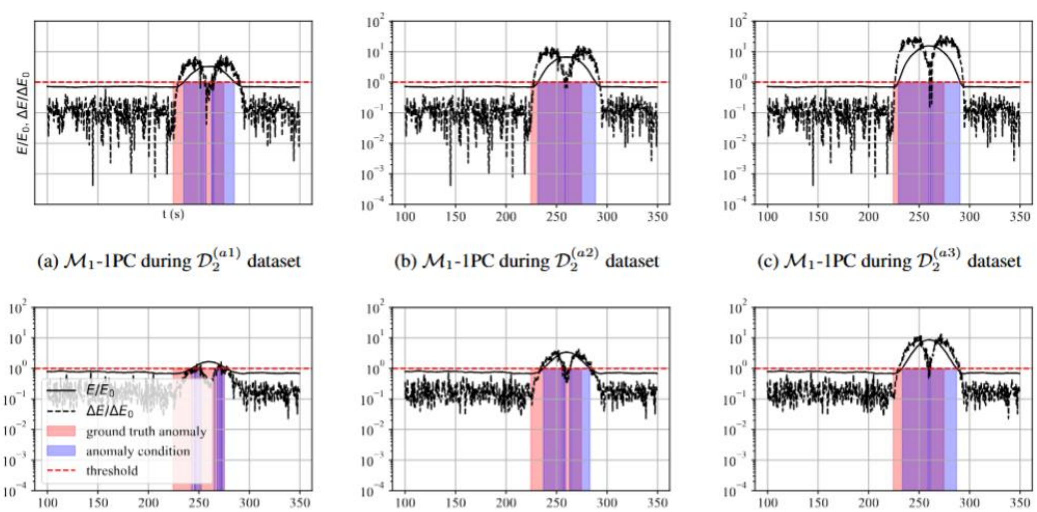
where N is the number of channels included in the model. All subsequent testing dataset (whether healthy or anomalous) are standardized using these scaler parameters. This ensures that the resulting transformed values might reflect deviations from the training dataset, allowing the model to identify anomalous behavior.

7) According to figure 11, it seems 1PC outperforms MPC despite MPC containing all information from 1PC plus additional components. Please investigate whether certain channels hinder rather than help anomaly detection, as this would strengthen the proposed methodology's motivation.

We believe the reviewer is raising an interesting point of view. In the manuscript, we clarified that the anomaly was introduced in the tower base signal, which has a relatively strong loading in PC1. This likely contributed to 1PC performing better than MPC despite having MPC containing all information from 1PC plus additional components. This is because this added information can dilute the influence of specific anomalous channels, especially when the anomaly is strongly represented in the leading component but has minimal contributions in subsequent components. Conversely, if an anomaly were introduced in a channel with weak or near-zero loading in PC1, its detection would likely require the inclusion of additional components. Thus, while MPC offers broader coverage across the feature space, it may also distribute the reconstruction error in a way that reduces sensitivity to certain localized anomalies. Therefore, the use of additional PCs may become necessary to capture anomalies in channels that do not have high loading in PC. Deeper investigation into how channel loadings influence anomaly detectability is an interesting direction for future research.

That said, we also note that in a physically coupled system such as the one studied here, localized anomalies may inherently manifest across multiple correlated channels due to system dynamics. As a result, even anomalies originating in channels with low PC1 loadings could still influence leading components via cross-correlations.

The 1PC variation demonstrates overall enhanced coverage (highlighted in blue) and reduced detection delay relative to the onset of the ground-truth anomaly (highlighted in red). ~~The reduction in detection delay is also represented~~ This is particularly evident in Fig. 12 ~~when comparing~~ which compares detection delays for 1PC ~~to MPC~~ and MPC under various anomaly



(a) $\mathcal{M}_1$-1PC during $\mathcal{D}_2^{(a1)}$ dataset (b) $\mathcal{M}_1$-1PC during $\mathcal{D}_2^{(a2)}$ dataset (c) $\mathcal{M}_1$-1PC during $\mathcal{D}_2^{(a3)}$ dataset

(d) $\mathcal{M}_1$-MPC during $\mathcal{D}_2^{(a1)}$ dataset (e) $\mathcal{M}_1$-MPC during $\mathcal{D}_2^{(a2)}$ dataset (f) $\mathcal{M}_1$-MPC during $\mathcal{D}_2^{(a3)}$ dataset

Figure 11. Anomaly detection based on $\Delta E \wedge E$ criterion during synthetically altered dataset variations of $\mathcal{D}_2$ and $\mathcal{M}_1$ model detection response, with (1) 1PC - $\mathcal{D}_2^{(a1)}$, (b) 1PC - $\mathcal{D}_2^{(a2)}$, (c) 1PC - $\mathcal{D}_2^{(a3)}$, (d) MPC - $\mathcal{D}_2^{(a1)}$, (e) MPC - $\mathcal{D}_2^{(a2)}$, and (f) MPC - $\mathcal{D}_2^{(a3)}$ variations.

scenarios. Although the MPC model includes more principal components, this added information can dilute the influence of specific anomalous channels, especially when the anomaly is strongly represented in the leading component but has minimal contributions in subsequent components. Conversely, if an anomaly were introduced in a channel with weak or near-zero loading in PC1, its detection would likely require the inclusion of additional components. Thus, while MPC offers broader coverage across the feature space, it may also distribute the reconstruction error in a way that reduces sensitivity to certain localized anomalies. Additionally, detection performance generally improved with increasing severity of the synthetically introduced anomaly. This is indicated in Fig. 12 which shows detection delay in seconds between synthetically introduced anomaly and the predicted anomaly by the models. The figure also shows a sensitivity analysis of the models to the timestep at which the data is sampled. Small timesteps (high sampling frequency) can provide reduce anomaly detection delay but at the expense of computational cost.