Replies to comments RC1

Dear reviewer,

Thank you for taking the time to review the manuscript and for the useful feedback. Below you can find my replies to your comments (your comments are in bold and italic, the remaining texts are my replies).

In this paper, the authors describe a methodology to detect malfunctioning or inoperable wind turbine drivetrain SCADA sensor readings and then correct for their effect on other diagnostic indicators. Two examples of "masking" of these faulted sensors are given, and the accuracy of the historical fault diagnoses are recalculated. The article addresses an important real-life problem and the first few Sections are well written. However, not being well-versed in autoencoders I had difficulty interpreting the Figures shown in the Results section and related text, which is also relatively brief. My biggest desire in revision would be to expand the Results section in this respect. I also offer additional comments for consideration below.

Title

•The Title does not mention drivetrains, while the Conclusion states "This paper presented a sensor-error-robust methodology for failure prediction in wind turbine drive-train components". I do believe it is important and relevant to revise the Title to be something like "Sensor-Error-Robust Normal Behavior Modeling for Wind Turbine Drive-train Failure Prediction Using a Masked Autoencoder".

- The authors of the manuscript acknowledge that there is indeed an inconsistency between the title and the statement in the conclusion. For this reason the title has been modified by adding the word drive train as suggested. However, in the conclusion a statement is added where it is explained that although the methodology has been validated on drive train data only, there is no specific methodological reason to assume that the same methodology would not work on other components or machines. In this sense the methodology should be general applicable (see lines 484-489 in the revised manuscript).

Abstract

• Line 12: I recommend this sentence be revised to "…method to improve the accuracy of data-driven failure prediction systems in practical wind turbine drive-train applications." The term "accuracy" is used in line 2 and the addition of "drive-train" mirrors my recommendation for the Title.

- The revised manuscript has been updated according to the suggestions.

3.1 Input data

• Figures 1 to 3 seem misplaced and should be moved back to where they are cited in text back in Section 3.4. Either that, or they could be cited where they are placed as examples of measurements.

- Do you mean the three figures showing examples of sensor errors (Figures 3-5) since those are referred to in Section 3.4? These were indeed in the wrong place. The problem has been corrected in the revised manuscript.

3.2 Data preprocessing

• Lines 156-176: The first bullet for Data aggregation makes what is done in the step clear by leading with "The original … data are aggregated…" and the same for the last bullet for Feature engineering with "Several additional features are derived…" However, the steps inbetween lines 156-176 are unclear in this respect. I typically understand the text, but it's not made clear what's actually done in the step. Can this be explicitly added for each?

- We've updated the manuscript according to your suggestions. The revised manuscript states more explicitly for each preprocessing step what is done.

4.1 Distributional similarity

• Figure 6 and Figure 7 (top): Here reconstruction error is shown, and although it was described to some extent in the previous text I will admit having a hard time understanding the y-axis units and, more generally, its significance.

From these figures, is it a correct understanding that one or more SCADA measurements (in this case, gearbox oil sump temp) are used to predict another related quantity (in this case, gearbox RE carrier bearing of planetary stage A1 temp)?

- The MAE uses for each wind farm as input all signals mentioned in Table 1. The task of the MAE is then to reconstruct the signals as well as possible. The MAE contains however a bottleneck, which forces it to focus on the most relevant aspects of the data. The MAE is basically a multi-input-multi-output model. It ingest all the signals and spits out reconstructions of all the signals (see Figure 1 which gives a schematic overview of the methodology). Figure 6 in the manuscript shows in the bottom plot the gearbox RE carrier bearing of planetary stage A1 temperature. The reason this signal is shown is because it suffers from sensor errors around M22-23. This does not mean that this signal was the only signal used as input. We just wanted to show what the impact is of a sensor error in one signal on the reconstruction errors of another signal (top plot). For this we selected the gearbox RE carrier bearing of planetary stage A1 temperature. We could have selected another signal, because the impacts are also clear there,

however, we chose one for the sake of clarity of the figure. The text of the manuscript has been extensively updated to make the explanation of the results more clear (see lines 356-375 in the revised manuscript).

Further, that the plot here is the error of this prediction?

- The figure at the bottom of Figure 6 shows the normalized gearbox oil sump temperature (the normalized values of the signal are shown for confidentiality reasons at the request of the industrial partner). We saw that the legend of this subplot contained an error. It mentioned the wrong signal. This has been corrected in the revised manuscript. The figure at the top of Figure 6 shows the reconstruction errors (orange line) for the model with no signals switched off (mask is completely set to 1) which corresponds more or less to a vanilla autoencoder, and the reconstruction errors (blue line) for the model with the gearbox oil sump temperature switched off when it suffers the sensor error. The differences between the two are clear. The orange line shows at the time of the sensor error a clear downward (Figure 6) or upward (Figure 7) jump in the reconstruction error. This is not the case when the sensor error is masked. This is exactly what we wat to achieve.

If I am completely off, it is probably because I am not well versed in normal behavioral models or autoencoders. Some more practical explanation of these figures, I believe would help more general wind industry readers. I believe the text in line 287 is related to this, where it asks "whether the reconstruction errors produced by the MAE when at least one sensor has failed remain comparable to those obtained under fully healthy sensing conditions".

- Yes that is correct. To see whether the MAE is truly able to neutralize the sensor errors we look to how the reconstruction error behaves when the sensor error occurs versus when it does not occur. We acknowledge that the text in the manuscript did not explain the figures well. For this reason the text was extensively rewritten to improve the explanation (see lines 356-375 in the revised manuscript).

• Figure 6 and Figure 7 (bottom): I believe the legends of these two figures are incorrect. According to the text and the figure titles, these should be gearbox oil sump temperature. The legend says "gearbox RE carrier bearing of planetary stage A1 temp." It would also be helpful to make the captions more specific, for example "Gearbox oil sump temperature sensor failure (bottom plot) and its impact on the reconstruction effort of the gearbox RE carrier bearing of planetary stage A1 temperature (top plot) without…"

- There was indeed an error in the legends. We've corrected it. The captions have been made more clear in the updated manuscript.

• Figures 8 and 9: It would be helpful to list the specific signal names on the y-axes corresponding to Table 1. For example, which temperature is "gearbox temp" – is it "gearbox main bearing temp", "gearbox oil temp", or "gearbox bearing temp"? Or some combination of them all? Figure 9b simply uses "generator" as well – it's not even listed as a temperature. I realize that it may not particularly matter which one, but I prefer consistent terminology be used throughout. More importantly, I am having trouble interpreting the significance of these 2 figures. What do they show? I understand to some extent that they relate to the text in Section 3.3.2 lines 236-250, but I am having trouble understanding how. A better summarization would be very useful here. It could also be that the figures and text be simplified. Are the particular turbines just examples, or are they each meaningful? As is, the figures contain a lot if information. Is it all important?

- Thank you for the remark. The figures are indeed somewhat complex, however, all aspects are important. To clarify the figures the captions have been updated with extra information. The terminology has been made consistent. The y-axis labels have also been made uniformly. The text that discusses the figures has also been extensively modified (see lines 376-408 of the revised manuscript).
  The figures show how well the distribution of the reconstruction errors of signals (that did not suffer from sensor errors) during sensor error windows (this is a window where at least one sensor has failed) compare to the distribution of the reconstruction errors of the same signals outside of the sensor error windows. There are different ways to quantify similarity between distributions. In this paper we selected several metrics, e.g. mean, median, sd, iqr, …. To test whether they are the same for both distributions, the Kolmogorov-smirnov test is used. If the p-values are larger than 0.025 this indicates that there is no evidence that the two distributions are different according to that specific metric.

4.2 Discriminative performance

• I believe I understand the objective of this Section, but I'm puzzled by statements like "Through detailed data inspection and expert consultation, the failure modes were identified as accurately as possible, and several confounding cases were removed. Nonetheless, some unknown factors may still affect the results." Were "failure modes" identified as one might do in a FMEA, or actual failures identified through historical failure records? How should I interpret these sentences with respect to the validation statistics presented at the end of the Section? Why is "Evaluating this capability is considerably more challenging than assessing sensor error robustness", if reliable historical failure data is provided as implied later in the Section? I am left assuming the authors selected 22 failures that should have been detected by any system, while there are others that are just left untouched for one

reason or another (not enough information, a failure that is unlikely for any system to detect, etc).

- We acknowledge that this section might create some confusion. To clarify it the text in the manuscript has been updated. The 22 failures used for failure prediction validation were given by the industrial partner and are based on historical failure records. Each case is based on an actual failure of a turbine. An inspection made clear which component had failed. The 22 failures were given to us because the industrial partner considered them as most economically relevant. This of course means that the list is not a complete collection of all failures that occurred in the turbines. This also means that there are failures that are not in the list that still impact the temperature signals and might result in anomalies being generated. This makes that determining how well the methodology can identify the gearbox and generator failures becomes more complex. To clarify the nature of the component failures used for validation Section 3.5 has been added to the manuscript (lines 335-347 in the revised manuscript).

• Figures 10 and 11 (bottom) show "Anomaly concentration score" but this is not described in the text anywhere. Please describe. Additionally, although it's described as fact that Figure 11 clearly shows a failure being identified, it is not clear to me at all (or at least less clear than Figure 10, in which there is at least a somewhat consistent rise in the reconstruction error and anomaly concentration score – although moving means might show this more clearly). What should be made of the rise in anomaly concentration score at Y2, compared to at Y5.5?

- We acknowledge that anomaly concentration score was not sufficiently explained. We added the explanation to the revised manuscript (lines 449-451). The rise in Y2 is at this time unclear. It is likely that an unknown turbine event or problem with the turbine caused the temperatures to rise. This is the subject of further investigation. It is unlikely that it is caused for example by seasonal fluctuations because this would also be visible in other years.

• The caption of Figure 11 describes the reconstruction error for the generator phase 2 temperature and ties it to a "Gearbox failure", which is then referred to in more detailed fashion in the text as the "gearbox rotor end main bearing". If I understand correctly, does this mean that a main bearing failure was detected through a generator phase temperature?

- Thank you for pointing this out. This was a typo in the caption. The top plot does not show the reconstruction error for the generator phase 2 temperature but for the gearbox RE main bearing temperature. We've corrected the mistake.

Conclusions

Line 354: Here it is stated "The validation data examined in this work contained a substantial number of sensor failures, many of which persisted for extended periods." However, Section 4.1 only shows 2 examples and Figures 8 and 9 are less than clear to the reviewer. Is there a better way to demonstrate the ability of the MAE to detect sensor errors than Figures 8 and 9 and accompanying text?

- The figures 8 and 9 do not show that the MAE is able to detect sensor errors. The purpose of the manuscript is not to show that the MAE can detect sensor errors, but that it can detect reliably component failures in a context with sensor errors. The figures show that the impact of sensor errors can be neutralized by the MAE. To clarify the figures 8 and 9 the text in the manuscript has been extensively revised (see lines 376-408). It was not possible to show all sensor errors in figures like 6 or 7. This would take up to much place if done in separate figures or make the lines difficult to see if all sensor errors were plot is the same figure.

• Line 361: Here it is stated "The results demonstrate that the model successfully maintains consistent reconstruction-error distributions when one or more sensors fail." I believe this is from Figures 8 and 9, but I'm left not quite understanding the matter. The same for "Even in the extreme scenario where all gearbox signals were absent for more than two years, the model produced generator signal reconstruction errors of acceptable quality, although with some signs of performance degradation" from Figure 9b, but I do not know how.

- We acknowledge that the explanation of figures 8 and 9 was insufficient. For this reason we've extensively rewritten the text accompanying the figures (see lines 376-408).
  "The results demonstrate that the model successfully maintains consistent reconstruction-error distributions when one or more sensors fail.": This is shown by the p-values of the KS-tests which are in general larger than the 0.025 threshold.
  "Even in the extreme scenario where all gearbox signals were absent for more than two years, the model produced generator signal reconstruction errors of acceptable quality, although with some signs of performance degradation": This is visible in the p-values in Figure 9b which are in most cases larger than the 0.025 threshold. However, the values are somewhat smaller than for the other wind farms where less signals suffer from sensor errors. For this reason we can conclude that there is some degree of degradation.

• Line 364: Here I believe the "reliably distinguishing healthy from degraded operating conditions", which is vague and could be interpreted as still discussing

sensor errors, is discussing the gearbox and generator faults. If so, I recommend changing this to "accurately distinguishing healthy from damaged drive-trains".

- We agree that the sentence is unclear. We've changed it in the updated version of the manuscript to your recommendation (lines 496-497).

I hope this sufficiently addresses your comments.

Sincerely,