Response to reviewers’ comments:

Thank you for reviewing out paper at this late stage in the review process. We are grateful for your valuable comments and hope that they improved the quality of our paper. In particular, based on your comments, we added several clarifications and give additional explanations regarding the methodology applied in the revised version of the paper.

In the following, you can find our answers to the comments. Line numbers in the reviewers’ comments refer to the previous version of the paper (initial submission); whereas line numbers in our answers refer to the revised version.

A revised version of the manuscript is prepared and the corresponding paragraphs are marked in the manuscript.

Reviewer 3: (marked cyan in the manuscript)

1.) My biggest concern with the paper in the current form is a bit the downside of its ambitious nature. The large number of techniques used, intermingled with one-another, sometimes makes it difficult to keep track of the exact methodology followed. In particular this plays a role when the performance and the uncertainty are discussed. Questions like: "Are we looking at the error on the total damage (sum over a year) or at the 10 minute errors on damage?)", "Do we retrain the ANN for every iteration of our 1000 uncertainty assessments?" regularly popped up.
At times it was hard to follow what the boxplot meant that were presented to us. I feel some extra clarifications here and there how some methods are exactly employed might be beneficial. In particular when uncertainties are addressed.

We completely understand that, due to the various techniques used, it is not always easy to follow. We revised the paper in order to improve the comprehensibility.

Frist, we tried to answer your questions as good as possible (see for example answer 11 (unclear content of box plots), 13 (not stated EOCs), 14 (insufficient description of the ANN applied) and 15 (incomprehensible descriptions in the context of uncertainty and performance)) and revised the paper accordingly. We hope that these points are much clearer now.

Second, we added some more explanations/clarification to the paper. These additional comments should make clear, for example, what errors we are looking at in each figure/section, what type of uncertainty is meant and where exactly the results discussed are shown in the box plots.

2.) *Perhaps it might be a consideration to drop certain segments (e.g. the one on computation time felt unnecessary in my book) or even reduce the role of GPR (which due to their timely exit, I don't feel ended up contributing much to the paper) in favor of a more clear narative. But I leave this up to the authors.*

Thank you for proposing several parts of this work that could be dropped. We carefully thought about your suggestions. Nonetheless, we still think that both parts are quite relevant. GPR is, if not performing any statistical evaluation, more suitable compared to ANN. The results are slightly more accurate (see for example Figure 19). Hence, if computing time is not crucial, GPR is definitely an alternative and could be analysed in more detail in future, for example, in the context of 1s SCADA data. If we keep the analysis of GPR, it is also essential to keep a section on computing time. Otherwise, there is no reasoning for focusing on ANN instead of GPR. We hope that the additional explanations and clarification, see, for example, answer to comment 1, still lead to a clearer narrative.

3.) *How the authors relate their work to e.g. the work of the group of Prof. John D Sørensen? Who approaches this problem more from a probabilistic angle?*

Thank you, for pointing out that we have not discussed the work of Sørensen, which is definitely relevant in this context.

In general, a probabilistic approach, e.g., Mai et al. [1], is a fourth option in addition to the simple extrapolation, the binning approach and machine learning approaches. It uses basically the same idea of correlating EOCs and fatigue damage and is relatively similar to the binning approach. For the probabilistic approach in Mai et al. [1], bins of EOCs are defined based on theoretical statistical distributions that are fitted to the EOC data instead of using empirical distributions as it is done for the binning approach. Moreover, the probabilistic approach goes one step further and considers the entire fatigue (or stress-range) distributions in each bin, whereas the binning approach just uses mean values. The advantage of the probabilistic approach is that less information is lost (no averaging in each bin). The disadvantage is that additional statistical uncertainty is introduced (especially for limited data) when fitting the distributions for the EOCs and the fatigue in each bin.

We added a statement regarding this approach to the introduction. However, as you pointed out, this work already covers a larger number of extrapolation techniques. Therefore, we do not further consider the probabilistic approach in this work.

4.) One concern when working with 10 minute damages is the role of long term fatigue cycles, (E.g. See "Marsh e.a. - 2016 - Review and application of Rainflow residue process" and recently confirmed in "Sadeghi e.a. - 2022 - Fatigue damage calculation of offshore wind turbines' longterm data considering the low-frequency fatigue dynamics"). Obviously such information is lost when binning as the temporal sequence is lost, moreover it can be difficult to replicate when drawing from a random distribution in forecasting. A similar concern applies for sequence effects. Do the authors have any thoughts on the matter?

This is a very interesting point of discussion. Although a thorough discussion of this topic is definitely out of the scope of this work. We added some thoughts regarding this topic to Section 3 and the outlook.

In general, the challenge of long-term fatigue cycles is the more pronounced the higher the material exponent *m*. In this work, a material exponent of 3 (at least for the relevant first part of the S-N curve) is used. Hence, according to Marsh et al. [2] and Sadeghi et al. [3], the error should be below 10%, which is acceptable for the current application.

Still, there are some possibilities to improve the Rainflow counting, which become even more relevant if such approaches are applied to rotor blades for which much higher material exponents have to be applied.

First, instead of using 10-min short-term damages, short-term damages for longer periods, e.g., 3 hours, could be used. On the one hand, this reduces the error due to long-term fatigue cycles by about 50% [2]. On the other hand, the correlation of relatively quickly changing wind conditions and fatigue damages might become less accurate and less short-term damages will be available. Hence, when increasing the length of the short-term intervals, these trade-offs have to be carefully balanced. More detailed analyses regarding the optimal length of the short-term periods are left for future research.

And second, it might be possible to derive a correction factor. For this purpose, analyses similar to those by Marsh et al. [2] or Sadeghi et al. [3] could be conducted for the entire measurement period. The measurement period is normally at least several months long. After 2-6 months the error in damage levels due to long-term fatigue cycles has converged [2]. Hence, using the entire measurement period, it should be possible to derive a correction factor that is applied to the extrapolation afterwards. Surely, this only works if the correction factor remains fairly unchanged for different EOCs and during the entire turbine life. Although this assumption has not been tested so far, it could be a valid one in the first place.

For sequence effects, a similar approach could be applied. Again, for rotor blades, sequence effects are much more relevant.

5.) Line 132 : authors state that mean values are not relevant for fatigue damage. In fact they are saying to ignore mean stress effects, this is a (valid) assumption, but I wouldn't say mean values aren't relevant.

You are completely right. We changed the statement to make clear that neglecting the mean stress is just an assumption.

6.) Figure 8, 9: Scatter plots with such a big dataset, don't show the density differences of the data which I assume is more densely packed towards the middle of the cloud. Perhaps working with a different style plot.

Thank you for pointing out that the scatter plots do not provide the relevant information. We totally agree. Therefore, we changed the plot style. Histograms are used now.

7.) How does 315 relate to the dominant wind direction?

315° means that the wind comes from northwest. This does not correspond to the dominant wind direction, which is southwest. However, this should not be problematic for this study. As explained in line 189 to 196, a use of other strain gauges, interpolations between the different strain gauges etc. are possible and would be required in an industrial context, but not for this study. Actually, the strain gauge approximately perpendicular to the dominate wind direction probably even leads to less pronounced correlation effects with the wind speed. Hence, the performance of the binning approach and the ML approaches might even be slightly reduced by the choice of the strain gauge. More detailed analyses are out of the scope of this work.

8.) Line 215 : the discussion on the "horizontal part" is interesting, but perhaps is phrased a bit too simplistic given there is a common terminology for it: 'Fatigue limit'. Perhaps a wording "the S-N curve does not account for a fatigue limit in the material, i.e. a horizontal part at low stress cycles, …" is more appropriate as I do appreciate the authors' effort of linking it with such a strong visual cue

The statement was changed accordingly.

9.) Line 280 : I personally find the statement "fill up about 40% of the bins (cf. Fig.9)" overly dramatizes the problem. Yes Figure 9. Does show a lot of "empty space" but in fairness those missing conditions also represent improbable conditions (e.g. very high waves at low wind speed), so while 40% of the bins might be empty, their combined probability will lay well below 40%, probably even reasonably below 1%. The story of empty bins is correct, but might not have too much impact on the total outcome due to low probabilities

We completely agree with your comment and reformulated the statement. Now, it should be clear that on the one hand, there are a lot of empty bins, but on the other hand, these empty bins feature very low occurrence probabilities so that they do not influence the total outcome significantly.

10.) Eq 13: the authors opt to work with a absolute error on their prediction. But I'm curious to see the results with a sign. E.g. given the conservative nature of filling empty bins do we end up with a conservative damage estimate? In contrast ANN doesn't have this built in conservatism, does it still produce conservative predictions (UPDATE: they later present results without a signed error)

Yes, signed errors are quite interesting in order to analyse conservatism. However, box plots as shown frequently in Section 4 are quite useless when using signed errors. Positive and negative errors would cancel out, so that low mean errors can occur even if the prediction in inaccurate in all situations. Hence,

for the box plots, we stick to unsigned percentage error. Nonetheless, in Section 4.1.1, we now refer the reader to Section 4.3 for signed errors.
For a discussion of the conservatism of the different approaches, see answer to comment 16.

11.) Figures 14: just to be 100% sure, each box in the boxplot only contains 13 samples?

This is correct. We added a brief explanation to make this clearer.

12.) Figure 15 : is it possible to have the axis the same size as the adjacent Figure 14? I feel it is fair to compare these results.

The axis of Figure 15 has been changed to be in accordance with Figure 14. Nonetheless, for a direct comparison, it is referred to Figure 19.

13.) In section 4.1.4: for the binning and functional extrapolation, I assume you are still using the wind speeds for binning and extrapolation? Perhaps this should be emphasized a bit stronger.

Thank you for pointing out that it was not clear enough that we still use wind speeds for the binning/correlation in Section 4.1.4. We added a brief clarification to the section.

14.) Section 4.1.4. : how are the "discrete" status's fed into the ANN? As numeric values (operation=1, parked=2)?

Using discontinuous inputs together with a continuous one in an ML approach is not straightforward. Therefore, we decided not to consider the turbine status directly in the ML approach, but to train an ANN or GPR model for each turbine status separately. Using this approach, discontinuous turbine statuses are not problematic. Since our approach has not been explained sufficiently so far, we added an explanation to Section 4.1.4.

15.) Section 4.2 : I'm a bit confused about this whole segment. For the binned technique you explicitly mention that the processing times goes up dramatically because you need to fill in the empty bins. I conclude that in your 1000 prediction, for every prediction you refill the bins and calculate the means of each bin? (i.e. this is the training phase of the binning method, retraining the damage map for every iteration) Do you do something similar for the ANN and GPR? So each iteration drawing new training data and then fitting a new model to said training data, so basically training a 1000 different models? Or are you just 1000 times evaluating into a single well performing model?
I then understand why computing time blows up. But if you did so, I don't think I would ever consider the uncertainty of ANN by training 1000 models (without much oversight). I rather have an uncertainty come from feeding a small number of ANNs 1000 different validation datasets (e.g. drawn using bootstrapping) and quantify the uncertainty from those predictions.
If you don't train an ANN for every evaluation then it is perhaps an unfair comparison as the training of the binning strategy is building these damage lookup tables.

Thank you for pointing out that the explanations given in Section 3.2.4, 4.2 and 4.3 have not been sufficient to understand the exact approaches applied for the uncertainty assessment. Probably, the most confusing point is that two different types of uncertainty are considered in this work. First, the uncertainty due to limited strain data is considered (Section 3.2.4 and 4.3). This type of uncertainty exists for all extrapolation approaches and is determined using a bootstrapping approach. And second, there is some kind of model uncertainty for the ML approaches. This uncertainty is due to randomly chosen initial weights (ANN) or subsets (GPR). It does not exist for the simple extrapolation and binning approach and is mainly relevant if there is no assessment of the uncertainty due to limited strain data. It is discussed in Section 4.2.

The uncertainty due to limited strain data is due to changes in the training data. For example, how probable is it that we get the same extrapolated damage for 2011 if we use training data from 2017 instead of 2016? This question is quite relevant to be answered. Only if it is probable, we can trust our results. If not, we need more training data (or another extrapolation approach).

Since we only have three years of measurement data here, we cannot determine the uncertainty directly, but have to use bootstrapping. This means that we generate new training data from the same year, e.g., we sample 52596 10-min intervals with replacement (this sums up to one year) from 2016 and repeat this 1.000 times. This gives us 1.000 "different" years, which we then extrapolate to 2011.

Each of these 1.000 "different" years in now treated as if it is our entire available measurement data. Hence, for the binning approach, for each run, mean damage values are determined, i.e., 1.000 damage lookup tables are determined. For the ML approaches, for each run, a new model is trained, as new measurement data is available.

The model uncertainty is due to randomly chosen initial weights (ANN) or subsets (GPR). Hence, if an ANN or a GPR is trained using precisely the same measurement data, the model will slightly differ each time it is trained. Since the trained model differs, the extrapolation results differ as well (cf. Figure 15, 18 and 19 in the paper). Therefore, for a single measurement data set (i.e., all sections except 4.3), the training of the ANN/GPR should be done more than once (e.g., 100 times) to guarantee some statistical evidence. The uncertainty of the 100 runs is averaged out.

If both types of uncertainties are considered at the same time, i.e., an assessment of the uncertainty due to limited strain data for ANN/GPR, it is not necessary to train 100 models for each of the 1000 "different" years, since the model uncertainty is already implicitly averaged by the 1000 "different" years.

We added new explanations and clarifications to 3.2.4, 4.2 and 4.3 and hope that the procedure is much clearer now.

16.) Figures 20 and 21 : how come that in Figure 20 the binned strategy is non-conservative (predicted damage > actual damage) and in figure 21 it is the other way around? Where the empty bins not conservative enough?

This is a good question.

First, the empty bin filling is in general conservative, but not always. The reason for this is that empty bins are filled using the highest value of the bins around the empty bin. Since the highest value is used, it is conservative in most cases. However, in some cases, it can also be non-conservative, e.g., a bin of really extreme conditions is empty and it is filled with the damage value of less extreme conditions since there are no filled bins for more extreme conditions.

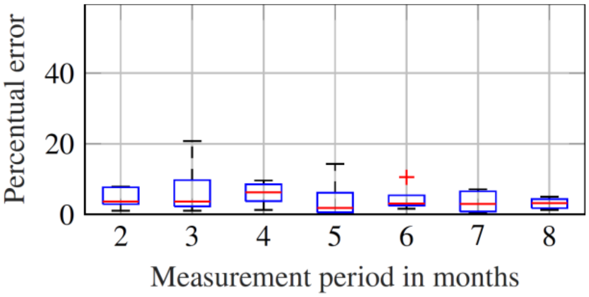
However, since the results presented in Figure 20 and 21 are based on one-dimensional binning (featuring no or nearly no empty bin filling), the reason for the non-conservative and conservative prediction is probably a different one. Here, the binning approach determines the mean damage in each wind speed bin for the given training data. Effects of other EOCs are not taken into account explicitly, but are just represented by the mean damage in the wind speed bin. For Figure 20, the damage values in the training period for a given wind speed are probably lower than the damage values in the extrapolation period for the same wind speed. This could be caused, for example, by higher turbulence intensities during the extrapolation period or less down-times of the turbine. For Figure 21, it is just the other way round. Now, the conditions during the training period were more damaging for the same wind speed and therefore, the extrapolated damage value is lower than the real one.

Such (random) changes in conservatism are normal for approaches using correlations, since they do not take into account all effects influencing damage.

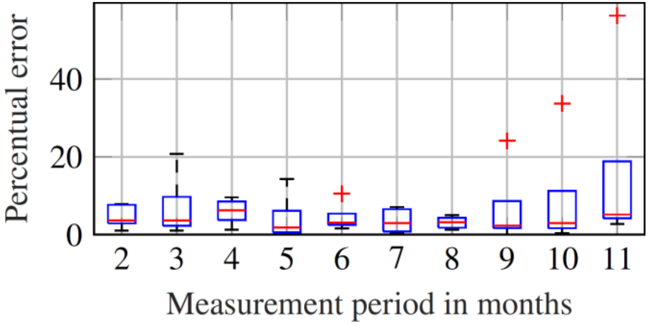
If it is necessary to be conservative in all cases, an option could be, for example, to use a high percentile of the distributions shown in Figure 20 and 21 (e.g., 99th percentile), yielding conservative extrapolations in both cases.
Some explanations regarding this topic are added to the paper.

17.) Figure 25 : the role of when the measurements start is an interesting comment (and understandable), I assume a similar mechanism may also reduce the amount of training data required for the ANN to converge?

In general, you are correct. A start of the training period in winter can also be beneficial for the ANN. However, as shown in Fig. 24, the ANN convergences quite quickly to percentage errors of 0-20%. After about 5 months, there is no significant improvement anymore. Hence, starting the measurement in winter, reduces this time from 5 months to 2 months (see figure below). The final error should remain unchanged, since – after about 9-12 months – there is not difference between measurement campaigns starting in winter or in summer.



However, comparing the results shown in the previous figure and in Fig. 24 of the paper, it becomes apparent that measurement campaigns starting in winter, i.e., the previous figure, seem to yield smaller percentage errors. However, as this is not logical, we assume that this is only a statistical artefact due to the fact that only 5 "different" measurement periods for campaigns starting in winter are available. When increasing the measurement period to 9-11 months, the percentage error increases again to the already known value of 0-20% (see figure below).



To summarise: Yes, measurement campaigns starting in winter can also be beneficial for the convergence behaviour of the ANN. However, as the ANN convergences quite quickly anyway, this is less relevant. The final error for long measurement periods is not influenced.
In the paper, we added a brief statement that starting in winter is also beneficial for the ANN as well, but we abstain from discussing it in detail.

[1] Mai, Q. A., Weijtjens, W., Devriendt, C., Morato, P. G., Rigo, P., & Sørensen, J. D. (2019). Prediction of remaining fatigue life of welded joints in wind turbine support structures considering strain measurement and a joint distribution of oceanographic data. *Marine Structures*, *66*, 307-322.

[2] Marsh, G., Wignall, C., Thies, P. R., Barltrop, N., Incecik, A., Venugopal, V., & Johanning, L. (2016). Review and application of Rainflow residue processing techniques for accurate fatigue damage estimation. *International Journal of Fatigue*, *82*, 757-765.

[3] Sadeghi, N., Robbelein, K., D'Antuono, P., Noppe, N., Weijtjens, W., & Devriendt, C. (2022, May). Fatigue damage calculation of offshore wind turbines' long-term data considering the low-frequency fatigue dynamics. In *Journal of Physics: Conference Series* (Vol. 2265, No. 3, p. 032063). IOP Publishing.