Response to reviewers' comments:

We want to thank all reviewers for their comprehensive reviews that helped us to improve the quality of the paper especially with respect to a more detailed discussion of the short-term fatigue calculation and of the most relevant environmental and operational conditions.

In the following, you can find our answers to the comments. Line numbers in the reviewers' comments refer to the prevision 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.

<u>Reviewer 1:</u> (marked cyan in the manuscript)

1.) A main result of the paper is that 'wind speed' is the most important parameter for temporal extrapolation of fatigue damage. An important part of the storyline is that the paper uses data which is readily available over the life of the offshore wind turbine. However, the authors used mostly data from the met mast FINO1 to determine the wind speed. Most wind farms do not have access to a nearby met mast. The content of the paper is only useful, if the parameter 'wind speed' is purely extracted from SCADA (with acknowledgement of its limitations). Can you add the results for wind speed based on SCADA only?

Thank you for this valuable comment. It is definitely correct that the presented approaches are only useful for real applications if they yield accurate results even if only SCADA data is used. For the validation of the approaches in this paper, we still think that the usage of "high quality" data is expedient. Hence, for our analyses, we use met mast and SCADA data. Nonetheless, we agree that it has to be shown that the usage of SCADA data only is possible as well. This is why we added a brief discussion of this topic to Section 2.2 (l. 140-142) and show a comparison of some results (met mast + SCADA versus SCADA only) in Appendix A.

2.) *The authors use environmental conditions and the turbine status for damage extrapolation. However, they do not consider continuous operational SCADA data, such as power output and pitch angle (which may even be more robust parameters than wind speed). Can you explain why you do not take this data into account?*

We agree that it could be a valuable alternative or addition to consider continuous operational SCADA data. This is why we added comments on these operational conditions (OCs) to Section 5 (l. 652-657) and Section 6 (l. 675).

Nonetheless, there are also some good reasons, why we do not use them in this work. First, the turbine status already combines several of these continuous OCs. There are 15 different turbine statuses from "normal production" over "run-up" to "grid loss". Hence, the mentioned OCs are already partly covered. Perhaps, this fact was not clear enough so far, so that we added a brief explanation regarding the turbine status to Section 2.2 (l. 146/147). Second, wind speed and power output are (at least during normal operation) highly correlated. Hence, using both can be problematic, e.g., machine learning algorithms would have to handle depend inputs. And third, using OCs instead of wind speeds is no alternative. In non-operational conditions, OCs have little to no informational value, e.g., for wind speeds below cut-in and above cut-out, the power output is constantly zero although loads are quite different.

After all, the usage of continuous operational conditions is an interesting alternative, which has definitely to be discussed in this paper. Nonetheless, it is not straightforward to incorporate additional OCs, which is the reason why we focus on environmental conditions and the turbine status in this work.

3.) *I disagree with the conclusion that long-term extrapolation is possible even if the OWT has been modified. This may be the case for the example of this paper but a generalization can be wrong and dangerous. A change of operational conditions can have a significant effect on tower loads. For example, aerodynamic imbalance due to blade pitch misalignment increases tower loads. Such an imbalance may occur after some years of operation (e.g. after end of the measurement campaign) and stays often undetected. If such factors are not represented in the measurement period, the extrapolation will underestimate loads. Being unaware of such effects is, in my opinion, the largest limitation of the presented extrapolation approach.*

We totally agree that extrapolations are only possible if no significant changes occur. This means, if the learned correlation between EOCs and fatigue damage is no longer valid, none of the approaches can yield accurate results. Hence, we reformulated the corresponding conclusions (l. 578-580 & 644/645).

With our conclusion we wanted to express that long-term extrapolations might be possible. A significant repair action does not necessarily change the correlation, as shown for the present example. This fact cannot be generalised, but it is a hint that it might be possible to use such approaches for extrapolation in more situations then expected.

In an industry application, the procedure could look as follows: Run a measurement campaign for at least one year. Then, after the strain gauges have failed, it might not be necessary to conduct a full new campaign. First, it is tested whether the correlation has changed. For this purpose, a few measurements (e.g., only strain gauges at one position) are done for a few weeks. If the correlation has clearly changed, a full new campaign is required. However, if the test campaign indicates that the correlations are still valid, no new measurements are required.

We added several new statements and explanations to the paper (l. 562-564, 578-580, 583-589 & 644/645) in order to clarify the limitation of this work, but also to explain the possible gain.

4.) L. 27: This is rather an industry guideline, not an international standard. It was written by DNV GL instead of an international expert committee.

It is true that the DNVGL-ST-0262 is not a classical international standard, as, for example, the draft standard 61400-28 by the IEC. Nonetheless, DNVGL-ST-0262 is called “standard” by the DNVGL and is used within the context of certifying wind turbine lifetime extensions.

To prevent misunderstandings, we slightly reformulated the sentence in the paper (l. 31 & 46).

5.) L. 83: The literature review is nice for all aspects but machine learning. There has been much more work to estimate loads based on SCADA data. Can you add these, please?

Thank you for pointing out this deficit. We added several additional citations regarding wind turbine load estimation to the introduction (l. 79-84).

6.) L. 84: In my personal opinion, this is rather a result and should not be part of the introduction.

The statement was removed from the introduction.

7.) L. 123: How was the post-processing done? Please explain your methodology, how replaced unrealistic data, etc.

Thank you for pointing out that we have not given any details regarding our post-processing. We added some information to the paper (l. 129/130 & 151/152).

For your information: First of all, the provided data was already post-processed by the RAVE initiative (i.e., the institution providing the data) to exclude some clearly wrong data. Then, we excluded erroneous data using semi-automatic methods. More specifically, for strain data, “zero-values” are removed automatically. Moreover, for strain data as well as all EOCs, unrealistically high and low values are also removed automatically. The thresholds to exclude a value are chosen manually after a visual inspection of the data. For example, for significant wave heights, all data above 10m and below 0m is removed, after having checked, that significant wave heights above 10m do not occur in reality. In addition, for the environmental conditions, if several consecutive values are precisely the same, they are removed. Finally, we performed a visual inspection to exclude some other, obviously wrong data values manually.

8.) L. 130: Data from met masts are not available for most wind farms. Why did you use this instead of the SCADA data? Of course, the data has better quality, but it is not really applicable outside research. Can you show results with wind speed from SCADA only?

See answer to comment 1

9.) L. 135: Why are you not using continuous SCADA data such as power output, rotor speed, pitch?

See answer to comment 2

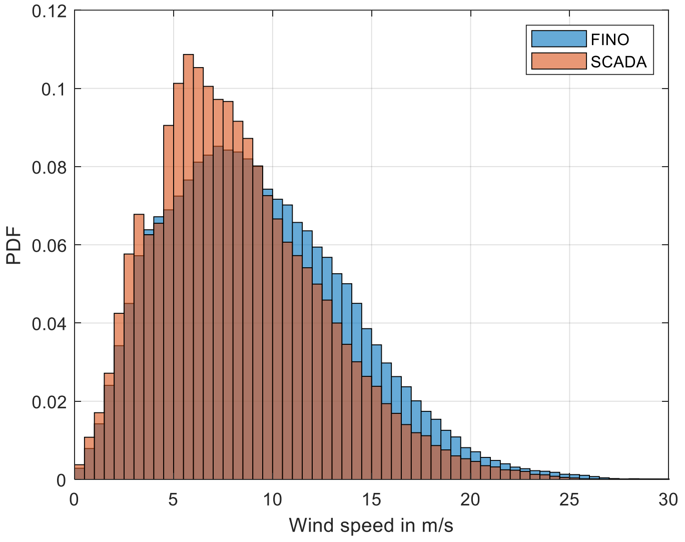
10.) L. 139: How much data has been replaced? Is there any statistical difference between FINO 1 and SCADA data?

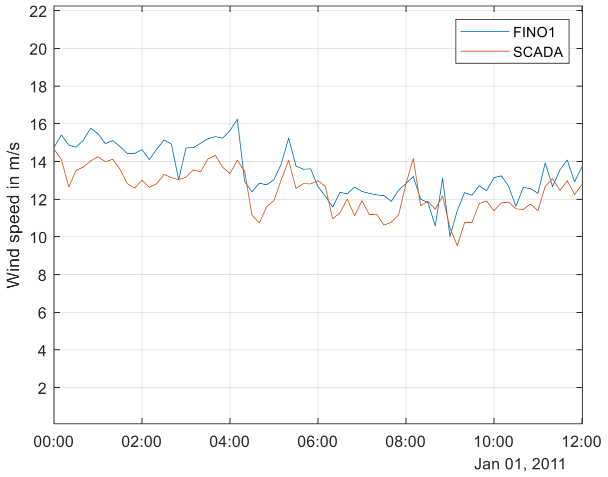
For the environmental conditions, for about 95% of the ten-minute intervals, data are available from FINO1. An additional 3% of the ten-minute intervals are filled up using SCADA data. Only for about 1.6% of the data, no useful data are available. A short comment about the amount of replaced/available data was added to the paper (l.137/138).

Regarding the statistical differences, the mean values are slightly different ($\mu_{SCADA} = 8.3m/s$ and $\mu_{FINO} = 9.3m/s$), but the distribution (see below), scattering (coefficients of variation: $CV_{SCADA} = 0.50$ and $CV_{FINO} = 0.50$) and time series (see below) are similar. The difference in mean values is due to a reduced wind speed behind the rotor (SCADA data). Some information regarding the statistical differences are given in the paper now (l. 156-160).

Surely, the difference in mean wind speeds does have an influence. However, here, EOCs are only used for the correlation, e.g., binning. Hence, using inconsistent data (i.e., FINO + SCADA) or “poor” data (i.e., SCADA) might lead, for example, to some ten-minute intervals that are sorted into the “wrong” bin (e.g., a wind speed of 9.3m/s in the measurement period is sorted in the bin 8-9m/s). However, the same data source is used for the measurement and extrapolation period. Hence, for the extrapolation, the “wrong” bin is also weighted according to the “wrong” sorting (e.g., a measurement of 9.5m/s in the extrapolation period increases the probability of the bin 8-9m/s). This is why these effects should mostly cancel out each other (see answer to comment 1 as well).

Problematically would be the usage of FINO data for the measurement period and SCADA data for the extrapolation period. In this case, the effects do not cancel out each other anymore (e.g., 9.3m/s counts for 9-10m/s in the measurement period, but for 8-9m/s in the extrapolation).





11.) L. 156: Do you have a reference for this?

There is no real reference for this, since this is an assumption that is discussible especially when taking facture mechanics into account. Hence, we reformulated the sentence (l. 171-174) to make clear that using the Palmgren-Miner rule for steel components is common practice. Moreover, the error due to the assumption of linear damage accumulation is less severe for steel components compared to the error when assuming it for composite materials in rotor blades.

12.) L. 159: Why do you use data from only one strain gauge instead of calculating resulting stresses (conservatively gives maximum stress at some fictive spot around circumference)?

Thank you for pointing out that our intension was not clear so far. The idea is to keep the verification process as simple as possible without including any unnecessary effects. Surely, for a real fatigue analysis, a maximum around the circumference (conservative) or various values around the circumference would be calculated by superimposing the strain signals of the four sensors. Hence, some kind of spatial interpolation would be used.

For the validation of the temporal extrapolation, it is not necessary to apply a spatial interpolation. Surely, it would lead to different fatigue values, but in the end, it is still a strain signal that would be treated exactly the same way in all further steps. Moreover, it might lead to unwanted additional effects, for example, complex correlations with the wind direction.
We added some explanations to Section 3 (l. 177-182).

13.) Fig. 10: How do you treat residual cycles?

Since all time series have a duration of ten minutes and most cycles with significant amplitudes have frequencies of at least 0.1 Hz (leading to at least 60 cycles in 10 minutes), the effect of residual cycles is relatively small. Still, you are right that some more information could be helpful. For the rainflow counting, an algorithm by Nieslony [2] is applied (l. 191). This algorithm counts half and full cycles. In our work, we round residual cycles to full cycles if they are at least half cycles and neglect them if they are less than half cycles.

14.) L. 244: What is the minimum number of data points that a bin must? Do you accept bins with only one data point?

Since the process of filling up empty bins adds quite a lot of uncertainty to the extrapolation process, we decided to set the minimum number of data points in a bin to 1. We know that an extrapolation based on bins with a single value in it is not very reliable, but the alternative of filling up bins conservatively is even worse.
We added a sentence clarifying that only bins with no data are considered to be empty to Section 3.2.2 (l. 284/285).

15.) L. 246: How much of the data set was filled up?

The number of empty bins that are filled highly depends on the bin types and sizes. For one-dimensional bins, no or nearly no bins are filled up (e.g., for $1D_{v_s}^{10}$, no bins are filled up and for $1D_{v_s}^{60}$, one bin is filled up). For three-dimensional bins, about 80% of all bins are filled up, i.e., are empty in the measurement period. On the one hand, it is logical that a high number of bins remain empty, since, for example, low wind speeds with high wave heights do not occur (see Fig. 9 in the paper). On the other hand, this sounds a lot. However, at the same time, about 80% of the bins feature an occurrence probability of zero during the extrapolation period. For those bins, the conservative filling is not relevant.
A short comment on the dependency between bin types and the number of empty bins is added to the paper (l. 282-284).

16.) L. 277: For all three methods, I miss a discussion on how to deal with newly incoming data. How you update your extrapolation once you get new data (continuously or discontinuously). How would this work with each method?

That you for this comment. It is true that we have not discussed this interesting topic. The reason why we have not done this so far is that all approaches are not very time consuming (at least when they are suitable for uncertainty analyses as well; cf. Section 4.3). Hence, the most straightforward way to update the extrapolation once receiving new data, is to just rerun the entire extrapolation with more data. For the binning approach and the machine learning approaches, it would also be possible to just update the occurrence probabilities of the EOCs continuously and the “load levels” discontinuously, e.g., once a month.
Since this is definitely an interesting and quite relevant topic, we added some information to Section 3.2 (l. 233-238).

17.) L. 313: Why does it matter that the years are consecutive?

You are right that it is not relevant that the two years are consecutive. However, it is relevant that long-term effects are not disturbing the analysis. Since the two consecutive years are from 2015 to 2017 and the third year is 2011, in between, long-term changes might have occurred.
We added a statement to Section 4.1 (l. 352/353) to clarify why we only take the two consecutive years into account at this point. The third year is analysed in more detail in Section 4.5.

18.) L. 331: How did you choose these?

The environmental conditions considered in this work are chosen based on the available data and some sensitivity analyses conducted by the authors and being available in literature, e.g., [1].
Since it has not been stated so far, we added a brief clarification to Section 2.2 (l. 144/145).

19.) L. 333: How did you choose the discretization? How sensitive is the result to the discretization?

Thank you for pointing out that we have not discussed our selection for the discretisation.
A first guess regarding the bin sizes was done based on expert knowledge and literature values (e.g., bin sizes recommended for the design of wind turbines in current standards). For example, for the wind speed, we started with a bin size of 1m/s.
Then, we tested different bin sizes and took into account how the percentage error changed and how many bins remain empty. For example, small bin sizes that might be appropriate for one-dimensional bins, are not adequate for three-dimensional bins, since ten bins per dimension already lead to 1000 overall bins in three dimensions.
In general, if reasonable bin sizes are chosen, the sensitivity is relatively small. This observation has already been stated in the paper before (line 386 of the revised paper) and was visible in Fig. 14 for bin sizes of 5m/s ($1D_{v_s}^{6}$), 3m/s ($1D_{v_s}^{10}$) and 0.5m/s ($1D_{v_s}^{60}$) for one-dimensional bins. We now added some additional bin sizes ($1D_{v_s}^{15}$, $1D_{v_s}^{30}$, $1D_{v_s}^{120}$) for the one-dimensional bins to Fig. 14 to clarify the limited sensitivity. Moreover, we reformulated the corresponding statement in the text (l. 386/387).

20.) L. 340: Wind speed is not a very robust measurement. How about power output and pitch instead?

See answer to comment 2

21.) L. 346: Can you please show the sensitivity of the results to the bin size, at least for the 1D bins?

See answer to comment 19

22.) L. 362: This may be different for other points of the structure. The further you go down, the more important wave loads may become relative to wind loads.

We completely agree that the conclusions in this section are – to some extend – limited to this turbine and the considered location. We added such a clarification to Section 4.1.2 (l. 404-406). Moreover, we now refer to Section 5, where the limitations regarding the general applicability are discussed in more detail (l. 639 & 661).

23.) L. 393/400: This is a questionable conclusion. Do you understand why this is the case? I believe this is only the case in this example here because the turbine status is similarly distributed between the measurement period and prediction period. If the prediction period would have significantly more downtime (due to a long repair, grid requirements, etc.), then the turbine status will become important. Please explain.

You are correct that this conclusion is an unfounded generalisation of the results. Only for this turbine, these measurement periods etc., we showed that OCs are of minor importance. This is why we reformulated the statements in Section 4.1.4 (l. 436-438). Moreover, it is referred to Section 5, where the limitations of this work are discussed in detail.

Regarding a potential explanation for this fact, we partly agree with you. One reason is a similar distribution of operational and non-operational conditions. In our case, the occurrence probability of operating conditions differs by up to 5.6 percent points between the three years. Hence, there is a moderate difference. A second reason is the difference between the fatigue damage behaviour during operational and non-operational conditions, i.e., the mean short-term damage for different turbine statuses. Again, for the considered turbine, these differences are moderate (approximately 25%). Due to these two reasons, turbine statuses do not have to be considered within the extrapolation. In a real application, the second point can easily be checked during the measurement period. The first point is more complicated. Still, in most cases, the duration of non-operational conditions will be known – at least approximately. Hence, in a real application, it will be possible to make an educated guess whether OCs have to be taken into account or not. A brief discussion about this topic was added to the paper (l. 445-451).

24.) Fig. 18: Should (1) not be the same as Fig. 15 for ANN_1D? This looks different to me.

You are correct, that they look different. This has something to do with the random initial weights used for ANN. This topic is discussed in Section 4.2. There, it is stated that Fig. 15, 18 and 19 show non-deterministic realisations for ANN, i.e., they have to look different. Since it seems as if this fact has been unclear so far, we added a short additional explanation to Section 4.2 (l. 463/464).

25.) Fig. 19: I am not an expert on machine learning, but would be very careful to state that binning is better than machine learning approaches since it strongly depends on how you set up the models. I think you made that acceptable clear, though, in the following paragraph.

We totally agree that “binning outperforms machine learning” cannot be a general statement. It will always be possible to find a machine learning approach outperforming the binning approach. The question is only how complicated this will be and which data will be needed. As you already pointed out, we tried to clarify this in the paragraph following Fig. 19 (l. 461-474 of the revised paper). Still, it might be sensible to add a statement regarding the performance of ANN and GPR before showing the figure. Such a statement was added to the beginning of Section 4.2 (l. 457/458).

26.) L. 484: I do not see convergence in Fig. 22. The error increases again after nine months?

Thank you for this comment. Yes, there is no actual/complete convergence visible for the simple approach. We assume that the slight increase after nine months is due to the small number of “different” measurement periods (13) making a statistical analysis complicated/unreliable. This is especially the case if a single month with severe wind conditions can change results significantly. For the simple approach, this is the case.

We reformulated the statement and added a short explanation regarding the limited data in Section 4.4 (l. 537-540 & 543).

27.) L. 520/581: I disagree. Long-term extrapolation (but also short-term extrapolations) is only possible with the provided methods if the loading situation of the turbine are represented in the measurement dataset and do not change in the prediction time. This is very important to be aware of.

See answer to comment 3

28.) L. 575: I disagree with this conclusion due to several reasons. 1) You have only investigated one point at the tower, quite high up. This may change for other points of the structure. 2) Other support structure types, other circumferences (e.g. deeper water) may also change this. 3) Wind speed from met masts is typically not available.

We agree that the conclusion is not correct in that way it was stated so far.
Regarding your first point, we changed the statement to clarify that this conclusion is only valid for points at the tower (l. 639). For other points, e.g., at the monopile, wave loads might become more relevant (see answer to comment 22 as well).
The limitation to similar turbines/support structures is discussed in Section 5.
Regarding your third point, see answer to comment 1.

29.) L. 590: One main limitation of the work is that you did not use 10min SCADA operational variables like power output, pitch, … Please explain why.

See answer to comment 2

30.) L. 606: Before this I believe it makes sense to look at 10-min SCADA operational variables like power output, pitch. Please explain why binning only applies to aggregated EC (wind speed) but not to aggregated OC?

See answer to comment 2

[1] Hübler, C., Gebhardt, C. G., & Rolfes, R. (2017). Hierarchical four-step global sensitivity analysis of offshore wind turbines based on aeroelastic time domain simulations. *Renewable Energy*, 111, 878-891.

[2] Nieslony, A. (2009): Determination of fragments of multiaxial service loading strongly influencing the fatigue of machine components. *Mechanical Systems and Signal Processing*, 23, 2712-2721.

Reviewer 2: (marked green in the manuscript)

1.) *The WT industry claims for a fatigue lifetime of substructures of only 20 years which is ridiculous from technical and environmental viewpoints. Private companies are obviously not interested in durable infrastructure; replacing WT every 20 years is very profitable for companies but has a very negative impact on the environment. The authors are invited to comment on the issue of useful service life to be expected from (environmentally friendly) WT.*

We completely agree that designing wind turbines for lifetimes of 20 or 25 years is not environmentally friendly or politically intended. Unfortunately, this is the standard today. Nonetheless, possible lifetime extensions become more relevant, so that we might come up with service lifetimes of 30-35 years. We added, a paragraph briefly discussing this topic to the introduction (l. 23-29).

2.) *Measured strain (stress) values in WT towers show relatively small values but the number of stress cycles is very high. The fatigue issue related to WT towers is in the domain of Very High Cycle Fatigue. This should be discussed. F.ex., fracture mechanics based fatigue theories indicate a threshold of fatigue stress intensity at (welded) details for which stresses, beyond this threshold, do not contribute to fatigue damage. This item should be discussed in more detail and considered in the fatigue damage accumulation.*

It is correct that most cycles wind turbines tower see are very small. It is also right that these small cycles might not influence the actual fatigue behaviour at all, e.g., if a horizontal S-N curve is assumed for small stresses.

In this work, we stick to the standards and recommendations used in the wind energy sector, e.g., [1]. These standards and recommendations recommend the usage of S-N curves without a horizontal part. Nonetheless, we added a paragraph to the paper (l. 213-216), which discusses this issue.

3.) *Related to the previous comment, typical (welded) details in WT towers and their fatigue resistance should be discussed in more detail, because of the obvious link between linear fatigue damage accumulation and the assumed S-N curve of fatigue vulnerable details.*

Surely, the (welded) details have a significant influence on the fatigue resistance. In the design process according to DNVGL [1], the details are taken into account by choosing different values for the stress concentration factors (SCF). Here, a SCF of 1.0 is assumed. However, this SCF highly depends on the precise detail investigated. Therefore, a short discussion of the uncertainty in determining adequate SCFs is added to the paper (l. 201-208).

4.) A safety factor is introduced but it is not discussed on how a safety margin should be fixed for the present case of monitoring-based fatigue assessment approach.

Thank you for this comment. We agree that the safety factor would definitely change if monitoring-based fatigue assessment is used. We added such a comment to the paper (l. 205/206).

However, for the validation of the extrapolation approaches (in time), the precise value of the safety factor is not essential, as it is only a (more or less linear) factor in the end. Surely, it should be in the correct range. Otherwise, unrealistic effects might occur, e.g., all cycles are shifted to the wrong part of the bilinear S-N curve.

[1] DNVGL: Fatigue design of offshore steel structures, recommended practice (DNVGL-RP-0005:2014-06) 2014