Response to reviewers / Manuscript ID: wes-2026-65 'Interannual variability in fatigue damage estimation from short-term strain monitoring of offshore wind turbines'.

The authors thank the reviewers for their insightful comments and suggestions, which will contribute to further strengthening and enhancing the quality of the proposed research. Based on the feedback provided, the manuscript has been carefully revised, and the modifications are marked in the tracked version.

Reviewer 1

In this work, binning-based approaches to extrapolate strain measurement-based short-term fatigue damage estimations for offshore wind turbines to longer periods of interest (for which only EOC data is available) are evaluated. For this purpose, for the first time, long-term measurements (here: eight years) are used. Furthermore, in contrast to the state of the art, a clear differentiation between year-to-year uncertainty and uncertainty due to limited sampling is done. As many offshore wind turbines start to reach the end of their design lifetime, investigating the performance and uncertainty of fatigue lifetime estimations based on strain measurements is very valuable. The methods used in this work are not new. However, due to the much longer measurement period, new insights were gained, which are of great interest to the readers of WES and probably also to industry. The paper is well written and nicely structured. There are only a few aspects that could be clarified or where a more comprehensive discussion would increase the added value of this work.

Comment 1: In line 43/44 you discuss that neglected EOCs or non-stationary distributions of covered EOCs are the reasons for the year-to-year variability. Perhaps, you can add that structural changes, modified control algorithms, etc. might also lead to changes over several years.

Response: *The authors thank the reviewer for this comment and have modified the manuscript accordingly.*

Modifications in the manuscript:

Page 2 (Lines 44-45)
"However, in practice, only a subset of EOCs can usually be considered, as some may not be available, measured, or identified as relevant, whereas exhaustively conditioning fatigue damage on numerous EOCs can result in sparse or empty bins. Consequently, if relevant EOCs are neglected and their distributions or effects vary over time, or if residual variability in damage is non-stationary due to factors such as structural changes or modified control systems, the resulting models may not be representative of long-term fatigue damage over the estimation period."

Comment 2: Several times you discuss that previous work does not differentiate between year-to-year uncertainty and limited sampling uncertainty. It is correct that so far this differentiation has not been made explicitly. Especially, results like the one in Fig. 14 in your work have not been presented so far and are very valuable. However, it is not fully correct that previous work has not addressed this topic. For example, Hübler & Rolfes (2022) also applied bootstrapping and a sliding window approach to identify separately both types of uncertainty.

Response: *We agree that previous studies, including Hübler and Rolfes (2022), have addressed year-to-year variability using sliding-window analyses. In the manuscript, we cite this study when discussing the results of the sliding window analysis in Section 5.1.1. In addition, we have revised Section 2 to acknowledge the contributions mentioned by the reviewer.*

Modifications in the manuscript:

Page 5 (Line 129-130), Page 6 (Lines 166-168)
"However, most of these approaches implicitly assume that the selected EOCs capture the dominant drivers of fatigue damage and that the conditional damage–EOC relationship remains stable over time."

"Only a limited number of studies explicitly model the stochastic behavior beyond the measurement period by propagating uncertainty from short-term data to longer estimation horizons. For example, Hübler and Rolfes (2022) applies a sliding window analysis to quantify the effect of yearly variability, and Mai et al. (2019) addressed predictive fatigue damage estimation"

Comment 3: Section 2 gives a nice overview of the state of the art and research gaps. However, there is a lot of overlap with Section 1. You might think about writing the introduction more concisely.

Response: *Following this comment, we have synthesized Section 1, making it more general, and moved detailed discussions to Section 2.*

Modifications in the manuscript:

Page 2 (Lines 33-41), Pages 2-3 (Lines 55-61), Page 4 (Lines 122-124)
"A common strategy for long-term fatigue damage estimation is to extrapolate or map short-term fatigue damage from strain measurements to longer periods, often leveraging information about EOCs (Hübler et al., 2018; Mujtaba et al., 2026). ~~In its simplest form, this mapping assumes that the measured damage scales linearly in time, an assumption that can logically lead to significant estimation inaccuracies when short-term strain data are not representative of the longer estimation period. More~~ Available methods range from simple linear mapping to more advanced data-driven approaches conditioning short-term damage on selected EOC parameters using binning or other conditional models (Hübler et al., 2018; Mujtaba et al., 2026; Hlaing et al., 2024). ~~In such approaches, damage calculated from 10-minute strain measurements is related to the corresponding EOC conditions, and the estimated damage over the available EOC period is obtained by weighting these conditional damages according to the EOC distribution during that period.~~"

"Regarding the duration of the effective short-term strain monitoring period for accurate long-term damage estimation, the literature presents conflicting views~~(Hübler et al., 2018; Marhadi and Skrimpas, 2015; Marsh, 2016; Hübler et al. (2018) and Hübler and Rolfes (2022) suggest that 9-10 months of strain data may already yield representative mean values with long-term damage estimation errors below 10%, whereas Marhadi and Skrimpas (2015) and Marsh (2016) argue that seasonal and interannual variability requires periods of at least 18 months , or even multiple years , to capture rare but influential events, such as storms~~, suggesting from 9 months to multiple years (Hübler et al., 2018; Marhadi and Skrimpas, 2015; Marsh, 2016; Loraux and Brühwiler, 2016; Mai et al., 2019) ."

"This linear mapping assumes that the measured damage scales linearly in time, an assumption that can logically lead to significant estimation inaccuracies when short-term strain data are not representative of the longer estimation period."

Comment 4: Typo in Fig. 2: It should be "Compute bins' estimators".

Response: *Based on this comment, we have adjusted Figure 2.*

Modifications in the manuscript:

Page 8 (Figure 2)

Comment 5: Eq. 6: You only apply the bin filling if there is no measurement in the bin. This is state of the art. Nonetheless, bins with only one or two measurements do not yield very robust estimates. Therefore, just out of interest: Have you tested to increase the minimum number of measurements in a bin?

Response: *Thank you for this remark. In the present study, bin filling is only applied to empty bins. We agree that bins containing only a very limited number of measurements may lead to less accurate estimates. However, the effect of imposing a higher minimum sample threshold per bin was not investigated within the scope of this work. This is an aspect that could be explored in future studies.*

Comment 6: Section 3.3.1: In this section, you introduce θ as a "general" damage estimator. However, in the entire work, you only use the mean value. Hence, it would be much easier for the reader to understand if you just stick to μ and remove θ from the paper. I know that θ was originally used in Sadeghi et al. (2024), but for the current work, I do not see the benefit of introducing it in Eq. 12 and then "removing" it again in Eq. 14.

Response: *Thank you for the suggestion. We agree that, since the present study only evaluates the mean fatigue damage, the notation could be simplified by using μ throughout. However, we have chosen to retain the more general estimator notation θ because the proposed framework is not limited to the mean and can be readily extended to other damage estimators in future work. Maintaining this notation also ensures consistency with the formulation introduced in Sadeghi et al. (2024) and facilitates future extensions without requiring modifications to the underlying methodology. To improve clarity, we have added this information to the paper.*

Modifications in the manuscript:

Page 10 (Lines 268-270)
"Re-estimation: For each replicate, recompute the conditional bin statistics (e.g., bin-wise mean damage) using Eq. (5), and evaluate a damage estimator $\widehat{\theta}(\cdot)$ based on the resampled data. Although the framework is formulated for a general damage estimator $\widehat{\theta}(\cdot)$, only the mean fatigue damage ($\theta = \mu$) is considered in this study. "

Comment 7: Fig. 3: If you have time steps of 90 days, the second year is not actually M(5) but starts about 5 days earlier (360 days vs. 365 days). How exactly do you handle the remaining 5 days of the year?

Response: *Thank you for pointing this out. The implementation uses calendar-based three-month windows derived directly from the time series timestamps, rather than fixed 90-day windows. Therefore, the yearly segmentation correctly accounts for the actual calendar duration and does not introduce a cumulative 5-day shift between years. To avoid this misunderstanding, the manuscript text has been revised from "90-day steps" to "3-month steps".*

Modifications in the manuscript:

Page 11 (Line 291), Page 12 (Figure 3), Page 16 (Line 370 and Line 380)
"using a circular sliding scheme with step Δ (e.g., ~~90 days~~3 months), to use the measurement"

"are constructed using a sliding scheme with time step ~~$\Delta =$~~$\Delta = 3$ months, resulting in overlapping segments"

"using a circular sliding-window scheme with a step size of ~~$\Delta = 90$ days~~ $\Delta = 3$ months ."

"we again construct $K = 32$ windows using the same circular sliding scheme with step ~~$\Delta = 90$ days~~$\Delta = 3$ months, to preserve a constant number of windows across durations."

Comment 8: Fig. 4: Are the precise positions of the strain gauges confidential or can they be provided? I assume that the strain gauges are evenly distributed every 60°?

Response: *The sensors are installed at 60° intervals, with S5 located at 265°. This is shown in Figure 4.*

Comment 9: Fig. 6: If I am not mistaken, the "Geometrical information" is already required in step "Convert the sensors' 10-minute stress ...".

Response: *We appreciate this remark and have revised the flowchart accordingly.*

Modifications in the manuscript:

Page 15 (Figure 6)

Comment 10: L. 344/345: You state a hierarchy for the conditional binning models. Have you checked the relevance of these EOCs or why do you use exactly these EOCs and this order. If I remember correctly, your group has analysed this or something similar before. Perhaps, you can refer to this work. Otherwise, the selection/order remains quite arbitrary. Similar applies to the bin sizes. For the wind speed, you analyse the effect of the bin size. However, how did you choose the bin size for the other EOCs? And why did you only investigate it for the wind speed?

Response:
Thank you for this comment. The selection and ordering of the conditioning variables are based on engineering judgment and prior knowledge of the main drivers of fatigue loading in offshore wind turbines rather than on a formal feature-selection procedure within this study. Wind speed, operational state, wind direction, and turbulence intensity are among the key environmental and operational conditions commonly considered in both fatigue assessment and turbine design practices.
The hierarchical ordering was chosen such that variables expected to explain the largest share of fatigue variability (operational state and wind speed) appear at the coarser conditioning levels, while wind direction and turbulence intensity are introduced at finer levels. The objective of this work was not to identify an optimal set of EOCs, but rather to investigate how increasing conditioning dimensionality affects fatigue damage estimation uncertainty. We agree that the choice of variables and their ordering can influence the results and have clarified this point in the manuscript. We have also added references to previous studies from our group that investigated the relationship between fatigue damage and environmental and operational conditions.
Regarding the bin widths, the selected values (2 m/s for wind speed, 30° for wind direction, and 0.1 for turbulence intensity) were chosen as representative values commonly used in wind energy analyses and to maintain sufficient sample populations within bins. A dedicated sensitivity analysis was only performed for wind speed because wind speed is the primary conditioning variable across most investigated models and is generally the dominant driver of fatigue loading for smaller offshore wind turbines. The goal was to demonstrate the effect of smaller bin sizes, and a general trend can be expected of any other parameter chosen as the binning parameter.

Modifications in the manuscript:

Pages 16-17 (Lines 360-363), Page 22 (Lines 507-508)
" These conditioning variables are selected based on their established influence on fatigue loading in small offshore wind turbines. A more detailed investigation of feature selection and variable importance is presented in Mujtaba et al. (2026) . It should be noted that the objective of this study is not to identify an optimal set of binning variables, but rather to evaluate how increasing the conditioning dimensionality affects interannual variability. "

"The purpose of this study is to demonstrate the effect of smaller bin sizes, and a general trend can be expected of any other parameter chosen as the binning parameter. "

Comment 11: Fig. 8: Perhaps you can briefly discuss the outliers for 0D and 1D-Case for year 4, 5, and 6. I assume in year 5 and 6, the missing data during winter is responsible. Similarly, in year 4, a lot of data is missing in summer. Hence, the higher damage during winter is overweighted.

Response: *A brief discussion is added to the manuscript.*

Modifications in the manuscript:

Pages 19-20 (Lines 459-466)
"Bin imputation is examined in Appendix A. Although it affects the 3D and 4D models, it is not the source of the observed interannual variability. The largest outliers that occur for 0D and 1D-Case are strongly attenuated once wind speed is included as a conditioning variable, indicating a mismatch between the wind-speed distribution within those measurement windows and that over the period of interest. Such a mismatch may arise from genuine year-to-year differences in the wind climate or from non-uniform strain data availability within the window. "

Comment 12: Table 1: It seems as if high-dimensional binning approaches start to become slightly biased estimators, i.e., $\bar{e}_{norm}$ is no longer "close" to zero. I assume that is due to empty bin filling. Perhaps, you can discuss this briefly.

Response: *A brief discussion is added to the manuscript.*

Modifications in the manuscript:

Page 21 (Lines 480-482)
"In some cases, higher-dimensional approaches also exhibit a slight bias, reflected by non-zero values of $\bar{e}_{norm}$. This behavior is likely related to increased data sparsity and the associated bin filling required in high-dimensional binning models. "

Comment 13: Fig. 9: It is very interesting to see that measurement periods longer than one year improve the results, as previous research sometimes revealed no further reduction for periods above 9 months. A reason could be that previous work mainly looked at measurement periods up to 12 months, where an increase from 9 to 12 months did not have an influence. Hence, previous work never considered two winter periods. Probably, you have not analysed this, but (perhaps for future research) it would be interesting to check whether the improvement is actually due to long measurement periods or more data during winter.

Response: *Thank you for this observation. We agree that seasonal coverage, and winter coverage in particular, is a plausible contributor, and we have retained the suggested experiment as future work. The new Appendix B indicates that a second mechanism is also at work in this dataset. All conditional models underestimate the ground truth in years 1 to 3 and overestimate it in years 5 to 8 (Figure B1), consistent with a step change in the conditional damage-EOC response near year 4. Multi-year windows that span this transition average across both regimes, which reduces the dispersion of the error independently of how many winters they contain. Conversely, a window confined to one side of the step remains unrepresentative of the period of interest regardless of its length, which is why longer monitoring reduces but does not eliminate the error in Figure 9. Disentangling the two effects thus requires separating seasonal coverage from coverage of the regime change. We have added this distinction to the future work discussion, alongside the fixed-window experiment with progressively added winter periods that the reviewer suggests.*

Modifications in the manuscript:

Page 29 (Lines 615-618)
"Investigating this coverage effect merits systematic study, since the reduced variability observed for longer windows may (i) reflect the duration itself, (ii) the recurrence of winter conditions, or (iii) the inclusion of data from both sides of the step change identified in Appendix B. A fixed one-year window progressively augmented with additional winter periods would separate the first two mechanisms, while restricting windows to a single regime would isolate the third. "

Comment 14: Fig. 10: Is there a reason, why you use the mean value here? For the other box plots, you used the median. Changing this within the paper might be confusing for the reader.

Response: *Thank you for your attention. The caption was inaccurate, and it is now modified to align with the rest of the paper.*

Modifications in the manuscript:

Page 23 (Figure 9), Page 24 (Figure 10)
"box plots summarize the error distribution for window lengths of 1, 2, and 3 years based on 32 windows. The boxes represent the interquartile range (IQR, defined as the 25th–75th percentiles), the grey line denotes the median error, and the whiskers extend to values within 1.5×IQR."

"corresponding yearly values. ~~comprises 8 points from the 8 consecutive individual years of the S5 damages shown as the markers.~~ The boxes represent the interquartile range (IQR, defined as the 25th–75th percentiles), the grey line denotes the median error, and the whiskers extend to ~~the most extreme points~~

values within 1.5×IQR.~~Grey lines denote the mean error.~~"

Reviewer 2

The manuscript presents a comprehensive analysis of the effect of training data selection for long-term fatigue damage estimation from short-term strain monitoring based on a binning approach. The investigated parameters are mainly the effects of different training data lengths and binning approach settings.

The paper is well-written, good to understand, and is interesting for the readers of WES. However, some aspects could improve the study:

Comment 1: The introduction and Section 2 are well-written but contain some repetitions (e.g. 55-60 and 140-145).

Response: *The introduction has been shortened and made more general to improve readability and avoid repeating the detailed discussions presented in Section 2.*

Modifications in the manuscript:

Please refer to response to Reviewer 1, Comment 3.

Comment 2: It is unclear whether the conditional binning model is novel or derived from prior work. If it is indeed new, the paper should address the effects of dimensionality filling and the order of binning. Otherwise, please cite the corresponding study.

Response: *Thank you for this comment. The conditional binning framework is not introduced as a novel methodological contribution in our study and has been adopted from previous fatigue extrapolation studies, which are now cited in the manuscript. The novelty of this work lies in evaluating the effects of limited monitoring data and interannual variability on fatigue damage estimation using an eight-year dataset.*

The revised manuscript addresses dimensionality by comparing conditional models from 1D to 4D and shows that increasing dimensionality does not necessarily improve performance because it increases sparsity and reliance on bin filling. The effect of bin filling is quantified by comparing the results before and after hierarchical imputation. Because this imputation uses the next lower-dimensional parent model, the order of the conditioning variables may affect the results. The adopted order was selected based on physical relevance, while a systematic comparison of alternative binning orders was not included in the present analysis and could be examined in future work.

Modifications in the manuscript:

Please refer to the response to Comment 10 posed by Reviewer 1 regarding the selection of binning variables.
“Table 2 , and Appendices A and B discuss the effects of data imputation.”

Comment 3: The explanation of the interannual variability of the DEL estimation mentions a non-stationary system or neglected EOCs. It would be useful to determine whether a statistical test can confirm non-stationarity. What could be the causes (such as structural aging or scour, etc.)? Which neglected EOCs could be the reason? Discussing potential future research directions for handle these effects would enhance the study.

Response: *Thank you for this valuable comment. Based on this feedback, we have revised the manuscript and included an explicit comparison of interannual variability with respect to finite-sample statistical uncertainty as well as a new appendix, where we provide insights on the residual variability structure.*

Regarding statistical tests, Table 2 now reports, for each binning strategy and damage direction, the lower bound of the 95% confidence interval on the interannual standard deviation of the normalized error (chi-squared, seven degrees of freedom, from the eight non-overlapping annual windows). For every strategy and direction this lower bound exceeds the corresponding within-window bootstrap standard deviation (Table B1), which stays near or below 1%. For 1D-WS in the S5 direction, for instance, the interannual standard deviation is 5.0 with a lower bound of 3.3, against a bootstrap value of 0.4. The year-to-year dispersion is thus not attributable to finite-sample statistical uncertainty, which is the alternative hypothesis relevant to what we examine. We refrain from a formal stationarity test on the damage series itself, since *eight annual blocks provide limited power and the quantity of interest is the conditional damage-EOC mapping rather than the marginal damage distribution.*

Appendix B shows that the year-wise errors are not random scatter. All conditional binning models underestimate the ground truth in years 1 to 3 and overestimate it in years 5 to 8 (Figure B1), consistent with a step change in the conditional response near year 4. This also explains why longer monitoring windows do not remove the error. Resolving the step across the EOC space provides insights on the possible causes (Figure B2). The relative change in conditional bin-mean damage between the late and early periods concentrates in the 270 to 330 degree and 60 to 120 degree wind-direction sectors and grows with wind speed, while the prevalent south-westerly sectors remain largely unchanged. Any mechanism consistent with this observation must act non-uniformly across direction sectors. This narrows the candidate explanations to direction-dependent quantities, whether environmental, operational, or structural. We thus report the directional structure as an empirical observation and do not attribute it to a specific cause. Identifying that would require information beyond the present dataset. Appendix A further confirms that empty-bin imputation governs the accuracy of the mean estimate but leaves the interannual spread essentially unchanged, so the observed variability is not an artifact of the bin-filling rule.

On future research, the conclusions now indicate potential directions: tracking of natural frequencies and damping to test for structural mechanisms such as scour or soil stiffness degradation; inclusion of wave parameters and wake-related descriptors as conditioning variables, together with cross-turbine and farm-wide validation to separate site-specific from turbine-specific effects; and Bayesian updating frameworks that assimilate monitoring data sequentially while accounting for model discrepancy.

Modifications in the manuscript:

Table 2 and Appendices A and B

Comment 4: Consider merging Section 4.3 and Section 5 to improve coherence, possibly including a table summarizing the studies.

Response: *We consider that keeping Sections 4.3 and 5 separate improves the structure of the manuscript, as Section 4.3 describes the experimental protocol and methodology adopted in this study, whereas Section 5 presents and discusses the corresponding results. However, to improve readability, we have added a summary table in Section 4.3 that provides an overview of the analyses conducted in the paper.*

Modifications in the manuscript:

Page 16 (Table 1)

Comment 5: Section 5.1 should include a table illustrating bin filling rates per model, indicating how the different dimensional models vary in terms of empty bins.

Response: *Thank you for this suggestion. We have revised Table 2 and added Appendices A and B to quantify bin occupancy across the different model dimensionalities and to examine the effect of empty-bin imputation on the estimated errors and their interannual variability.*

Modifications in the manuscript:

Page 22 (Table 2), Appendix A, and Appendix B

Comment 6: Clarify the rationale behind the choice of specific models (e.g., sensor heading, binning dimension) for respective studies.

Response: *This information is included in Table 1.*

Modifications in the manuscript:

Page 16 (Table 1)

Comment 7: For Figure 14, it's unclear why only 0D and 1D-WS models are shown. They have been shown to not the best choice to capture the interannual variability. It would be beneficial to include results for all applied model dimensions. Why are the results of only one and two years window length shown and not the three years window?

Response: *Thank you for pointing this out. Figure 14 illustrates how the bootstrapping results should be interpreted, rather than emphasizing the predictive performance of the models. Additionally, Table B1 shows how the width of the bootstrapping confidence intervals changes with increasing measurement sample size. To provide a more comprehensive overview and ensure consistency across all models and measurement periods, Figure 14 and Table B1 have been updated accordingly.*

Modifications in the manuscript:

Pages 26-27 (Lines 545-564),Page 27 (Figure 14), Page 33 (Lines 711-716), Page 35 (Table B1)
"~~Table 2~~ Figure 14 reports normalized errors for all the binning models based on eight non-overlapping calendar-year windows. For each window, 95% bootstrap confidence intervals are computed using $N_{\mathrm{boot}} = 1000$ replicates under the BootWhole scheme (Sadeghi et al., 2024) .

Additionally, Table B1 in Appendix B reports the standard deviation of the normalized error in the estimated mean damage from bootstrap resampling within each ~~measurement~~ sliding window, together with the corresponding standard deviation computed across ~~multiple~~ 32 windows without resampling. The ~~latter is based on eight consecutive individual years and eight 2-year~~ results are based on 32 one-year sliding windows with a ~~1-year step . Results~~ step of 3 months. Values are reported for the S5 damage direction and expressed as percentages for ~~both the 0D and 1D-WS~~ all binning models. The results show that the standard deviations obtained from bootstrap resampling are negligible relative to the overall variability observed across measurement windows. This indicates that, when long-term mean damage is estimated from limited-duration measurements, interannual variability constitutes the dominant source of uncertainty, whereas sampling variability due to finite data within a given window plays a comparatively minor role. ~~Comparison of normalized error standard deviations from bootstrapped samples for each individual year, together with the standard deviation of the normalized mean errors across the eight consecutive windows for window lengths of 1 and 2 years. Results are reported for the 0D and 1D-WS binning strategies in the S5 direction. Columns Btrp. (w1) to Btrp. (w8) denote the standard deviation of bootstrapped normalized errors for each yearly window, while std(e_{norm}) reports the standard deviation of the eight normalized mean errors across windows. Values are expressed in percentage. Strategy Window lengthBtrp. (w1)Btrp. (w2)Btrp. (w3)Btrp. (w4)Btrp. (w5)Btrp. (w6)Btrp. (w7)Btrp. (w8)std(e_{norm}) (Table 2)0D 1 year0.60.50.70.80.70.60.7 0.613 1D-WS 1 year0.40.40.40.50.50.50.4 0.450D 2 years 0.40.40.50.50.40.40.4 0.48 1D-WS 2 years0.30.30.30.30.40.30.3 0.34~~

~~Figure 14 reports normalized errors for the 0D and 1D-WS models based on (left) eight non-overlapping calendar-year windows and (right) eight two-year windows. For each window, 95% bootstrap confidence intervals are computed using $N_{\mathrm{boot}} = 1000$ replicates under the BootWhole scheme (Sadeghi et al., 2024) .~~"

"Separating the two sources of error confirms that interannual variability dominates over finite-sample noise. Table B1 shows that the within-window sampling variability, obtained by bootstrapping, stays near or below 1% for all conditional models and window lengths, whereas the standard deviation of the error across windows is several times larger (e.g., 5% versus 0.4% for 1D-WS at one year). The 95% confidence interval on this cross-window standard deviation lies well above the bootstrap values, confirming that the interannual variability is statistically distinguishable from finite-sample noise and is the dominant source of uncertainty."