Response to Reviewer Comment RC2

We sincerely thank Reviewer #2 for their constructive and insightful comments on our manuscript *CMIP6 Multi-model Assessment of Northeast Atlantic and German Bight Storm Activity*. The comments greatly helped us to improve the manuscript and clarify key points.
In the following, we will give a point-by-point response to the reviewer's comments and describe how we plan to address the issues raised.

General comments:

A A number of analyses in this study lack any assessment of robustness, e.g. by checking for statistical significance or provision of confidence intervals. In particular this is the case for results presented in Figs. 4, 6, 7, 8, 9, 10, and 11. I would be willing to accept this in case of Fig. 11 in order to ensure readability but for the other Figs. an analysis of statistical significance (nothing mentioned in the text either regarding these results) as well as including any indication in this respect in the Figs. seems necessary and possible. The authors themselves discuss the strong sensitivity of such studies depending on choice of metric, integration period, storm identification method, … (lines 259-262). Especially in these cases, a thorough assessment of robustness and statistical significance is unavoidable for a proper scientific study.

Response: We agree with the reviewer that the mentioned figures require an assessment of significance or confidence. We will reproduce the figures from the analysis with added indicators of statistical significance. We will also expand the results and discussion section with thorough analysis of the significance test and implications for our conclusions.

B The authors introduce very briefly, why the analysed parameters matter. Apart from lines 14-19 which sketch a few possible impacts of Northeast Atlantic storms, there is nothing. In particular, they provide no reasoning why wind direction is interesting besides wind speed/storm activity alone. A number of reasons are very clear to me but readers not familiar with the specific of potential storm impacts will have no clue why wind direction is relevant beyond academic exercise.

Response: We thank the reviewer for this observation. We will revise the introduction to include a paragraph that motivates the relevance of wind direction analysis. In particular, we will emphasize its importance for understanding the pathways of storm systems, storm surge impacts on coastal areas, directional wind stress on infrastructure, and compound flooding risks due to onshore winds. Furthermore, wave height and direction are dependent on the wind direction (via the wind fetch), and play a big role in determining coastal impacts such as erosion. This will help clarify the added value of this analysis for readers unfamiliar with storm impact mechanisms.

C I am a little worried regarding the compilation of the multi-model ensemble. The authors themselves stress a problem related to their procedure of selecting one ensemble member from each model which leads to thos models with just one member always remaining part of the multi-model ensemble. This is a major shortcoming of this study. I am grateful that the authors have the courage to outline this shortcoming themselves. This is a great example of good scientific practice and they present an approach to asses the potential influence by compiling a second multi-model ensemble, bootstrapping only from thos models with 5 or more members. Anyway, I am worried for one more reason: If I understand correctly, the authors perform the bootstrapping for those models with more than one member separately for the historical experiment and the individual scenarios. When doing so, a chosen scenario run is probably unrelated to its historical counterpart (at least in most cases). This yields inhomogeneities masking physical meaningful climate signals. In the end, it probably does not matter so much, given that the averaging over the multii-model ensemble is done, smoothing out this inconsistency between the end of the historical period and three separate beginnings of future scenarios. However, I would argue that choosing scenario simulations belonging to the chosen historical parent simulation had been a better solution. If my understanding is correct, I would ask the authors to include this issue in the discussion, too.

Response: We greatly appreciate this comment. The reviewer is correct in assuming that the bootstrapping is performed separately for the historical and the scenario runs. The main reason why we did not choose the same runs for historical and scenarios is that not all members continue from the historical to the

scenario period. In other words, the number of available runs changes between historical and the different scenarios for certain models, and some scenario runs do not have a clear parent in the historical period. A way to circumvent this is to only allow for the selection of those members where a scenario run is clearly connected to a historical parent run. We will investigate the feasibility of this approach and add a discussion of this issue in Section 4, including its potential implications for model consistency and the physical realism of trend transitions.

D lines 113-115: Choosing the nearest grid point from each model for a given observational site may lead to some distortion in areas of steep orographic gradients, namely Bodoe and Bergen. I understand that the authors use SLP, however, an extrapolation of modelled surface pressure into the ground of the one model (where the closest grid point is inland) may be somewhat different from SLP practically identical to modelled surface pressure of another model (where the closest grid point is flat terrain or even ocean). I do not insist a priori on including a discussion of this matter in the manuscript but I challenge the authors to think about this possible case (or perform some sensitivity test) and provide arguments here, why this should not be relevant for their findings.

Response: We thank the reviewer for this thoughtful comment. We will add a brief discussion in the methods section acknowledging that the selection of the nearest model grid point may introduce slight distortions in orographically complex regions such as Bodø and Bergen. We will mention this explicitly in the revised manuscript.

E I am pretty confused by the section title of Sec. 3.2. "Internal variability" refers to temporal fluctuations of various variables inherent to the natural earth system, observed or modelled. Single-model initial condition large ensembles (SMILE) are great tools to distinguish extrenally forced signals from internal variability. However, this is not what you are analysing in the large part of this section. Instead, the main focus of this section are heterogeneous climate change signals for different parts of the storm intensity spectrum as well as depending on wind direction. I would suggest choosing a more suitable title for Sec. 3.2 plus deleting a few sentences in the first paragraph of Sec. 3.2 (see my specific comment below).

Response: We appreciate this concern regarding the title of Section 3.2. In accordance to our response to the specific comment below, we will completely rework the first part of this section so that it does not mix up changes in the ensemble mean and internal variability anymore. We will also follow the reviewer's suggestion and rename the section to better reflect its focus on future changes in the extreme tails of the wind speed distribution.

Specific comments:

L26-28 Actually, the synthesis of Feser et al. is the other way around (increasing north of 60°N, decreasing south of 60°N) in line with a poleward shift of the NH storm track.

Response: We thank the reviewer for catching this error. Indeed, the synthesis of Feser et al. concludes that a majority of studies project decreasing storm activity south of 55-60°N and increasing storm activity north of that. We will correct this error in the revised version.

L38-39 I don't understand the line of argumentation here. The previous sentence is about low model agreement, hence large model-related uncertainty. This sentence here now seems to present a link to substantial changes in wind extremes when combined with changes in storm track locations. It seems something is missing here in between.

Response: We intended to argue that the large model-related uncertainty in both the location/shift of the storm tracks and the cyclone density leads to a very high uncertainty in the future evolution of (extreme) wind speed distributions at certain locations in the North Atlantic sector. We will rephrase this paragraph to clarify our line of argumentation.

L97 Did the authors analyse if discrepancies are introduced when using 3-hourly data for MPI-ESM compared to daily averages for the other models? An average of the eight 3-hourly values is usually in quite

good agreement with a respective daily mean. So, it might even haven been an alternative to calculate such a proxy daily mean for MPI-ESM before analysing its simulations consistently with the other models.

Response: We appreciate this concern. For the multi-model analysis, we used daily-mean pressure data only, also from the MPI-ESM model. The three-hourly data are used exclusively in the single-model large-ensemble analysis. We see that this is not explained clearly enough in the Methods section and will revise the respective paragraphs accordingly. We agree that comparing storm activity values calculated from three-hourly data with those calculated from daily values would introduce inconsistencies.

Sec. 2.2 It is not entirely clear from this section if MPI-GE is also used as part of the CMIP6 multi-model ensemble and hence included in respective analyses or not.

Response: We apologize for the unclarity in this regard. The MPI-GE is part of the CMIP6 multi-model ensemble. In the multi-model sections, we use daily-mean SLP data from all listed CMIP6 models, including the MPI-GE (MPI-ESM-LR). We see that assigning two names to the same model (MPI-GE and MPI-ESM-LR) might be confusing to the reader, and will improve the explanation and use of the terms MPI-GE and MPI-ESM-LR accordingly.

Sec. 2.3 This section lacks the information that an averaging over all ten triangles is performed to yield results for the Northeast Atlantic. This fact is explicitely written only in lines 271-272, that is the discussion.

Response: We agree with the reviewer that this important step is missing here. We will add a paragraph on the averaging that is performed over the ten triangles.

L115-117 Sory, I don't understand the procedure for "triangles" that basically fall onto a line. I am lost when you write about "the observation site that is most distant to the corresponding gridpoint". Which gridpoint? It is three observation sites with three associated grid points... Please rephrase (or correct?) your description here.

Response: We apologize for the lack of clarity in this section and thank the reviewer for pointing it out. The procedure aims to construct a triangular area for each model that mirrors the geometry of the three observational sites used to estimate storm activity. For each model, we identify the grid points that are geographically closest to the three observational stations. These three grid points define a triangle over which the geostrophic wind is calculated.

However, in some models, the three closest grid points can fall nearly on a straight line (e.g., sharing the same latitude or longitude), which would prevent a meaningful calculation of the wind vector due to an enclosed area of zero. In such cases, we slightly adjust the position of one grid point to form a proper triangle. Specifically, we move the grid point corresponding to the observation site that is geometrically furthest from the initially assigned grid point. This adjustment is limited to a single grid cell in the nearest orthogonal direction to preserve the original geometry as closely as possible while ensuring a valid triangle. We will revise the manuscript to explain this procedure more clearly.

L125&149: I would refrain from using the term "multidecadal oscillation" here. For me this term implies some type of natural variability associated with these fluctuations. But this is not possible given that you analyse the ensemble mean (and certain ensemble quantiles) here. These fluctuations must be either by chance (unlikely in these cases) or also result from external forcing common to all individual simulations. If you discuss these fluctuations, you have to provide possible drivers here.

Response: We agree that the term “multidecadal oscillation” is inappropriate here as the ensemble mean can only reflect variability that relates to an external forced climate signal. We will rewrite the respective sentences to avoid suggesting that the ensemble mean variability hints at an oscillation originating from within the Earth system.

Fig. 3 I would suggest swapping subfigures vertically in order to present the same order as in previous figures: first, the results from the bootstrapped MME, second, the results from all members.

Response: We will follow the reviewer's suggestion and rearrange the subfigures.

L178-179 "Internal variability" refers to fluctuations inherent to the natural earth system, observed or modelled. Single-model initial condition large ensembles (SMILE) are great tools to distinguish extrenally forced signals from internal variability. However, you cannot diagnose the internal variability from the SMILE's ensemble mean. It seems to me that you are doing that here. I would suggest eliminating everything from line 178 to line 196, replacing that with a better introduction for the direction- and intensity-dependent analyses, and then continue with line 197.

Response: We thank the reviewer for bringing this up. We agree that the ensemble mean of the SMILE is not a suitable tool to quantify internal variability. We see the need to disentangle this paragraph into a part on the general evolution of the SMILE ensemble mean across the scenarios (i.e., the forced signal), and the actual analysis on the internal variability (extremes, wind directions) which uses the pooled ensemble data and not just the ensemble mean. We will take up the reviewer's suggestion to completely rewrite the first part of this section.

L201 It looks to me as if the (positive) frequency changes are rotated clockwise (not counter-clockwise) compared to the historical frequencies.

Response: The reviewer is correct that the positive (red) frequency changes are rotated clockwise compared to the historical (gray) frequencies. However, in this sentence we refer to the differences in frequency changes between the MPI-GE (Figure 6b) and the CMIP6-MME (Figure 4b), which show a counterclockwise shift from CMIP6-MME to MPI-GE. We will rephrase this part of the results to clarify which wind roses we compare to avoid misunderstanding.

L206-208 ff. Why are the changes of upper percentiles of wind speeds only analysed from MPI-GE. This could have been done from the CMIP6-MME as well or not?

Response: The reason we performed the extreme wind percentile analysis only for MPI-GE is the availability of high temporal resolution data (3-hourly), combined with the large ensemble size within a single model, i.e. with consistent model physics. Most CMIP6 models only provide daily data, which is less sufficient for robust storm activity estimation at the high end of the distribution. We will clarify this constraint in the manuscript.

L210 Here you write about "significantly weaker". Have you checked statistical significance? If not, refrain from using this term here but as said in my general comment A) you will have to assess statistical significance.

Response: We agree that the term "significantly" should not be used without supporting statistical tests. In line with our response to general comment A, we will include significance tests for these results and update the language and figures appropriately.

L219-222 Why does only MPI-GE allow for the analysis of the extreme events? You could have done the same with the full CMIP6-MME, or not?

Response: Similar to the point above, we focus on MPI-GE for extreme event analysis due to its temporal resolution and ensemble size. However, we acknowledge that a subset of CMIP6 models could be used for similar analysis if suitable data are available. We will mention this limitation explicitly.

L304-305 Rephrase this sentence about the internal variability. This is not what you are looking at.

Response: We will remove the term "internal variability" here and rewrite the sentence to reflect that the ensemble spread is able to encompass the observed variability.

Outlook I would suggest at least one or two outlook sentences... What are possible next steps that could be done based on your findings?

Response: We thank the reviewer for this suggestion. We will add a short outlook section at the end of the discussion, highlighting potential next steps such as including a seasonal decomposition, applying percentile-based event attribution approaches, and investigating compound coastal hazard risks under future storm conditions.

Technical corrections:

L52 Insert "that is" before "most pronounced in the CMIP3...".

Response: We will reword this sentence as suggested by the reviewer.

L110 The reference to Krueger et al. (2019) makes no sense here. It is correctly placed in line 111, so please delete it here.

Response: We agree and will delete the reference in line 110.

L277 Replace "second part" by "final part". It's just this last element and not half of your study.

Response: We will replace the term as suggested.

We, the authors, would like to thank Reviewer #2 again for their careful reading of our manuscript and for the constructive comments. We hope that our responses and proposed revisions clarified all outstanding points and look forward to further feedback.

With kind regards,

Daniel Krieger and Ralf Weisse