Response to Reviewer #3

We sincerely thank Reviewer #3 for their constructive and insightful comments on our manuscript *Skillful Decadal Prediction of 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.

Comments

C1 One issue is that the paper focusses on some unusual forecast lead times (4-10 years and 7 years) without properly motivating why they use these. It would not seem the most interesting lead times for a user of a storm activity forecast. There is frequent reference to short and long averaging periods and I was not always sure whether that referred specifically to these two periods or had been generalised somehow. But if it is the latter, it is not defined. The language needs to be cleaned up around this.

> We thank the reviewer for making us aware of this ambiguity in the manuscript. Our intent was to show two exemplary lead year ranges, one for short averaging periods, and one for long averaging periods. The analysis could for example have also been done for lead years 1-7 and 4 (Fig. 1), leading to similar conclusions. We will clarify our intent in the methods and results sections and try to highlight that we only give two examples for reasons of brevity, but draw conclusions for more lead year ranges than just the two shown.

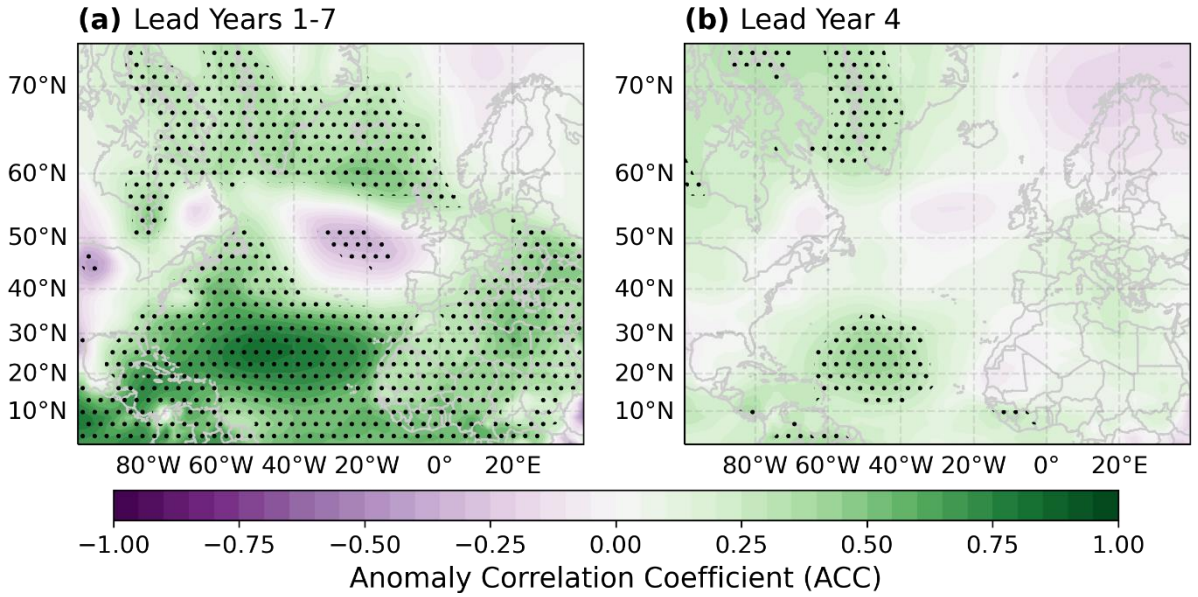

Fig. 1: Gridpoint-wise anomaly correlation coefficient (ACC) for DJF MSLP anomalies between the hindcast ensemble mean and ERA5 for lead years 1-7 **(a)** and lead year 4 **(b)**. Stippling indicates significant correlations (p<=0.05), determined through a 1000-fold bootstrapping with replacement.

C2 Figures which lead to firm conclusions are squirreled away in an appendix. I suggest all important figures need to be in the main paper. See minor comments below.

> We agree that the selection of figures in the main body of the manuscripts needs improvement. We will rearrange the figures, so that all figures which support the conclusions can be found in the results section, not in the appendix.

C3 Another issue is that there is no deterministic skill for mslp anywhere near the German Bight (Figure 1), so how do you explain that you have skill in predicting storm activity there? This needs to be covered in the discussion.

> We apologize for leaving this vague. One way to explain it could be that German Bight storm activity does not depend on the MSLP itself, but on the annual statistics (95^{th} percentiles) of the horizontal MSLP gradients, for which the model shows some skill. This might be due to the model being unable to predict fluctuations around the mean, but being able to predict sufficiently large deviations from the mean. We will add a paragraph on this contradiction to the discussion section.

C4 Negative skill is presented as useful skill. It is true, you could multiply the forecast by -1 and get a good forecast on average. The problem is that the skill is possibly negative due to a poorly modelled teleconnection and if there is an individual year when that teleconnection is not strong, multiplying by -1 could be the wrong thing to do. Better to assume negative skill is not useful even if it is significant.

> The reviewer is right; a significant negative correlation is just useful for predictions when the physical reason behind it is clear, which is not the case here. We will clarify that, while the correlation itself is significant, this is of little use for a skillful prediction.

C5 Finally, the text often refers to "tails" of the distribution and "extremes" when in fact the data refers to anomalies exceeding 1 sigma, which is neither in the tail or an extreme. These words need to be removed from the text.

> We agree that the terms “tail” and “extremes” can be misleading when talking about +/- 1-sigma events. We will replace these terms with better-suited vocabulary.

<u>Minor comments</u>

For reasons of brevity, we will refrain from repeating all comments in this document, and instead just refer to the line numbers in the original manuscript.

9 Here, “short lead years” should rather read “short averaging periods”. We apologize for this mistake and will rephrase this sentence.

126 We chose two exemplary lead year periods, one for long averaging periods, and one for short averaging periods. There is probably little interest in forecasts for lead years 4-10 specifically, but our intention was to use two cases to bring our point across. We definitely see the need to motivate this choice better and will enhance the respective paragraphs. Please also see our response to comment C1.

149 We calculated the means and standard deviations for each member separately to account for possible biases/shifts between individual members, and to force each member to be centered around a storm activity of 0. It would be equally valid to allow biases between members and use the full ensemble mean and standard deviation for the derivation of GBSA. Please see the figure below (Fig. 2), showing the model-observation ACC (deterministic skill) after using the individual means and standard deviations (left), as well as after using the full ensemble mean and standard deviation (right). The differences in ACC are negligible. This is also the case for the BSS of the probabilistic forecast.

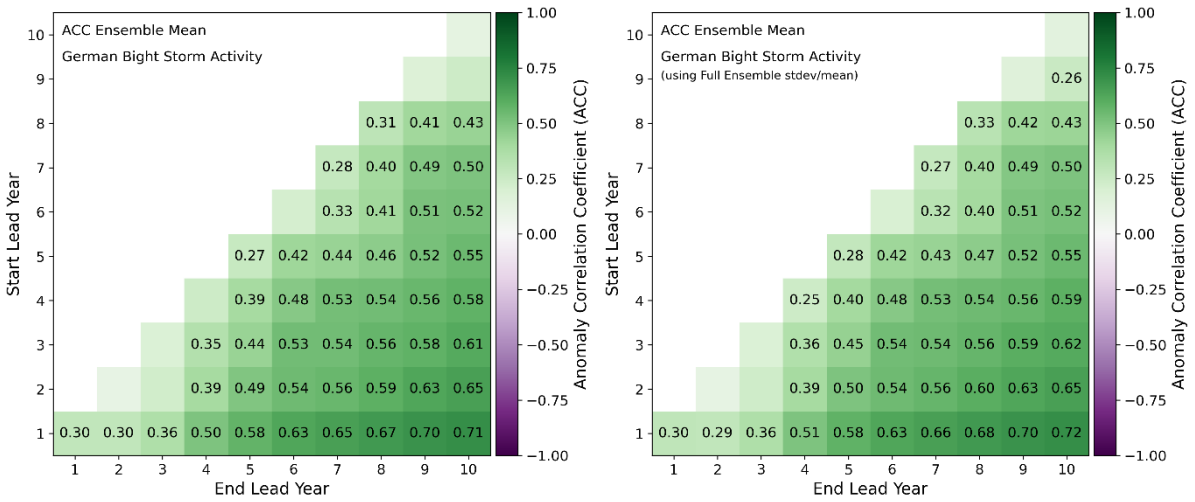

Fig. 2: ACC between observations and ensemble mean predictions of German Bight storm activity (GBSA) for all combinations of start and end lead years. GBSA predictions are based on individually standardizing members with their respective means and standard deviations **(left)**, and on standardizing with the mean/standard deviation of the full ensemble **(right)**.

162 The reviewer is correct. The Fisher-z method requires independent samples, which we did not account for. We will recalculate the significance with a bootstrapping approach, as especially the smoothed (multi-year average) time series are heavily autocorrelated.
We also agree that adding time series of GBSA would be helpful. However, we want to note that these time series would also likely be for exemplary lead times only, since including time series for all 55 possible lead time combinations would overload the manuscript. We will restructure the methods sections and then re-evaluate the possibility of including exemplary time series of derived GBSA, for example for lead years 4-10 and 7.

195 We will rephrase this sentence.

205/211 Please see our response to comment C4.

224 We will remove “anyhow” from this sentence.

241 We thank the reviewer for their thoughts on a possible initialization shock. In fact, the predicted geostrophic winds are lowest for lead years 3-5, and highest for lead year 1 (Fig. 3). While we use lead year 1 means and standard deviations to derive standardized GBSA from the absolute geostrophic winds for all data, we also tested whether standardizing each lead year with its respective mean and standard deviation has a notable effect on the results. We find that the ACC between model and observation is almost unaffected by the choice of our standardization reference (compare Fig. 4 and Fig. 2 left). Nevertheless, we will expand the discussion of our results with a paragraph on the effects of a possible initialization shock.

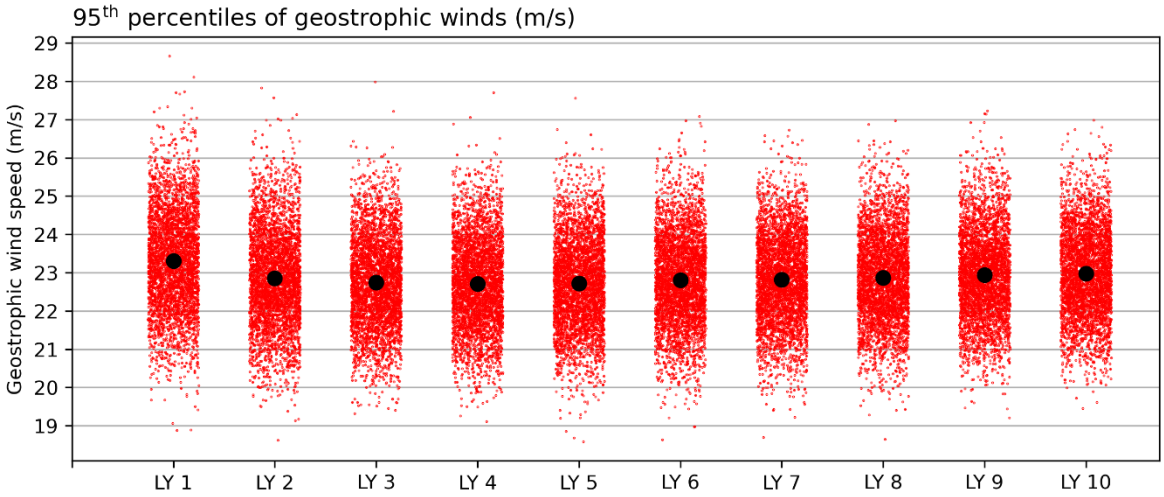

Fig. 3: Absolute 95th annual percentiles of predicted geostrophic winds per lead year. Red dots represent individual members and initialization years, black dots show the ensemble mean for each lead year.

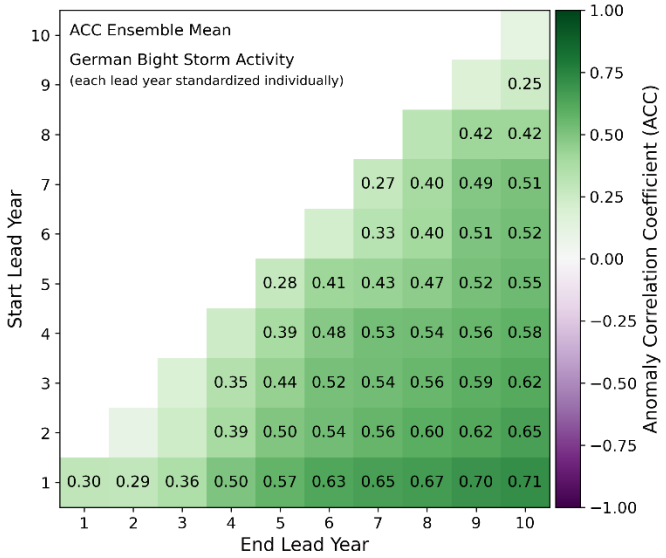

Fig. 4: ACC between observations and ensemble mean predictions of German Bight storm activity (GBSA) for all combinations of start and end lead years. GBSA predictions are based on lead years that are individually standardized by their respective means and standard deviations, i.e., absolute geostrophic winds for lead year 5 are standardized by subtracting the mean and dividing by the standard deviation of lead year 5, instead of lead year 1 like in the original manuscript.

246 The reviewer is correct here. If the entire distribution shifts, the shift would also be present in the ensemble mean. The shape of the distribution and specifically the tails needs to change in order for a probabilistic prediction to gain an advantage over the deterministic prediction. We will amend the paragraph to correctly reflect this explanation.

251 We apologize for being unclear. We will explicitly mention the lead year ranges and then draw conclusions from that.

260 We will highlight the German Bight on the skill maps.

271 We thank the reviewer for bringing up this point. We will rephrase our conclusions to be more precise in stating that we only show two specific lead year ranges for reasons of brevity, but have looked at other lead year ranges as well to draw conclusions that are more general. We will also give a clearer definition on short and long lead year periods in the methods section.

302 We thank the reviewer for that suggestion. Please see our response to comment C2.

319 We will rephrase that sentence to make it less informal.

323 Absolutely. After revisiting the manuscript and considering the points raised by multiple reviewers, we agree that there is the need to compare the model to a climatology-based prediction, which would be a prediction that uses the climatological probabilies of a year falling into a certain category. We originally opted against that to avoid an excessive number of different references in the results section, but we see that this is a more challenging test for the model than a fixed 50% probability forecast. We will test the model against climatology and add the results to the manuscript.

329-338 We will create a separate discussion section to discuss our results, as well as limitations and caveats of the model. Everything from line 304 onwards will be moved to this separate section.

345 That is correct. We will update this sentence.

361 We will expand on this paragraph and give a speculation on the origin of the skill for short averaging periods.

Fig. 1 We will add a figure to the methods section showing a map of the study region and the locations of the three model gridpoints.