Response to Reviewer #1

We sincerely thank Reviewer #1 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.

Data

D1 Just to clarify, you are not using the MiKlip data, but have constructed your own decadal prediction system? I was not sure until I got to line 102...

> Our system is indeed based on one developed within MiKlip, namely the "EnKF" system as described in Polkova et al. 2019. However, this system should not be confused with one of the central prediction systems used during the actual life time of MiKlip. These systems all used oceanic and atmospheric nudging for assimilation and lagged initialization for the ensemble generation.
In contrast to the "EnKF" system within MiKlip, our prediction system includes CMIP6 instead of CMIP5 external forcing, and the hindcasts are run with a total of 80 members, members 17-80 also with 3-hourly output. These 64 members are analyzed in our study.
We will add two sentences at the beginning of the corresponding paragraph.

D2 I do not quite understand how you constructed the 64-member ensemble (L104-111). Please describe this in more detail.

> The initialization of five members each are derived from one assimilation member, the only difference between those five members coming from the perturbation applied to the horizontal diffusion coefficient in the stratosphere. With a 16-member assimilation, this results in 5x16=80 members. However, 3-hourly output is only available for members 17 to 80, which comprise the 64-member ensemble used in our study.
We will adapt the description within the paragraph with a more distinct explanation.

D3 Please clarify which decadal runs you chose. If you are looking at the period 1961-2018, did you select all runs that include those years regardless of the lead time, or is the last run you selected the one that was initialized in 2008?

> We always select the maximum number of possible runs for each lead year range. This means that, for example, the last run used for a lead year 1 evaluation is the one initialized in November 2017, whereas for a lead year 10 evaluation the last run would be the one initialized in November 2008. For longer averaging periods, the last lead year is decisive, so the lead year 4-10 evaluation considers all runs up to 2008, whereas the lead year 4-6 evaluation includes runs up to 2012.
We will add two sentences at the end of the corresponding paragraph to clarify which runs we use.

Methods

M1 *Lead times, part 1:* The selection of lead times seems somewhat arbitrary. Why did you choose 4-10 and 7 and not 1-7 and 4 or 2-8 and 5 ...? Have you checked whether your results/conclusions would be different with a different choice of lead time?

> We thank the reviewer for raising this issue. We are aware that choosing lead years 4-10 and 7 is quite arbitrary, and it would be equally valid to choose 2-8 and 5, or 1-7 and 4. We also checked other combinations of the same averaging period and found that similar general conclusions can be drawn from these lead times (see Fig. 1). We will clarify that for reasons of brevity we just show one example for short (7) and long (4-10) averaging periods, respectively, but the conclusions hold for other lead time combinations as well.

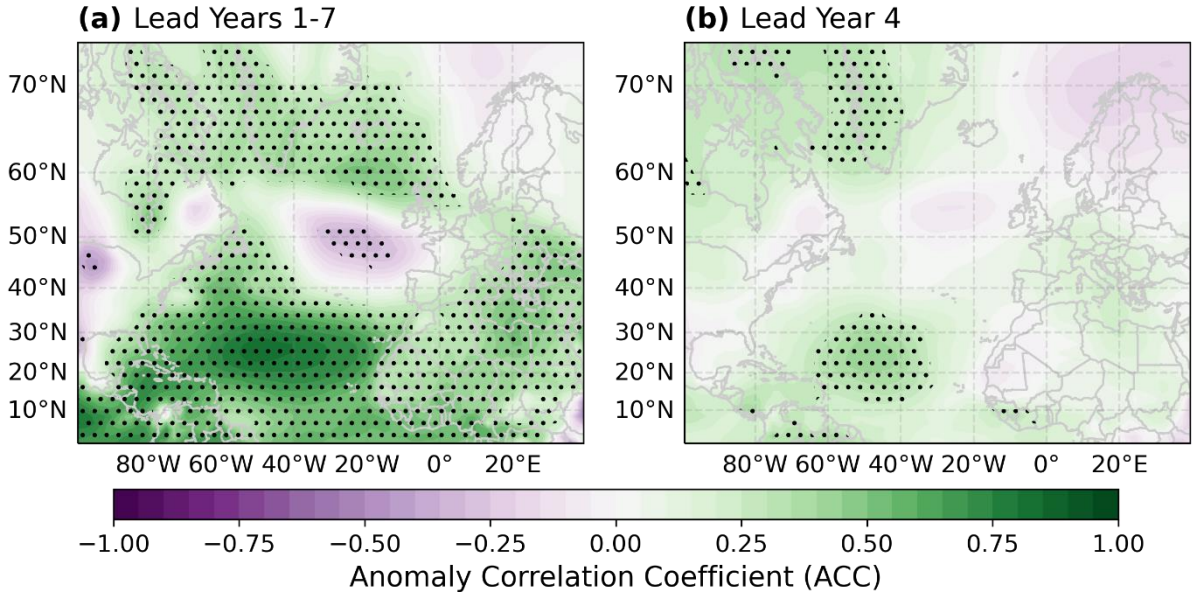

Fig. 1: Gridpoint-wise anomaly correlation coefficient (ACC) for DJF MSLP anomalies between the hindcast ensemble mean and ERA5 for lead years 1-7 **(a)** and lead year 4 **(b)**. Stippling indicates significant correlations (p<=0.05), determined through a 1000-fold bootstrapping with replacement.

M2 *Lead times, part 2:* In L126ff, you state that you focus on lead years 4-10 and 7. However, this only applies to the MSLP anomalies, since you show all possible lead year ranges for GBSA. Please be more specific in this regard.

> We will clarify that we only show lead years 4-10 and 7 for MSLP, but all combinations for GBSA.

M3 *Pressure reduction:* Is this a standard procedure for calculating MSLP from modelled surface pressure? Could you add a reference for equation 1? Does it affect the comparability of your results if you use direct MSLP for one half of the ensemble and calculate MSLP for the other half?

> We will add a reference for the pressure reduction formula, which is based on the US standard atmosphere and a fixed air density, as described in Alexandersson et al. (1998) and Krueger et al. (2019). We also performed a consistency check (Fig. 2) to quantify the MSLP difference between direct and derived output and found that it is negligible for low elevations.

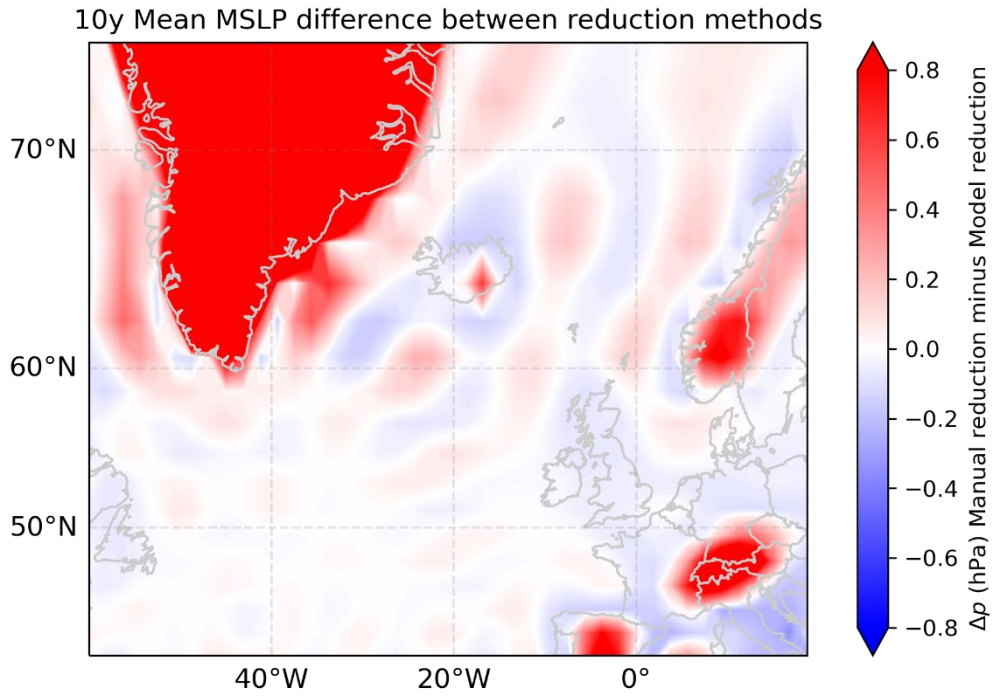

Fig. 2: Difference between manually reduced MSLP and model-output MSLP for one exemplary ensemble member, shown as a 10 year mean (2021-2030) of data from the 2020 initialization. Red colors indicate regions where the manual reduction results in higher MSLP than the automated model output.

M4 *Region of interest:* Please clarify that you are analysing MSLP anomalies for the entire North Atlantic basin (including the German Bight), whereas the GBSA analyses focus only on the German Bight.

> We thank the reviewer for making us aware that this is unclear. We will clarify this paragraph and explicitly state that we investigate MSLP for the whole North Atlantic basin.

M5 *Selection of grid points (L140-144):* This information refers to the generation of GBSA time series, correct? If so, either integrate it in the respective paragraph (L146ff) or clarify why you need to select three grid points. At the moment, the whole paragraph comes a bit out of nowhere, without a clear link to the preceding/subsequent paragraphs...

> We thank the reviewer for this suggestion. We agree that the structure of this part of the method section needs improvement to make it more comprehensible. We will rewrite and restructure the respective paragraphs.

M6 *Generation of GBSA time series:* Did I understand correctly that the time series cover the whole period 1960-2018, while you only use the period 1961-2010 for the standardization?

> That is correct. We base the choice of 1961-2010 as a reference period on Krieger et al. (2020), who also used 1961-2010 to standardize the timeseries. We decided to adapt this reference period in order to introduce as few inhomogeneities as possible.

M7 *Prediction skill:* Please add a short explanation of why it is important to consider both deterministic and probabilistic skill scores when assessing the skill of a decadal prediction system.

> We thank the reviewer for this suggestion. We will add a paragraph on the benefits of assessing the forecast skill of two different prediction methods.

M8 *ACC:* Although this should be common knowledge, please add the possible range of ACC and an explanation of what the different values mean.

> We will add a sentence on the characteristics of the ACC and the possible range in the respective paragraph.

M9 *ACC versus BS:* Be careful when using f and o in equations 2 and 4. You chose the same letters, but they have different meanings (value for ACC, probability for BS). Consider replacing f and o in equation 4 with capital letters.

> We thank the reviewer for making us aware of this unclear nomenclature. We will change the variables in Equation 4 to capital letters to avoid further confusion.

M10 *Choice of BSS:* Out of curiosity – why did you choose the BSS rather than the ranked probability skill score (RPSS)? Since you are interested in three categories (low/normal/high), the RPSS seems the more natural choice to me as it also contains some information about the distance between model and observations.

> We chose the BSS instead of the RPSS as we wanted to investigate whether the model is particularly skillful in predicting one of the three defined categories. By calculating three distinct skill scores for the dichotomous forecasts *high/not high, low/not low, moderate/not moderate*, we want to demonstrate the added value for forecasts of extreme periods, and the inability of the model for forecasts of moderate activity periods. This distinction would not have been possible by calculating the RPSS, which incorporates the skill for every distinct category into one single measure.
We will clarify our intention to use the BSS instead of the RPSS in the respective paragraphs.

Results

R1 *Some thoughts on L234-242:* Could it be that the initialisation has a "negative" impact in the first years (some kind of initialisation shock) – which would explain why the predictive skill is highest for lead time ranges starting in year 3 and 4? This would also fit (to some extent) to previous studies on wind-related variables like Kruschke et al. (2014) or Moemken et al. (2016). However, these studies use uninitialized historical simulations as reference and not persistence…
For temperature, several studies show high predictive skill for later/longer lead times (e.g. Feldmann et al., 2019). This increase seems to originate mainly from the longterm climate trend. However, I have never heard of the importance of climate trend for decadal predictions of wind-based parameters…

> 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 (Fig. 4).

Nevertheless, we will expand the discussion of our results with a paragraph on the effects of a possible initialization shock.
Regarding the climate trend, we agree that the prediction skill for longer lead times can be greatly impacted by the presence of a trend. However, as the reviewer already correctly states, there is little agreement on the response of storm activity to future climate change. Additionally, observational records indicate that, so far, there has not been a significant climate signal in storm activity in our study region, which leads us to believe that long-term trends don't play a major role in the prediction skill here.

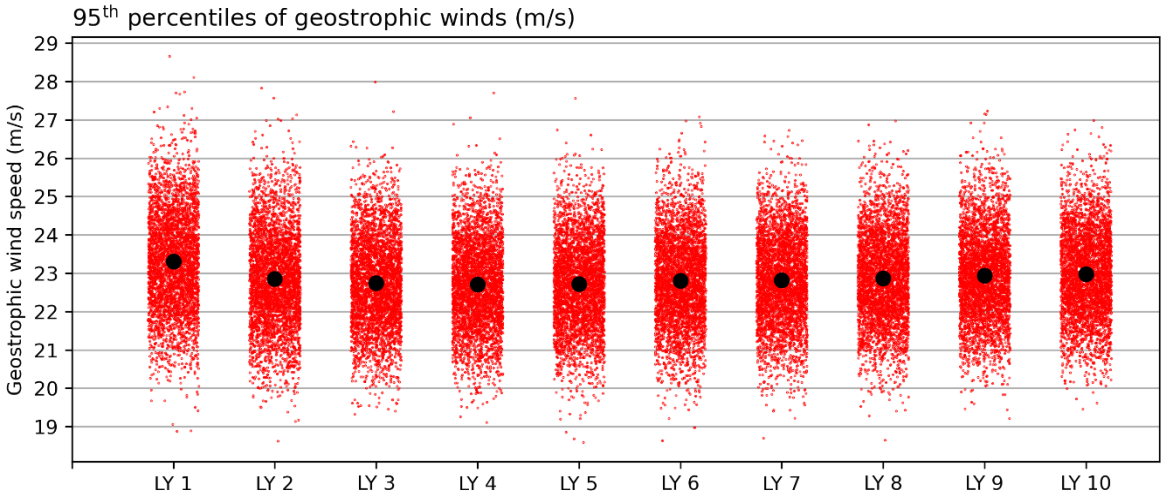

Fig. 3: Absolute 95th annual percentiles of predicted geostrophic winds per lead year. Red dots represent individual members and initialization years, black dots show the ensemble mean for each lead year.

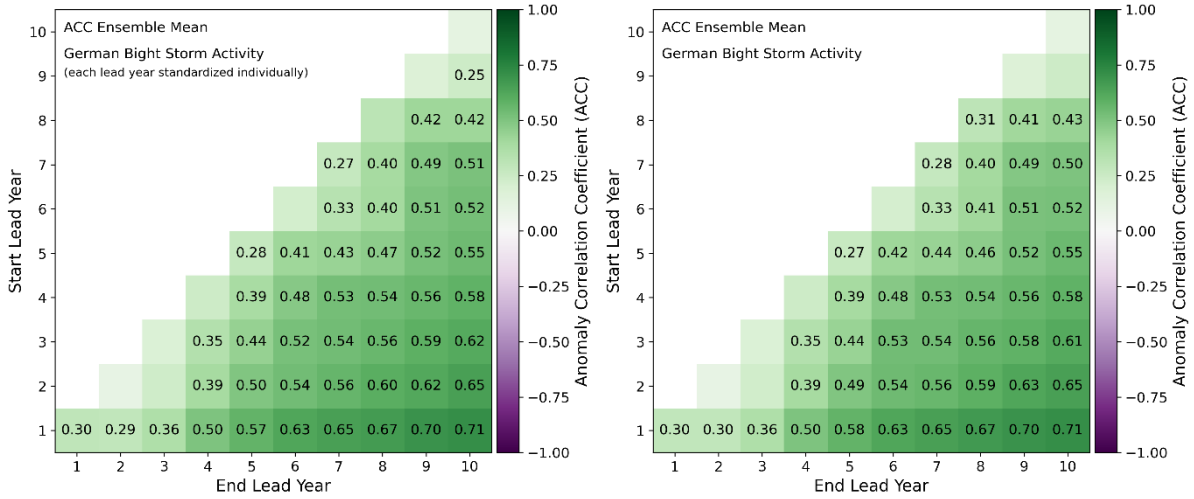

Fig. 4: ACC between observations and ensemble mean predictions of German Bight storm activity (GBSA) for all combinations of start and end lead years. **Left:** GBSA predictions are based on lead years that are individually standardized by their respective means and standard deviations, i.e., absolute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

R2 *L304-338:* These paragraphs seem to be more of a general discussion of your results and are not really related to the rest of section 3.2.2. Therefore, it might make sense to introduce a new section (3.3 Discussion) or new chapter (4. Discussion) for this part of the manuscript.

> We will separate the paragraphs discussing our results from the result description and create a separate section for discussion only.

R3 *Persistence as reference:* Many studies dealing with decadal prediction systems use uninitialized historical simulations of the same model or simple climatology as reference. Is there any particular reason why you have not tried this as well? Please do not get me wrong – I think it is a strength of your study that you consider persistence and random guessing. It just makes it harder to compare your results with other studies on decadal predictions.

> After revisiting the manuscript and reviews, we also see the need to discuss the performance of the model against climatology, as climatology proves to be a tougher challenge than random guessing. We originally opted for persistence and random guessing to not overload the manuscript with a large number of different comparisons, but we agree that using climatology as an additional reference simplifies the comparison of our results with those from other studies. We will restructure the results section and add comparisons to climatology where we see fit.

Figures

F1 For readers unfamiliar with Germany (and the German Bight in particular), it might be helpful to include a figure showing the region of interest. In this, you could also mark the grid points given in Table 1.

> We agree that a map will be helpful and will add one to the methods section.

F2 Figure 2: Please add some explanation in the text (L226-230) about the structure of the plot (that it shows all possible lead time combinations etc.).

> We will add a short introduction on the structure of the matrix plots before summarizing the key findings of Figure 2.

F3 Consider simplifying the captions of Figures 5 and 6 (the same applies to B3 and B4) by saying something like “Same as Figure 4, but for …”.

> We will shorten the repetitive figure captions wherever applicable.

Specific comments

> We will address all minor comments noted by the reviewer as suggested.

References

Alexandersson, H. et al. (1998): Long-term variations of the storm climate over NW Europe, The Global Atmosphere and Ocean System, 6.

Krueger, O. et al. (2019): Northeast Atlantic Storm Activity and Its Uncertainty from the Late Nineteenth to the Twenty-First Century, Journal of Climate, 32, 1919–1931, doi:10.1175/JCLI-D-18-0505.1.

Polkova, I. et al. (2019): Initialization and ensemble generation for decadal climate predictions: A comparison of different methods. Journal of Advances in Modeling Earth Systems 11 (1), 149-172, doi:10.1029/2018MS001439.