Dear Joaquim Pinto,

Thank you for the opportunity to submit a revised version of our manuscript #EGUSPHERE-2022-288 titled “Skillful Decadal Prediction of German Bight Storm Activity” to *Natural Hazards and Earth System Sciences*. We, the authors, would like to express our gratitude to the three anonymous referees for providing valuable feedback on our manuscript.

We would like to give a point-by-point response to the reviewers’ comments in the next section, followed by a list of additional changes that we felt were necessary to improve the quality of the manuscript.

Response to Reviewer #1

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 lifetime 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 expanded the section on our decadal hindcasts to clarify how our prediction system differs from the MiKlip data.

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 replaced 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 added a sentence to the end of the section on decadal hindcasts to clarify the choice of runs.

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. 1.1). We added a statement to the section on lead times 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. We also added similar reminders to the results section where we saw fit.

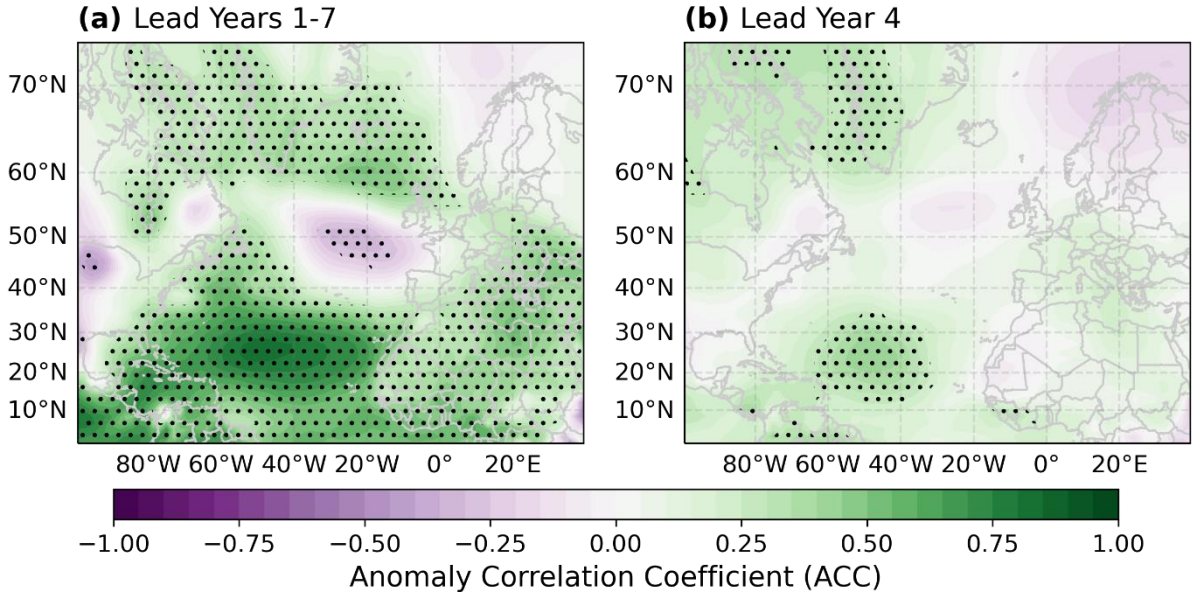

Fig. 1.1: Gridpoint-wise anomaly correlation coefficient (ACC) for DJF MSLP anomalies between the hindcast ensemble mean and ERA5 for lead years 1-7 **(a)** and lead year 4 **(b)**. Stippling indicates significant correlations (p<=0.05), determined through a 1000-fold bootstrapping with replacement.

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 clarified in the section on lead times 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 added 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. 1.2) to quantify the MSLP difference between

direct and derived output and found that it is negligible for low elevations. We therefore added a sentence to the respective section to clarify that we checked the model output for consistency.

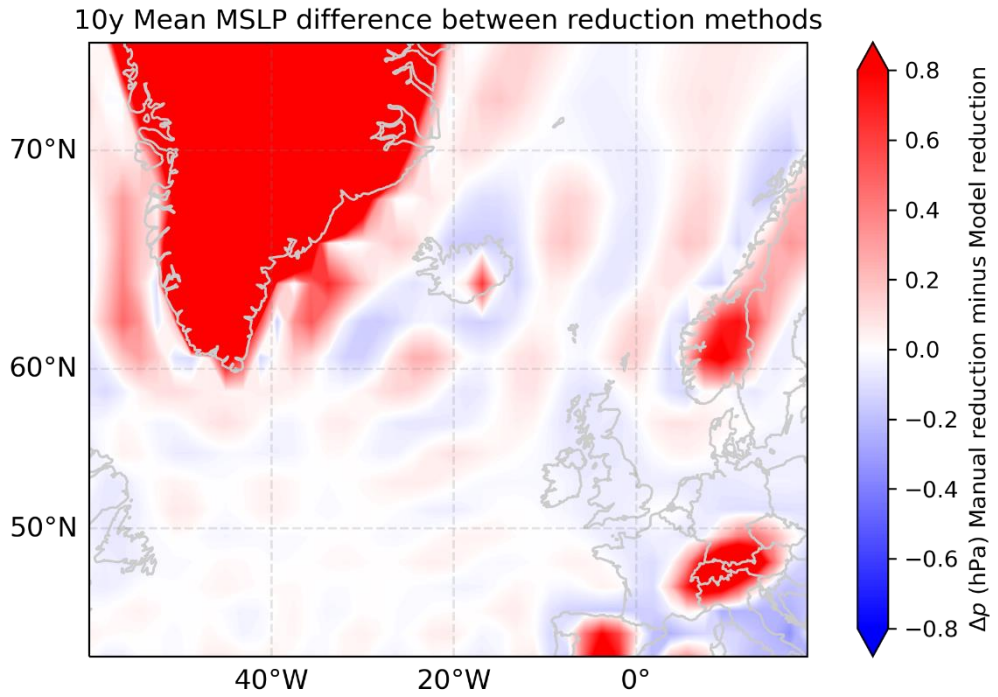

Fig. 1.2: Difference between manually reduced MSLP and model-output MSLP for one exemplary ensemble member, shown as a 10 year mean (2021-2030) of data from the 2020 initialization. Red colors indicate regions where the manual reduction results in higher MSLP than the automated model output.

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 clarified this paragraph and now explicitly state that we investigate MSLP for the whole North Atlantic basin including the German Bight.

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. Based on this suggestion and comments from other reviewers, we restructured the paragraphs on the derivation of GBSA from observations.

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 added two sentences on the importance of both types of predictions to the beginning of the "Evaluation of Model Performance" section.

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 added 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 changed the variables in Equation 4 and the subsequent paragraph 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 high activity 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 added additional explanation behind our intention to use the BSS instead of the RPSS in the respective paragraphs and expanded on the differences between the two metrics in the discussion section.

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. 1.3). 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. 1.4).
Nevertheless, we expanded 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.

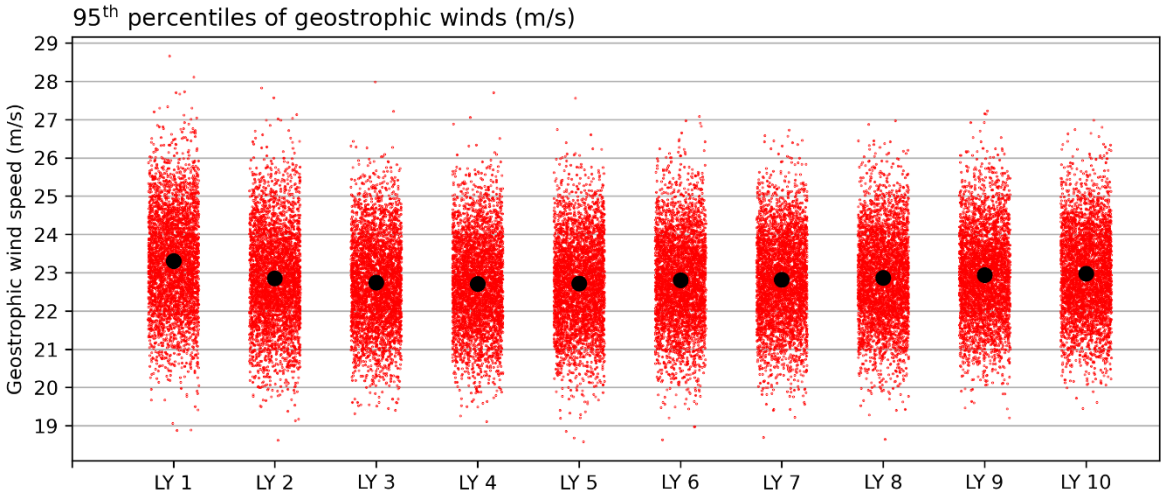

Fig. 1.3: Absolute 95th annual percentiles of predicted geostrophic winds per lead year. Red dots represent individual members and initialization years, black dots show the ensemble mean for each lead year.

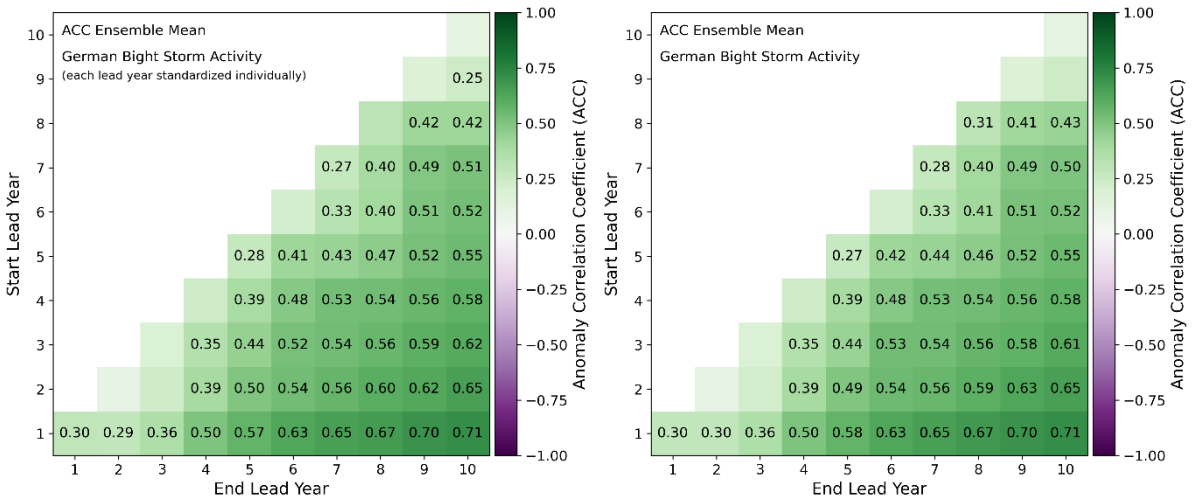

Fig. 1.4: ACC between observations and ensemble mean predictions of German Bight storm activity (GBSA) for all combinations of start and end lead years. **Left:** GBSA predictions are based on lead years that are individually standardized by their respective means and standard deviations, i.e., absolute 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 created a separate discussion section where we discuss our results.

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 saw 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 restructured the results section, added climatology as a reference, and removed the sections and plots discussing the coin-flip-based random guessing.

<u>Figures</u>

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 added a map of the German Bight that shows the location of the triangle 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 added a short introduction to the structure of the matrix plots before summarizing the key findings of Figure 6 (Figure 2 in the old manuscript).

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 simplified the respective figure captions.

Specific comments

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

Response to Reviewer #2

2.1 Conclusions

2.1A On the effect of autocorrelated time series on increased correlation coefficients for longer averaging periods

> We agree with the reviewer. We added a paragraph to the discussion section with additional thoughts on the effect of averaging window length on the correlation of associated time series. Please also see our reply to comment 2.4.1B.

2.1B On the choice of reference forecasts and the effect of estimating correlations from smoothed timeseries

> We agree that there is a need to discuss the effect of the choice of reference in greater detail. We improved the discussion section to include a comment on the lead-time dependence of persistence (comment 2.4.3A). We also replaced the coin-flip-based random guessing with climatology and discussed this choice of reference as well (more details in comment 2.4.3C).

2.1C Revisiting the conclusion that is based on the choice of the Brier Skill Score instead of the RPSS

> The reviewer is correct in assuming that we refer to the RPS/RPSS when mentioning "highly aggregated probabilistic skill scores". We added a paragraph to the newly formed discussion section to highlight our intent and elaborate more on the differences between the general concepts of the RPS and the BS, which we also explain in our reply to comment 2.4.2A.

2.1D On a possible initialization shock or model drift

> We thank the reviewer for suggesting the possibility of an initialization shock or model drift. In fact, the predicted geostrophic winds are lowest for lead years 3-5, and highest for lead year 1 (Fig. 2.1). 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. 2.2). Nevertheless, we expanded the discussion of our results with a paragraph on the effects of a possible initialization shock.

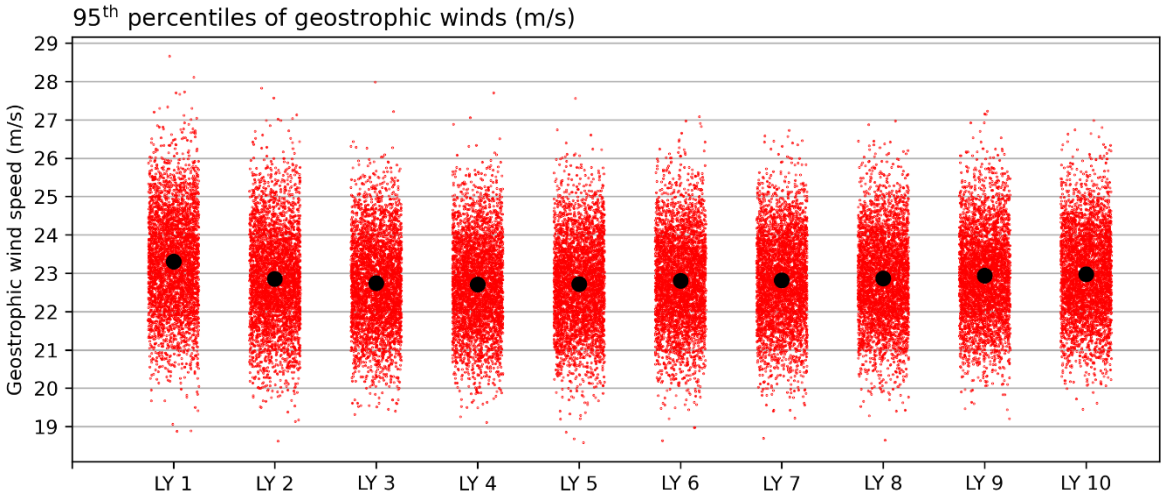

Fig. 2.1: Absolute 95th annual percentiles of predicted geostrophic winds per lead year. Red dots represent individual members and initialization years, black dots show the ensemble mean for each lead year.

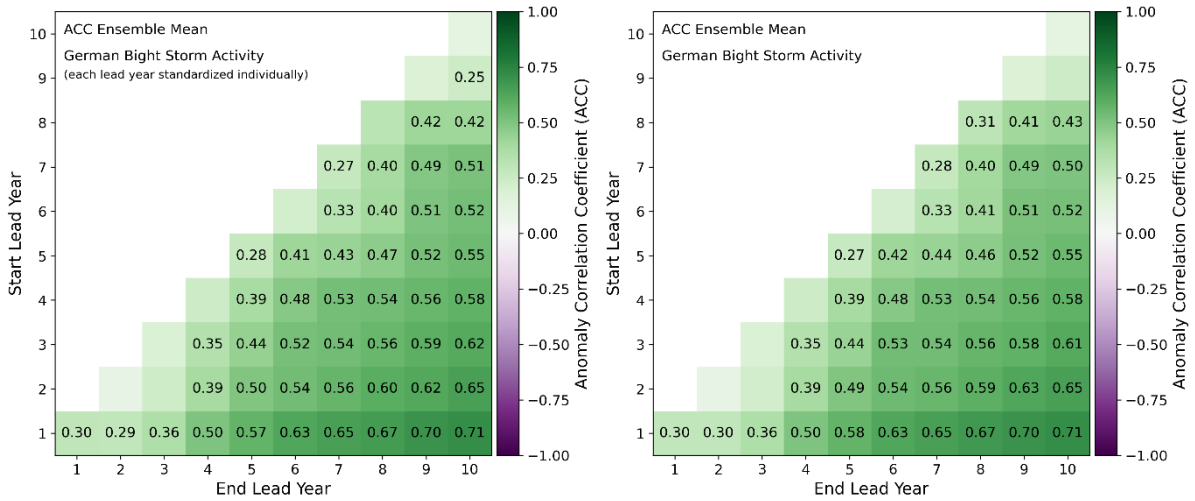

Fig. 2.2: ACC between observations and ensemble mean predictions of German Bight storm activity (GBSA) for all combinations of start and end lead years. **Left:** GBSA predictions are based on lead years that are individually standardized by their respective means and standard deviations, i.e., absolute 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.

2.2 Terminology

2.2A On the usage of the term “skill”

> The reviewer is correct that the correlation coefficient per se is not a measure of forecast skill, but much rather a measure of linear association. We replaced the term skill by simply referring to the ACC instead.

2.2B On the usage of the terms “deterministic skill” and “probabilistic skill”

> We agree that the terms “deterministic skill” and “probabilistic skill” are not precise, as “deterministic” and “probabilistic” refer to the forecast types. We changed the wording and now refer to the skill of the (probabilistic) prediction instead.

2.3 Structure

2.3A On the structure of the section on pressure reduction and the derivation of GBSA

> We agree that the title of this subsection needed to be changed to reflect its scope more accurately. We changed the title of the section to “Geostrophic Wind and German Bight Storm Activity” as suggested. We also restructured the entire section and added details to various aspects of the methodology.

2.3B On subdividing the model evaluation section into two parts.

> We agree that dividing this section makes it clearer to the reader that we are introducing two different concepts here. We split up the section into two parts, one on the ACC and one on the BSS.

2.3C On establishing a separate discussion section.

> We agree that the text from 304 onwards discussed the results rather than describing them. We therefore created a separate discussion section.

2.4 Statistical concepts

2.4.1 Anomaly correlation

2.4.1A On the effect of autocorrelation on significance

> We agree on this point. Calculating confidence intervals and significance levels via the Fisher-z transformation requires independent samples, an assumption that is not satisfied in our case due to autocorrelation. We recalculated the significance with a block-bootstrapping approach, as especially the smoothed (multi-year average) time series are heavily autocorrelated. We also revised the section on the calculation of anomaly correlations accordingly.

2.4.1B On the association between forecast and observations and its effect on correlation

> We thank the reviewer for the explanation and sample code on the effect of autocorrelation on the correlation coefficient of smoothed time series. While it doesn’t explain the entirety of the ACC increase for longer averaging periods, it might account for a part of it. We added our thoughts on this effect to our discussion section.

2.4.2 Nature of the probabilistic forecast and Brier score

2.4.2A On the choice of the Brier score instead of the RPS/RPSS

> We thank the reviewer for bringing up the issue of evaluating a 3-category forecast with the Brier (skill) score. We completely agree that the RPS/RPSS is the correct evaluation metric for a 3-category forecast. However, we are not aiming at correctly predicting which category out of the three will occur, but much rather whether the model shows skill for a 2-category forecast with different event thresholds. To use the reviewer's analogy, we are interested how often (out of the three options) the model succeeds in juggling with two balls, not whether the model succeeds in juggling with three balls. While the RPS/RPSS acts as a metric for how well the model performs for a 3-category forecast, it is unable to show whether, for example, a *high vs. no high activity* prediction is more skillful than a *low vs. no low activity* prediction. In our manuscript, the additional skill of the model for high storm activity predictions compared to persistence and climatology would not have been detectable with a single three-category forecast and the RPSS. We totally agree that we needed to clarify this intent better in order to avoid the impression that we aim at generating a 3-category forecast and appreciate your insightful thoughts on this matter. We dedicated a part of the discussion and method sections to our intent and revised paragraphs where our choice of evaluation metric had not been stated clearly before.

2.4.2B On the explicit clarification of the "standardization"

> We agree with the reviewer that we needed to clarify our standardization process more explicitly. We rephrased the part of the methods section that introduces the standardization to make this matter clearer.

2.4.2C On obtaining the probabilistic forecasts and re-standardization

> We apologize for being vague here. The probabilities are obtained by counting the number of members above/below a category threshold and dividing this number by the total number of members in the ensemble (64). The time series of moving averages for longer periods are standardized again, so that we always compare predicted and observed time series, which by definition have a mean of 0 and a standard deviation of 1. We agree that we needed to describe this more thoroughly, as it indeed makes a difference. We revised the method section and created a separate subsection on re-standardization to reflect this procedure.

2.4.2D On the choice of lead years 4-10

> We agree that the choice of lead years 4-10 (and 7) appears quite arbitrary. We could have also done the analysis for lead years 1-7 and 4 (Fig. 2.3) and drawn equally valid conclusions from those lead times. We added a statement to the section on lead times 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. We also added similar reminders to the results section where we saw fit.

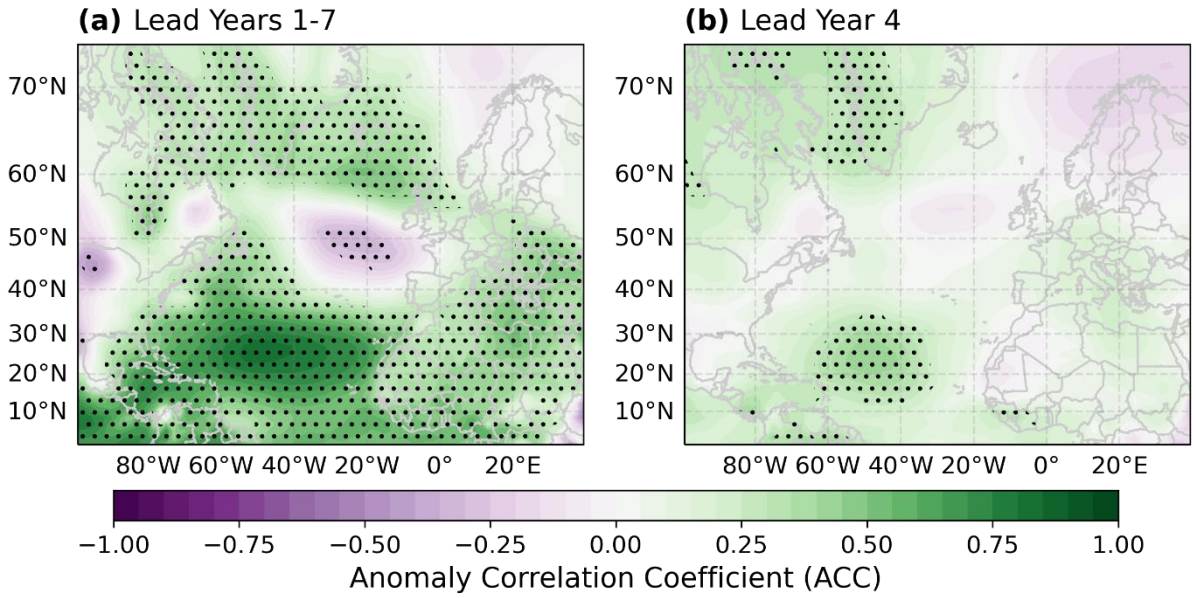

Fig. 2.3: Gridpoint-wise anomaly correlation coefficient (ACC) for DJF MSLP anomalies between the hindcast ensemble mean and ERA5 for lead years 1-7 **(a)** and lead year 4 **(b)**. Stippling indicates significant correlations (p<=0.05), determined through a 1000-fold bootstrapping with replacement.

2.4.3 Reference forecast

2.4.3A On the lead-time dependence of the skill of persistence forecasts

> We thank the reviewer for their thoughts on the lead-time dependence of persistence. We calculated the Brier Score for persistence (Fig. 2.4). While there are certain lengths of averaging periods, for which the BS is lower, there is no general decline in BS with increasing lead time or averaging period. We believe that including plots of the Brier Skill for all forecasts and categories would clutter the manuscript. However, we added paragraphs to the newly created discussion section where we discuss the performance of persistence and its effect on the BSS in more detail.

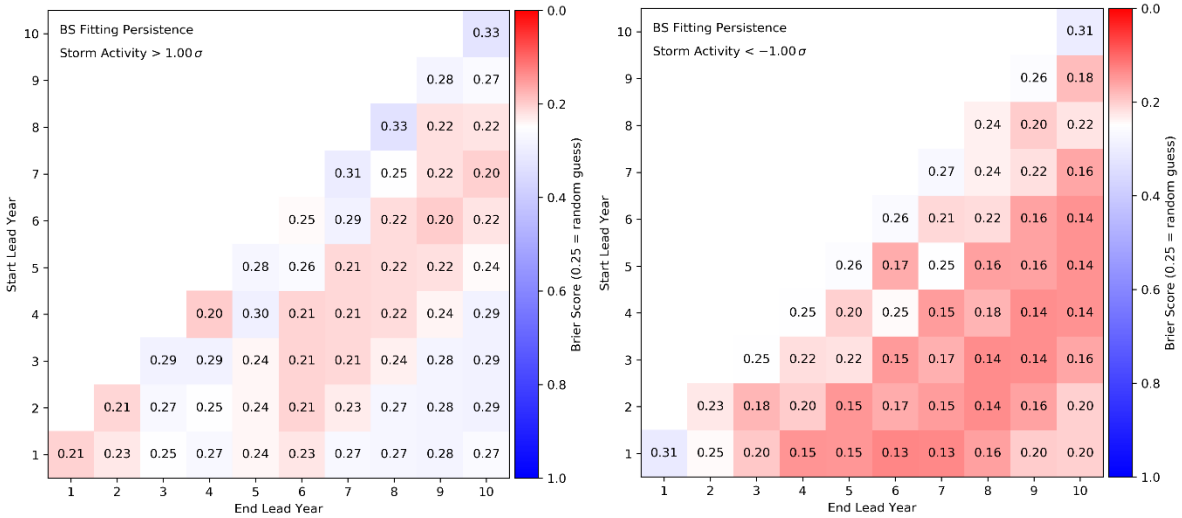

Fig. 2.4: Brier Score of persistence for all combinations of start and end lead years for high storm activity **(left)** and low storm activity **(right)**. Reddish colors indicate BS lower (= better) than that of a coin-flip forecast, bluish colors indicate BS higher (= worse) than that of a coin-flip forecast.

2.4.3B On random forecasts

> We are sorry for causing confusion by our ambiguous use of the term “random forecast”. Our intention was to create a reference forecast that always predicts a fixed probability of 50%, not one that randomly selects one of two outcomes every year. After consideration, we decided to drop the original 50%-“random” forecast from the manuscript entirely and focus on climatology as our second reference prediction instead. We agree that a reference that draws from the historical climatological distribution (as e.g. described in Wilks, 2011) reflects a random prediction more accurately. Please also see our reply to comment 2.4.3C below.

2.4.3C On climatology as a reference forecast

> After revisiting the manuscript and taking into account the points raised by multiple reviewers, we agree that there was the need to compare the model to a climatology-based prediction. 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 revised the results by adding climatology as our second reference prediction and removed the results and discussions related to the fixed 50% prediction. In this process, we also revised the methodology, discussion, and conclusion sections to fit the newly included climatology prediction.

2.5 Comments on specific lines in the manuscript

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 We thank the reviewer for making us aware of this issue. Our intention behind this statement was to mention that, for certain lead times, a probabilistic forecast can show significant skill, while a deterministic forecast produces insignificant correlation coefficients, and vice versa. We revised the abstract and in the process clarified the confusing statements.

13 We added an example for the benefit of decadal predictions a.

32 We addressed the methodical difference between evaluating a 3-category forecast with the RPSS (as in Kruschke et al., 2016) and evaluating three separate 2-category forecasts with the BSS and the implications of doing so in the newly created discussion section. Please also see our reply to comment 2.4.2A.

62 We thank the reviewer for bringing up this additional detail. The uncertainty of a forecast is a great advantage of probabilistic predictions.
With the quoted statement, we intended to explain that periods of high and low storm activity might be detectable through changes in the shape of the ensemble distribution and its tails. If the whole distribution shifts towards one direction, the shift should be notable in both a probabilistic and a deterministic (ensemble mean) prediction. However, if the shape of the tails in particular changes, a probabilistic prediction might hold an advantage over the deterministic prediction. We amended the paragraph to correctly reflect this explanation.

74 We thank the reviewer for pointing out the difference of using a skill score for deterministic predictions like the MSE and a measure for linear association like the ACC. We believe that the ACC is widely established as a metric to quantify the ability of a climate model ensemble mean to predict the temporal evolution of a quantity. We therefore decided to keep the ACC as our metric but replaced the term “deterministic skill” with ACC (compare comment 2.2A).

81 That is correct. We standardize time series by applying a z-transformation, i.e., by subtracting the mean and dividing by the standard deviation. We added a clearer definition of the standardization process to the respective paragraphs.
We also added an exemplary histogram of geostrophic wind speeds from one ensemble member, one model run, and one forecast year to illustrate the distribution of geostrophic winds. We also added a violin plot of the distributions of 95th percentiles of geostrophic wind speed in the model ensemble, separated by lead year. For observational GBSA, which is based on an average of 18 different standardized time series of geostrophic winds from overlapping triangles, we do not see fit for such a graphic in the manuscript. The absolute 95th percentiles vary depending on the size of the individual triangles, preventing any generalization of the absolute wind speeds. In order to attempt a rough comparison between model and observations, we compared the range and mean of the modeled 95th percentiles of geostrophic wind speeds to the average 95th percentiles of the observed geostrophic wind speeds from all 18 triangles in Krieger et al. (2020) and found that the mean values agree.

146 We apologize for leaving this unclear. We use the MSLP gradient of a plane through three grid points, as it was done in previous studies (e.g., Alexandersson et al., 1998; Krueger et al., 2019). We rephrased this sentence to avoid misconceptions.

147 We enhanced the explanation of the concept of using standardized 95th percentiles and added an exemplary histogram of absolute geostrophic wind speeds to the respective section.

154 We thank the reviewer for this suggestion and agree that this title would fit the section better. We updated the title of the section to “Evaluation of Model Performance”.

155 Please see our reply to comment 2.2A.

166 We amended that sentence to include the dichotomous nature of the Brier score. Please see also our reply to comment 2.4.2A.

173 We apologize for not going into enough detail here. We bootstrap by sampling different forecast/initialization years with replacement. We do not apply the bootstrap to sample ensemble members. We revised the wording in this paragraph.

174 We thank the reviewer for pointing out this imprecise wording. We amended the sentence to be more accurate.

176 We changed the sentence to be more accurate.

179 The reviewer is correct; our phrasing was ambiguous. We updated this sentence following the suggestion.

180 We thank the reviewer for this suggestion. We changed the wording and now elaborate more on the characteristics of the Brier Score.

182 Please see our reply to comment 2.4.3B.

185 That is correct. We removed the sigma from the category thresholds.

188 Yes, both GBSA time series from the model and the observations are standardized. We reworded this sentence to clarify.

189 We apologize for causing confusion here. GBSA is derived from the MSLP gradient of a plane through three grid points, as the reviewer correctly notes. The gradient of the plane can be interpreted as the average horizontal MSLP gradient of the triangle spanned by the three grid points, but there is no averaging of different gradients involved in the derivation of GBSA. We reworded this sentence to avoid misconceptions.

192 Please see our reply to comment 2.4.3A.

194 We apologize for being unclear here. The persistence forecast is not a probabilistic forecast, but rather a deterministic one. We use the average storm activity of the past *n* years as a persistence forecast for our target lead years and assign it a Brier Score of either 0 or 1, depending on whether the persistence forecast is on the same side of the threshold as the observation or not. In other words, the persistence always forecasts a probability of either 0% or 100% for an event to happen.
We rephrased the respective paragraphs to clarify that persistence per se is a deterministic prediction that is used as a reference to evaluate the skill of probabilistic predictions of our model.

204 Please see our reply to comments 2.2A and 2.2B.

205/209 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 removed these statements and replaced them by a note that, while the correlation itself is significant, this is of little use for a skillful prediction.

210 Yes. We rephrased that sentence to clarify that we refer to the absolute correlations being larger on average.

211 Yes. We thank the reviewer for making us aware of that. We removed the redundant statement.

212 We thank the reviewer for this suggestion. We added a paragraph on the effect of averaging periods on ACC to the discussion section.

262 We added thoughts on the performance of persistence to the newly created discussion section. Please also see our reply to comment 2.4.3A.

263 Please see our reply to comment 2.4.3C.

265 We apologize for this misleading terminology. We removed the word “absolute”, as it does not fit here.

266 Please see our reply to comment 2.4.3C.

267 We agree.

297 Please see our reply to comment 2.4.3C.

311 We agree with the reviewer that this part of the discussion needed to be enhanced. Please see our reply to comments 2.4.2A and 32.

316 We thank the reviewer for this suggestion. Please see our reply to comment 2.4.3A.

323 Please see our reply to comment 2.4.3C.

325 We agree that it is certainly also a deficit that comes with using persistence as a reference. We think that an improvement in skill over a reference can be seen in two ways, both as a valuable aspect of the DPS and as a deficit of the reference. We expanded on the performance of reference forecasts and the implications in the discussion section.

335 We thank the reviewer for this thought. Using the 0.16 and 0.84 quantiles would indeed eliminate the need for normal distributed quantities.

Minor comments

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

Response to Reviewer #3

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. 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. 3.1). We added a statement to the section on lead times 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. We also added similar reminders to the results section where we saw fit.

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 needed improvement. We rearranged the figures and, in the process of replacing random guessing with climatology, moved all figures from the appendix to the main body of the paper.

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 added a paragraph on this contradiction to the discussion section.

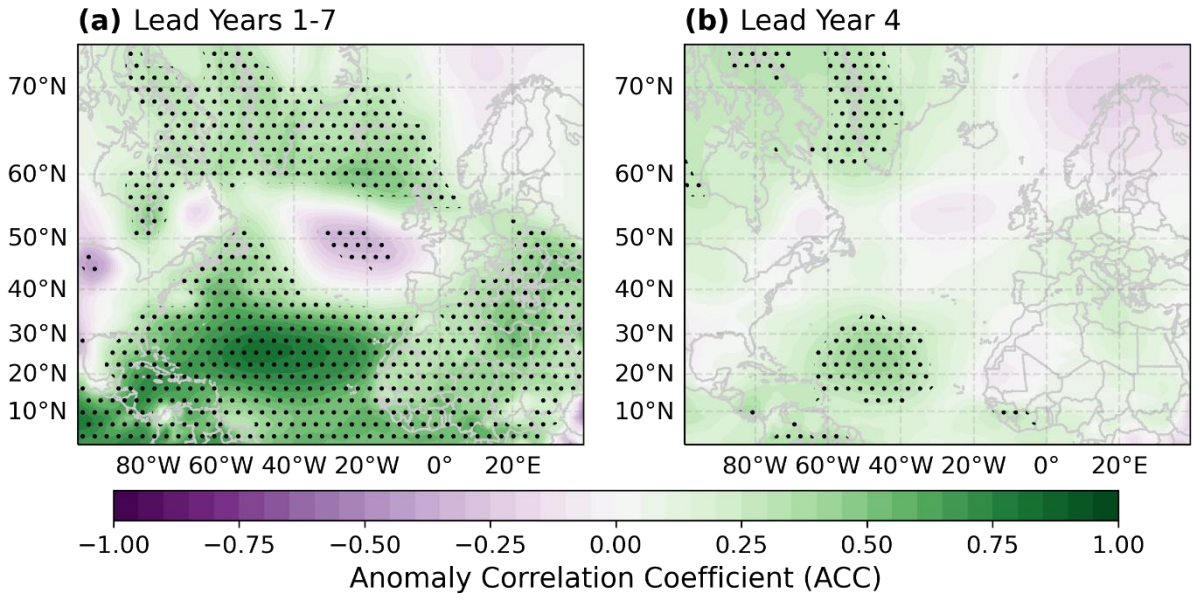

Fig. 3.1: Gridpoint-wise anomaly correlation coefficient (ACC) for DJF MSLP anomalies between the hindcast ensemble mean and ERA5 for lead years 1-7 **(a)** and lead year 4 **(b)**. Stippling indicates significant correlations (p<=0.05), determined through a 1000-fold bootstrapping with replacement.

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 changed the wording to point out 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 reduced the usage of these terms in the manuscript.

<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 rephrased the abstract accordingly.

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 saw the need to motivate this choice better and therefore enhanced 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. 3.2), 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.

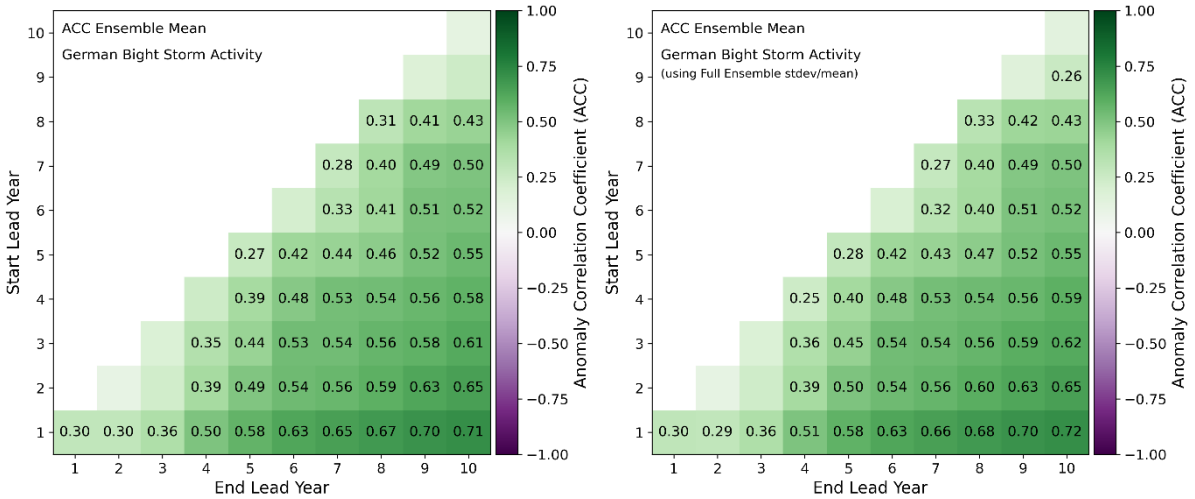

Fig. 3.2: ACC between observations and ensemble mean predictions of German Bight storm activity (GBSA) for all combinations of start and end lead years. GBSA predictions are based on individually standardizing members with their respective means and standard deviations **(left)**, and on standardizing 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 recalculated the significance with a block-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. Therefore, we decided to add exemplary time series of predicted and observed GBSA for lead years 4-10 and 7.

195 We rephrased this sentence.

205/211 Please see our response to comment C4.

224 We removed “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. 3.3). 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. 3.4 and Fig. 3.2 left). Nevertheless, we expanded the discussion of our results with a paragraph on the effects of a possible initialization shock.

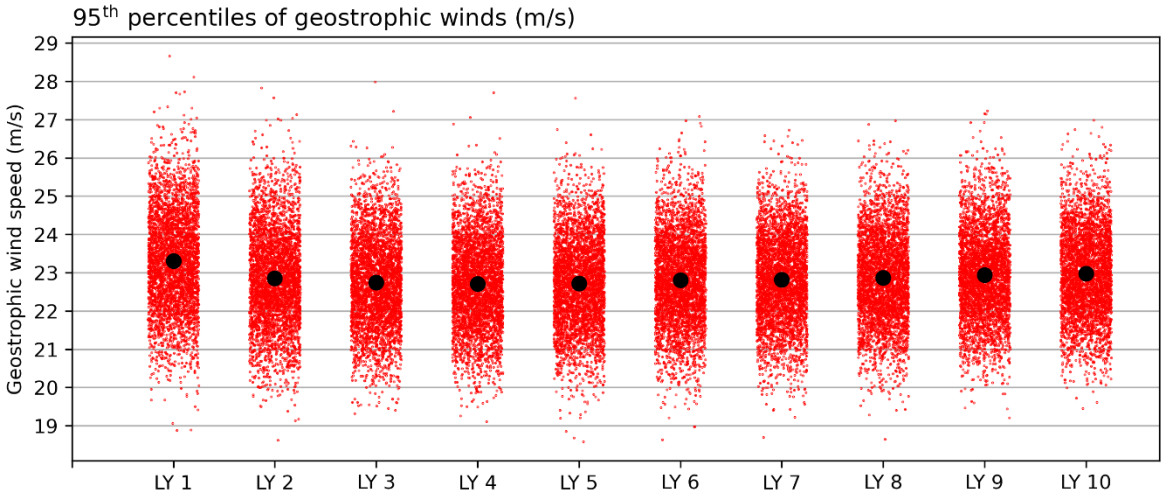

Fig. 3.3: Absolute 95th annual percentiles of predicted geostrophic winds per lead year. Red dots represent individual members and initialization years, black dots show the ensemble mean for each lead year.

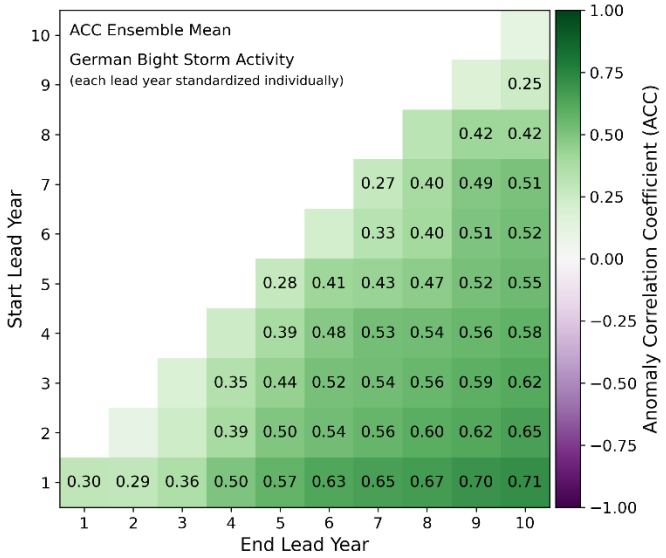

Fig. 3.4: ACC between observations and ensemble mean predictions of German Bight storm activity (GBSA) for all combinations of start and end lead years. GBSA predictions are based on lead years that are individually standardized by their respective means and standard deviations, i.e., absolute geostrophic 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 amended the paragraph to correctly reflect this explanation.

251 We apologize for being unclear. We changed the respective sentences to now precisely state whether we talk about specific lead years, averaging periods, or draw general conclusions.

260 We added red dots to mark the German Bight on the skill maps.

271 We thank the reviewer for bringing up this point. We rephrased the respective sections 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 more general conclusions. We also added a clearer definition of 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 reworded 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 a need to compare the model to a climatology-based prediction, which would be a prediction that uses the climatological probabilities 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 revised the choice of reference, removed random guessing from the manuscript, and replaced it with climatology.

329-338 We created a separate discussion section to discuss our results, as well as limitations and caveats of the model.

345 That is correct. We updated this sentence.

361 We expanded on this paragraph and added our thoughts on the origin of the skill for short averaging periods, also taking the performance of the reference prediction into account.

Fig. 1 We added a map to the methods section that shows the study region and the location of the German Bight triangle.

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.

Kruschke, T. et al. (2016): Probabilistic evaluation of decadal prediction skill regarding Northern Hemisphere winter storms, Meteorologische Zeitschrift, 25, 721–738, doi:10.1127/metz/2015/0641.

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.

Wilks, D.S. (2011): Statistical Methods in the Atmospheric Sciences. 3rd Edition, Academic Press, Oxford.

Additional Changes

We revised the abstract to include the findings from our comparison with climatology and removed some misleading statements.

We corrected values in the results section that were slightly affected by changing the way of calculating significance.

We reduced the length of the conclusion section, since we believe some of the points fit better into the new discussion session and do not have to be repeated in the conclusion.

We removed the second part of the appendix that contained the comparison with random guessing. As the comparison with random guessing is not part of the manuscript anymore, we see no need to keep this section. Furthermore, all new figures that show the comparison with climatology are now presented in the main body of the manuscript.

In addition, we corrected several typos and grammatical errors throughout the manuscript.

We would like to thank the reviewers again for their time and effort and their valuable and insightful comments on this manuscript.

We look forward to hearing from you regarding your decision on the manuscript and are happy to respond to any further questions or comments.

Sincerely,

Daniel Krieger
Corresponding Author