Response to Reviewer #2

We sincerely thank Reviewer #2 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.

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 will expand our discussion and conclusion sections 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 details. We will enhance the conclusions to also include the lead-time dependence of persistence (comment 2.4.3A) and a comparison with climatology (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 will enhance this section of the conclusions to bring in 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. 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).
Nevertheless, we will expand the discussion of our results with a paragraph on the effects of a possible initialization shock.

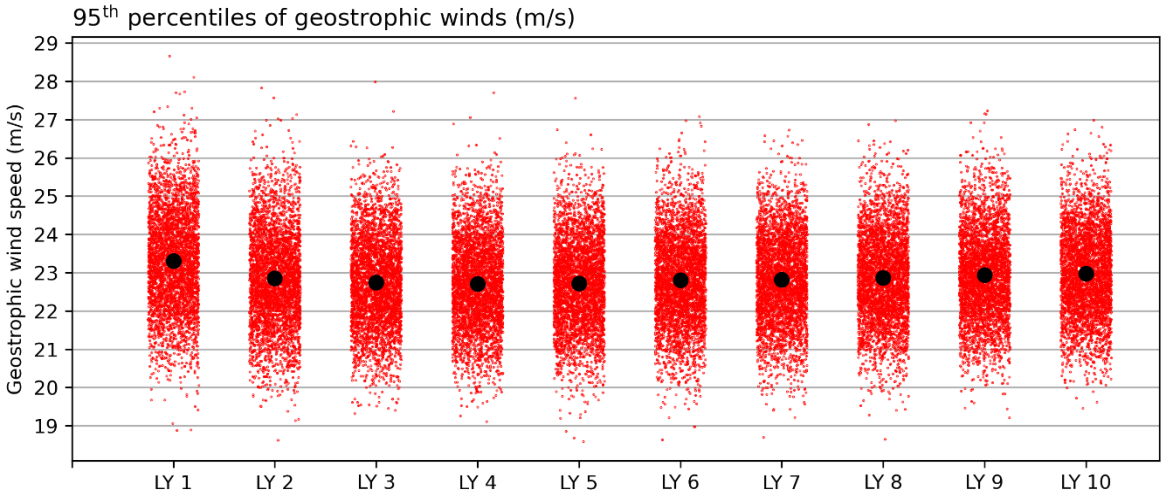

Fig. 1: Absolute 95th annual percentiles of predicted geostrophic winds per lead year. Red dots represent individual members and initialization years, black dots show the ensemble mean for each lead year.

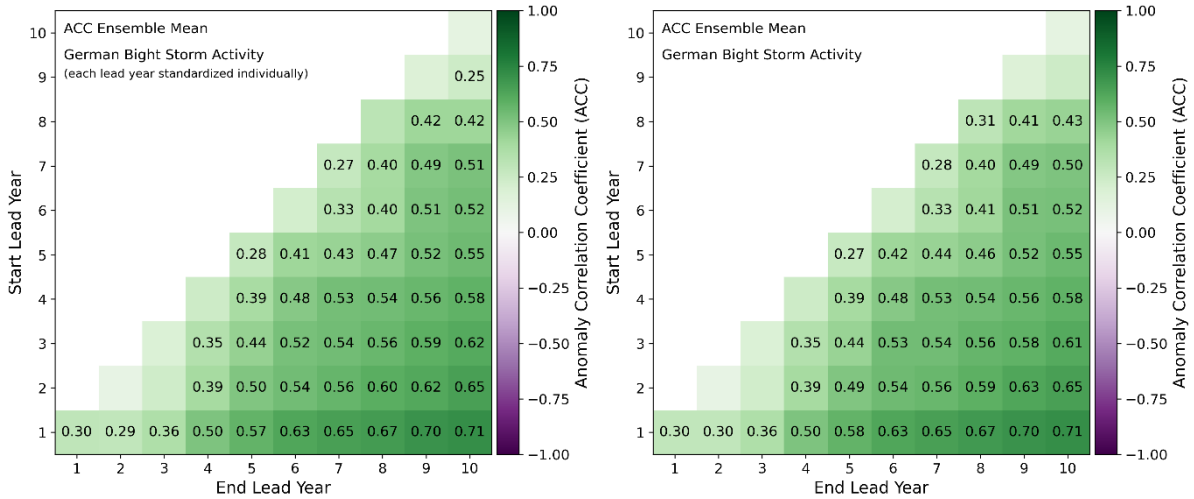

Fig. 2: ACC between observations and ensemble mean predictions of German Bight storm activity (GBSA) for all combinations of start and end lead years. **Left:** GBSA predictions are based on lead years that are individually standardized by their respective means and standard deviations, i.e., absolute 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 will refrain from calling the ACC a skill score.

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 will resort to a more precise terminology.

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 needs to be changed to more accurately reflect its scope. We will reword the title and also restructure this subsection.

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 will split up the section.

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 see the need to create a separate discussion section and will add one in the updated manuscript.

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 will recalculate the significance with a block-bootstrapping approach, as especially the smoothed (multi-year average) time series are heavily autocorrelated.

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 will add this effect to our discussion and conclusion sections.

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. We totally agree that we need to clarify this intent in order to avoid the impression that we aim at generating a 3-category forecast and appreciate your insightful thoughts on this matter.

2.4.2B On the explicit clarification of the “standardization”

> We agree with the reviewer that we need to clarify our standardization process more explicitly. We will rephrase the part of the methods section which 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 need to describe this more thoroughly, as it indeed makes a difference. We will improve the section on 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. 3) and drawn equally valid conclusions from those lead times. 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.

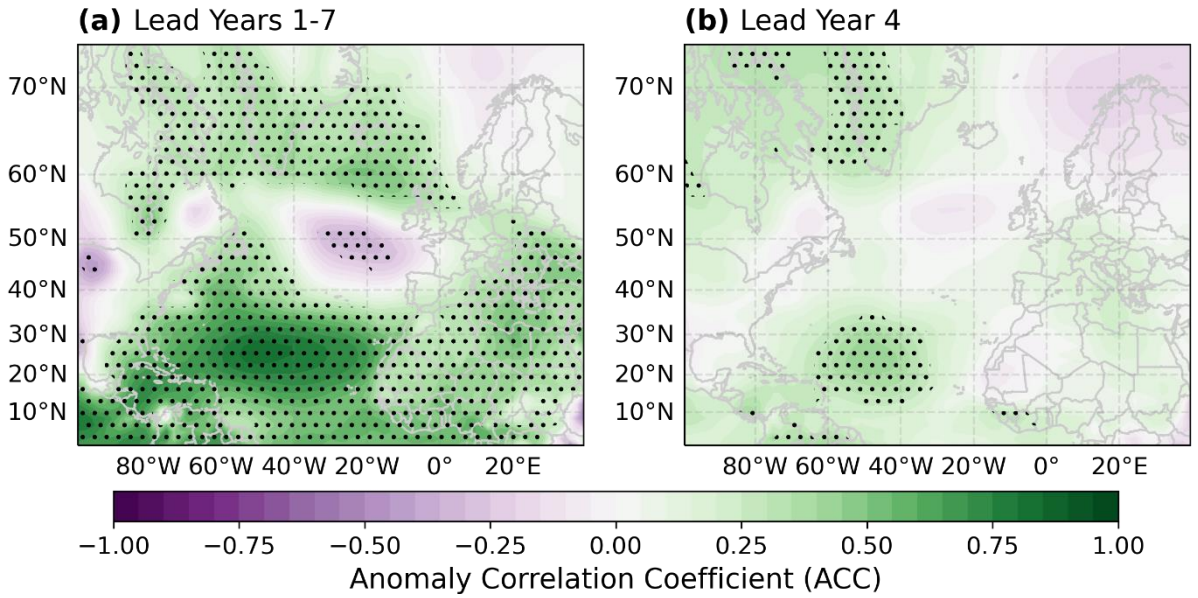

Fig. 3: Gridpoint-wise anomaly correlation coefficient (ACC) for DJF MSLP anomalies between the hindcast ensemble mean and ERA5 for lead years 1-7 **(a)** and lead year 4 **(b)**. Stippling indicates significant correlations (p<=0.05), determined through a 1000-fold bootstrapping with replacement.

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. 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 will improve the conclusions by discussing the performance of persistence in more detail.

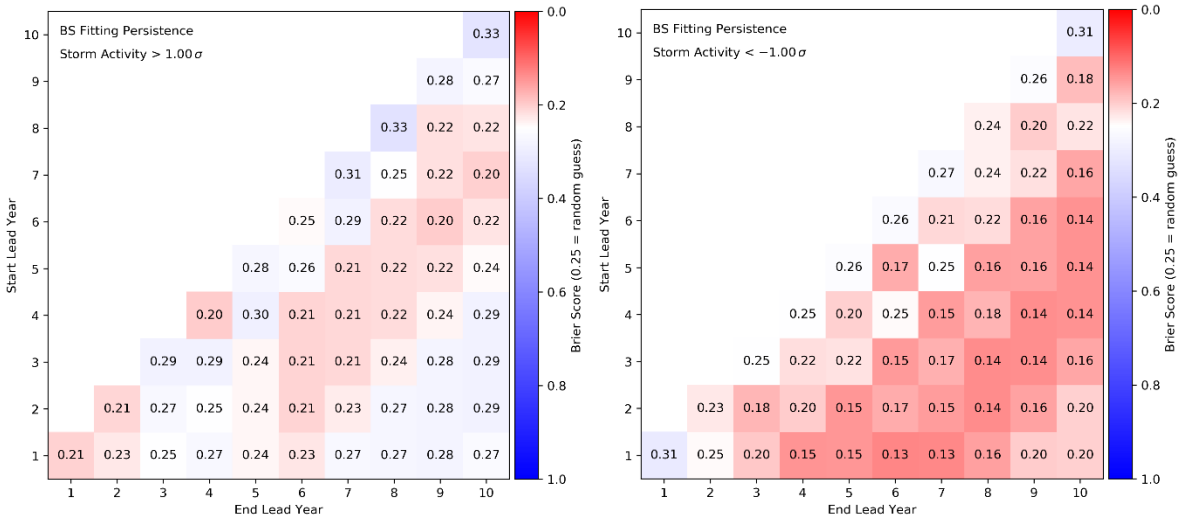

Fig. 4: Brier Score of persistence for all combinations of start and end lead years for high storm activity **(left)** and low storm activity **(right)**. Reddish colors indicate BS lower (= better) than that of a coin-flip forecast, bluish colors indicate BS higher (= worse) than that of a coin-flip 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. We will rename the previous “random forecast” to avoid confusion, and additionally discuss the performance of the model compared to a true random forecast with fixed climatological probabilities (as e.g. described in Wilks, 2011). 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 is 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 will test the model against climatology and add the results to the manuscript.

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 understand that this distinction needs to be made a lot clearer, and we will revise this paragraph accordingly.

13 We will add an example for the benefit of decadal predictions and a reference to this paragraph.

32 We will address 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 greater detail. 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 will amend 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 will therefore opt to keep the ACC as our metric, and instead stop referring to it as “deterministic skill” (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 will add a clearer definition of the standardization process to the respective paragraphs.
As for the histogram, we agree that a histogram of (absolute) wind speeds will benefit the methods section. However, we believe that it might be more suited to add it to the paragraph on deriving GBSA from model output, mainly because the observational GBSA is based on an average of 18 different standardized time series of geostrophic winds from overlapping triangles. The absolute 95^{th} percentiles vary depending on the size of the individual triangles, preventing any generalization of the absolute wind speeds. For the model output, it might be more consistent to show a histogram of absolute wind speeds.

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 will rephrase this sentence to avoid misconceptions.

147 We will enhance the explanation of the concept of using standardized 95^{th} percentiles and add an exemplary histogram of absolute 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 will update the title of the section.

155 Please see our reply to comment 2.2A.

166 We will amend 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 will clarify our approach in this paragraph.

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

176 We will change the sentence to be more accurate.

179 The reviewer is correct; our phrasing was ambiguous. We will update this sentence following your suggestion.

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

182 Please see our reply to comment 2.4.3B.

185 That is correct. We will remove the sigma from the category thresholds.

188 Yes, both GBSA time series from the model and the observations are standardized. We will rephrase this paragraph 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 will rephrase 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 will rephrase the respective paragraphs to clarify this.

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 will clarify that, while the correlation itself is significant, this is of little use for a skillful prediction.

210 Yes. We will rephrase 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 will remove the redundant statement.

212 We thank the reviewer for this suggestion. We will improve the discussion on the effect of averaging periods on ACC.

262 We will improve the discussion on the lead-year dependence of persistence. 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 will remove 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 needs 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 will update the respective paragraph to include both viewpoints.

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. We will include this in the discussion section.

Minor 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.

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.

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