Response to Lisa Degenhardt (Reviewer #2)

We sincerely thank Lisa Degenhardt for her 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 comments and describe how we plan to address the issues raised.

Comments:

a) I have a few questions about the methods. It does seem alright, it is more the description of the method section, where I would like to see some more details for a better understanding:

Response: We appreciate the questions on the methodology and will address them in the following in order to improve the understanding.

L95 My comment doesn't fit here, but maybe in the introduction or discussion, but I realised it in this line. Why are you choosing exactly these predictors? I think it would be nice to have a bit more details why you choose those and with references and why T70 is used from September.

Response: We apologize for leaving out details on the reasoning behind the choice of November Z500 and September T70 as predictors. The choice of stratospheric temperature and extratropical geopotential height is based on the proposed links between the European winter climate and the QBO state, as well as the Rossby wave pattern. The choice of month and vertical level stems from a systematic lead-lag correlation test at different vertical levels with winter GBSA, which, for reasons of brevity, is not shown in the manuscript. T70 and Z500 emerged from this analysis as the two variable-height combinations with the highest correlation with GBSA. Choosing Z or T at an adjacent vertical level (e.g., T50 or Z400) would have been similarly reasonable, as they are also positively correlated with GBSA. However, we decided to go with those fields that exhibit the maximum correlation with GBSA. Similarly, T70 from August or October also yields a similar correlation pattern to September T70, but the absolute correlations are slightly lower, thus leaving September as the optimal month for predicting winter GBSA. We will include a more thorough explanation of our choice in the introduction section.

L101-Eq.1 I am a bit confused with this paragraph. And is np=nx?

Response: We apologize for the confusion, the x_n in Equation 1 should read x_p to fit the description in the text.

L117-120 I think this part is a bit confusingly written or at least I got confused. I believe it does explain what Fig. 1 is showing. Which I did understood, but maybe need some more clarity

Response: Again, we apologize for the confusion. The text indeed is supposed to explain the workflow shown in Fig. 1. We will rephrase this paragraph to describe the subselection process more clearly.

b) Chapter 2.5 How can you train with ERA5 data, but use the decadal data as predictor? And what are you training for?

Response: We are thankful for this question. The states of the predictors T70 and Z500 are always taken from ERA5 data, both for the "training" period of 1940-1959 and the "prediction" period of 1960-2019. From the decadal prediction system data, we only take the winter GBSA

predictions of the 64 individual members, as well as the large-scale fields of winter MSLP, Z500, and U200 analyzed in Section 3.2.1. The training aims to establish a set of gridpoints or areas in which the respective predictors are positively correlated with winter GBSA before starting the first actual prediction with the 1960 model run. As we have an additional 20 years available from ERA5 and observational GBSA records, we can use these 20 years prior to the start of the first model run to "train" our predictor-based selection method into knowing which gridpoints to take into account for the calculation of the predictor state, once the first model run in 1960 is started.

c) I think the Introduction but especially the Discussion is missing some external references. I added a personal note at the bottom, which is just a suggestion as I worked in a similar field. But in general, I would like to see a bit more referencing to other studies.

Response: We appreciate this comment and the suggestion of the study below. We agree that the introduction and discussion sections require a more thorough relation to other studies in the field. We will expand these sections with several relevant studies from the field of seasonal predictions of the European winter climate and put our findings into context by comparing the results of our study to those of other papers.

Major corrections:

L198 I thought initialisation is November (see abstract), which I thought is mostly around the 1st and I thought you are then predicting core winter (DJF), why now "end of November"?

Response: That is correct, the initialization of the model takes place at the 1st of November. However, the predictor field of November Z500 anomalies is only available at the end of the month. Hence, we can perform the subselection no earlier than the end of November. An alternative would be to use October Z500 instead which would available at or shortly after the initialization, but the correlation between October Z500 and winter GBSA is much lower than for November Z500 and would therefore not lead to any significant improvement of winter GBSA predictability. We will remove this statement in this section and add clarification to the methods section to avoid confusion here.

Fig. 6 & 7 I don't understand the difference between Fig. 6b, e & h and Fig. 7b, e & h. I thought the sub-selected ensemble is by the 25 closest members? So for me both would be the same, but they aren't.

Response: Figures 6 and 7 are based on different sets of 25 members. In Fig. 6, the 25 closest members are determined via the absolute difference between predicted winter GBSA in the respective members and observed predictor fields (September T70 and November Z500). In Fig. 7, the 25 closest members are determined via the absolute difference between predicted winter GBSA and actually observed winter GBSA. Fig. 7 shows which members would have been chosen if we already knew at the start of the model run how stormy the winter would turn out to be. We realize that our description of the perfect test lacks clarity and will thus rephrase the respective section.

L308 I understand what a "training period" is, but when I reached this part of the paper, I realised that I don't know why you needed a training period at all.

Response: We are thankful for raising this point. In order to determine the area-averaged state of the physical predictors in each year (T70, Z500), we need to specify the areas that we include in or exclude from the calculation of said state. We do this by correlating the predictor time series at every gridpoint with GBSA and only choosing those gridpoints that correlate significantly over

a time period from 1940 to the year before the start of the model run. As the hindcast runs start in 1960, we define the remaining time from 1940-1959 as our “training period”, which acts as a starting base to get a first idea of which regions to take into account for the computation of the predictor states. This is not a training period in the classical sense, like e.g. in machine learning, but more a required “lead time” to obtain an initial view of relevant regions for GBSA prediction. We will clarify the respective section in the manuscript to make this point clearer.

L315 I could have missed it in the paper, but where is this result coming from?

Response: This result is not explicitly shown in the paper. The idea of selecting those members that match the observed November Z500 was rather meant to serve as an outlook towards possible future studies that might make use of this technique. In our case, we did not find any skill gain in selecting the members this way, and therefore did not dedicate an entire section or figure to it. We realize that we should clearly differentiate between what we see as a potential outlook and what we already tested, and emphasize that in the discussion section accordingly.

Minor corrections/suggestions:

L123 you only have 2 predictors, so you can say “both” instead of “multiple”, right?

Response: That is correct, we will change “multiple” to “both”.

L134 I found this sentence a bit unnecessarily confusing written. I do understand it and it is right, but maybe it is easier to say: “ACC values of 1 indicate a perfect correlation, 0 no correlation, and -1 a perfect anticorrelation.”

Response: We thank the reviewer for this suggestion. We will rephrase the sentence to make it less confusing.

L152 Shouldn’t Fi be Oi?

Response: Here, Fi refers to the probability of occurrence obtained from the prediction system. Oi, which we use for the observed state, is described further down in the paragraph.

L170 What are the 17-80? Are these IDs of members? If yes, why not using the first 16?

Response: Yes, these are the sequential numbers of the ensemble members. For the first 16 members of the prediction system, only daily MSLP output is available, while the remaining 64 members (17 through 80) provide three-hourly MSLP output. To calculate storm activity from this output, a high temporal resolution is required and we thus decided to disregard the first 16 members. Doing so, we also keep the study consistent with the storm activity calculations in Krieger et al., 2022. We will drop the member ID numbers from the manuscript and clarify this in the methods section.

Fig 2&3 Is there a specific reason why the dots have a white inside? The difference between white and black makes the dots quite blurry on my screen. Maybe test to make them fully black. Also in Fig. 2, I believe there is a significant area in the positive bloop over eastern Russia, maybe you could increase the density of the dots to make at least one or two dots visible there?

Response: We thank the reviewer for making us aware of this graphical issue. While we ourselves do not see any white in the dotted significance patterns with our PDF reader, we will try and use a

different, denser stippling pattern to avoid such visual artifacts in the future. We would greatly appreciate further feedback on whether the issue persists with the denser stippling patterns.

L213/214 I would add here that the first sentence is for the measure correlations. Even the 25 members is right for all measures at the end. "The optimal sample size is found at 25 members per predictor for correlations (r = 0.64)."

Response: We agree with the reviewer and will reword the sentence.

L219 What is Z500,sep now? I thought T70 is used from September.

Response: That is correct, we apologize for this mistake. Here, it should just read Z500 without the "sep". We will change this line accordingly.

L221 Maybe add an "individually" to make clear that you now talking about the sensitivity of each predictor alone.

Response: We agree with this suggestion and will add "individually" to the sentence.

L265 Is "perfect test" and "perfect ensemble" here the same? I think I would stick with one, maybe perfect ensemble?!

Response: We thank the reviewer for pointing out the inconsistent wording. We intended to use perfect test for the procedure of choosing the closest members to the observed GBSA, and perfect ensemble for those members themselves. We realize that this might be confusing, and will rephrase the section to explain our definition of both terms, and to avoid switching between test and ensemble too frequently.

L281 I believe "stark" is supposed to be "strong"

Response: We agree and will change "stark" to "strong".

Personal Point:

I did a similar skill study to seasonal predictions of European (wind-)storms and what could improve their skill. There are some publications available about that, other seasonal predictability and their influencing factors like atmospheric drivers. It would be nice to see a bit more of these studies in either the Introduction or Discussion. No need to use mine, but as we are both looking for storm activity over Europe, I wanted to mention it at least: Degenhardt, L., Leckebusch, G.C. & Scaife, A.A. Large-scale circulation patterns and their influence on European winter windstorm predictions. Clim Dyn 60, 3597–3611 (2023). https://doi.org/10.1007/s00382-022-06455-2

Response: We thank the reviewer for pointing us to this study. We will include the study in the introduction and discussion, as it is very relevant for this research field and should not be left out.