Response to Reviewer #1

We, the authors, sincerely thank Reviewer #1 for their insightful and valuable comments and suggestions on our manuscript entitled *Improving seasonal predictions of German Bight storm activity*. The comments greatly helped to improve the manuscript and to remove unclarities.
In the following, we would like to address the issues raised by giving a point-by-point response and propose how we plan to incorporate the comments in our manuscript.

Comments:

1 The paper lacks some basic details and references on, among others, storm activity and its variability, as well as decadal prediction systems (see for example your own introduction in Krieger et al., 2022). I understand that you are trying to tell a different story than in Krieger et al. (2022), but you should not expect the reader to know your previous publication.

Response: We apologize for leaving out important details on storm activity and the decadal prediction system. We agree with the reviewer in seeing the need for a more independent introduction of these concepts for readers not familiar with Krieger et al. (2022) without re-telling the same storyline. We will improve the respective sections that introduce storm activity and the prediction system and add more background information.

2 There are several relevant publications from the MetOffice colleagues on seasonal predictions, for example Athanasiadis et al. (2017), Scaife et al. (2014), or Scaife et al. (2016). Please include and discuss some of them in your paper.

Response: We thank the reviewer for bringing up the aforementioned publications. We will enhance the introduction and discussion and include more studies on seasonal predictability.

3 Can you elaborate a bit why you specifically use T70 and Z500 as predictors for GBSA? Is there any reference for this choice?

Response: We are sorry for leaving the reasoning for this specific selection unclear. 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. The two predictors T70 and Z500 emerged from this analysis as the two variable-height combinations that correlated best with GBSA. It would have been similarly reasonable to use temperature or geopotential height from an adjacent height level (e.g., T50 or Z400), as the correlations between these fields and GBSA were only slightly lower. However, we decided to use those vertical levels that yielded the highest correlations. We will include a more thorough explanation of our choice in the introduction section.

4 How do you define which ensemble members have a GBSA that is closest to the first guess?

Response: We thank the reviewer for this question. We select the members based on the absolute difference between the GBSA value of the respective member and the state of each predictor. We choose the *n* members that exhibit the lowest absolute difference to the first guess. We will add this explanation to the description of our methodology.

5 Based on line 170, it looks like the ensemble has 80 members, but you only use 64 of them. Please explain this.

Response: The reviewer is correct here. The original ensemble consists of 80 members, 64 of which provide MSLP output at three-hourly resolution. The remaining 16 (numbered 1 through 16) only provide daily output. For this reason, we confine our analysis to the 64 members numbered 17 through 80, following Krieger et al., 2022. We will drop the member ID numbers in line 170 and clarify the choice of 64 members in the methodology section.

6 L268-269: Can you please elaborate on why the perfect test includes future information and why this cannot be used operationally?

Response: Absolutely. In the perfect test, we do not select 25 members based on their proximity to the first-guess GBSA obtained from the predictor fields in September and November, but based on their proximity to the observed winter GBSA from the period December-February. While the regular subselection allows us to make a first-guess prediction for winter at the end of November (i.e., once we know the state of November Z500), the perfect test selection needs observational information from the winter itself, which, when viewed from a pre-winter standpoint, lies in the future. This perfect selection can therefore only be carried out after the winter has passed, i.e., at the end of February. For operational purposes, where a forecast for the winter is required before the start of the winter, this is impractical. We will clarify the implications of the perfect test and its limits with regard to operational use in the respective section.

7 Outlook: How could your approach be used in future studies?

Response: We thank the reviewer for this question. One example for a future study based on this methodology would be to link the statistical connection between, for instance, T70 and winter GBSA to a dynamical prediction of T70 at lead times of multiple years. By establishing skillful predictions of the QBO, which governs the state of tropical T70, one or two years ahead, and then applying our method, more skillful predictions of winter GBSA one or two years into the future may become possible. In a broader context, any climate state or extreme that can be associated to physical predictors may be more skillfully predictable with our approach. While we limit our study to T70 and Z500, other atmospheric or oceanic variables may be useful as predictors for certain climate extremes, also possibly at even longer or shorter lead times. We will expand the outlook part of our discussion section to discuss these potential applications in greater detail.