Response to Reviewer #3

We sincerely thank Reviewer #3 for their constructive and insightful comments on our manuscript *Improving seasonal predictions 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.

Major comments:

1 The study should provide a bit more historical context of their subsampling idea in the introduction. Also, the Discussion section does not refer to any other studies.
Historical context: Conceptionally, the approach returns to an old idea of "analogue forecasting" of future states that can be constrained based on their similarity to meaningful predictors of the preceding observed state (cf. Lorenz 1969; Barnett et al. 1978). As demonstrated earlier for weather prediction (van den Dool 1994) or weather field reconstructions (Schenk & Zorita 2012), the skill will depend on various choices that Krieger et al. test here for a large-ensemble prediction system. In particular, the skill of the subsampling approach will also depend on the number of spatial degrees of freedom of predictor and target variables.

Response: We appreciate the comments and input on the historical context of subsampling and analogue forecasting. We will expand the introduction to properly present the idea of subsampling and its origin. We will also discuss more studies in the discussion section.

2 Statistical context: In summary, the authors make largely appropriate efforts to provide robust statistical significance testing for both, identification of predictors as well as the resulting prediction skill. While the authors consider temporal autocorrelation using block-bootstrapping, I have some concerns that their locally significant results could be randomly significant owing to the potentially very low number of spatial degrees of freedom and hence very large spatial autocorrelation of fields like T70 and Z500. A quick field significance test is suggested below (cf. Livezey & Chen, 1983; Wilks, 2006) which would otherwise not change the results of the final prediction skill in this study. In the historical context, this study is even more remarkable as a quite good skill is achieved with using spatially rather homogeneous predictors.

Response: We appreciate the feedback on the statistical significance testing. We see the need to include a global significance test that checks the proportion of local tests which were erroneously considered significant. We decided to perform a global test that controls the false discovery rate (FDR) as it is considered one of the most powerful tools to check global field significance (e.g. in Wilks, 2006). We find that by controlling for the FDR at a level of 0.05, 21% of gridpoints for T70 and 82% for Z500 that were previously considered significant (7% and 13% of all gridpoints, respectively) are now insignificant. However, those areas that were deemed most relevant for the prediction of GBSA, namely the tropics for T70 and the extratropical Rossby wave train for Z500 are still significant globally. Thus, we decided to keep calculating the predictors from locally significant gridpoints. We will add this test to the methodology section and refer to it wherever necessary. We updated Figures 2 and 3 in the manuscript (see Figs. R1 and R2 below) to now display both locally-only and locally-and-globally significant gridpoints.

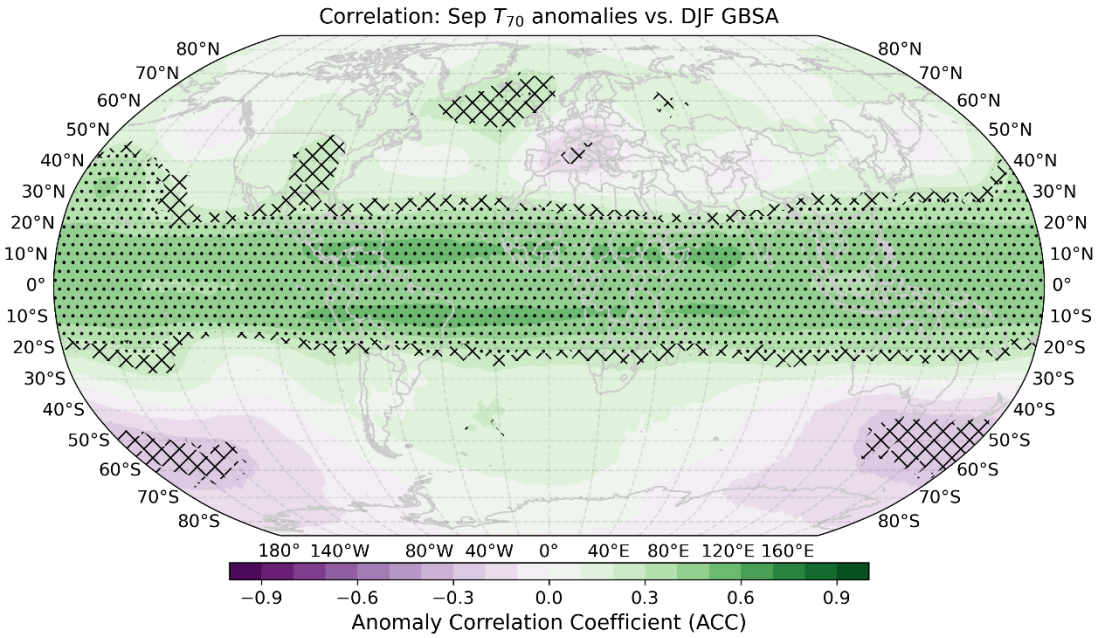

Fig. R1 Gridpoint-wise correlation coefficients between global T70 anomalies in ERA5 and observed winter (DJF) German Bight storm activity. Period 1940–2017 for temperature anomalies, 1940/41–2017/18 for storm activity. Hatching indicates local statistical significance ($p \leq 0.05$) determined through 1000-fold bootstrapping. Stippling indicates additional global (field) significance by controlling for the FDR at a level of 0.05.

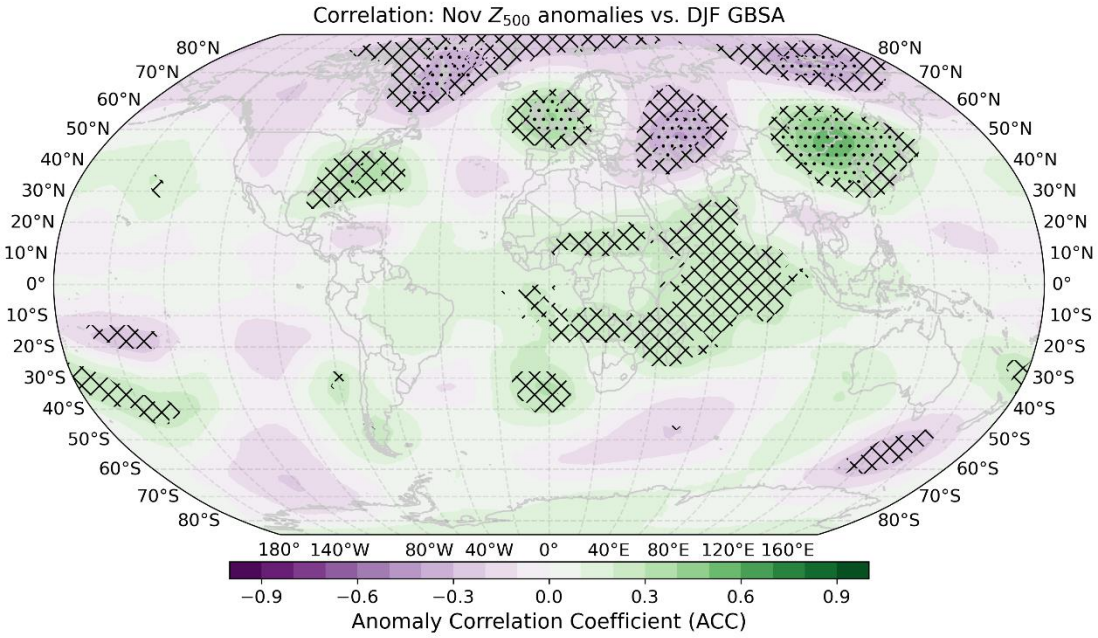

Fig. R2 As Fig. R1, but for Z500 instead.

Minor comments:

L28 You directly jump here to the concept of using specific physical predictors to aid in reducing the spread of model predictions and increase prediction skill. You later use this idea in this study to improve subsampling of certain ensemble members that are most similar to an initial predictor state. It might be worthwhile to briefly mention here some historical context mentioned above that this idea is very similar to analogue forecasting attempts already in 1970s to predict future weather (Lorenz 1969) or short-term climate fluctuations (Barnett et al. 1978) based on analogues that proceed from the present state to estimate future states. Also, the idea to predict unknown full field states based on incomplete low-order predictors via analogues was successfully applied in reconstructions (e.g., Schenk & Zorita 2012). Interestingly, in that study the skill improvement was tested in a very similar way as done here regarding the dependency on the number of predictors, use of multivariate predictors and benchmarking with an idealized model-dependent prediction skill.

Response: We greatly appreciate the input on the close relation between predictor-based predictions and analogue forecasting. As mentioned in our response to the 1st major comment, we will expand the introduction to properly introduce analogue forecasting and the idea of subsampling.

L97-99 It is a bit unclear how statistical significance is derived here. Fields of T70 and Z500 tend to have a very large spatial autocorrelation (low spatial degree of freedom). It is quite likely that far more than 5% with easily up to more than 20% of locally significant grid cells could be randomly significant globally. To be sure, you could test the global significance by estimating 1000x correlations from bootstrapping of these fields and its correlation with ERA5 fields to evaluate how many grid cells are randomly significant (Wilks 2006). If your T70 and Z500 predictor fields vs. ERA5 fields yield more locally significant grid cells than the randomly significant correlations, you can claim to use globally or regionally significant predictor fields. It should be noted that even if this test fails, predictor fields may still provide predictive skill if the locally significant areas are physically meaningful (e.g., linked to Rossby waves, NAO etc.). It looks like this is the case in your study. You could add a plot of the global field with random correlations and locally significant areas and provide the test quantity of "overall % of n significant grids x 100 / N total grids" which provides more context to Fig. 2 and 3 in the main text. Based on the constraints (line 113), global in this study could also be regionally 30-90°N here. Based on Fig. 2, this area may not be overall significant but appears to provide coherent regions of locally significant correlations. Fig. 3 might show a higher fraction of significant correlations due to wave propagation (N=4 areas) plus Arctic, total N=5 areas of predictive skill but perhaps only N=2 independent predictors?

Response: We agree with the reviewer that the manuscript is missing a global significance test. We will add a field test based which controls for the FDR at a level of 0.05, and add the results to the correlation maps in Figs. 2 and 3, as well as the corresponding text.

L102-106 Here, the randomly significant issue becomes obvious from using a gridpoint-wise testing with a 95% local confidence level. The selection process is correct locally but may not provide regionally or globally significant field results regarding the full fields with low spatial degrees of freedom. I would not change this procedure here but just add a small test described above to include a sentence whether you use regionally or globally significant predictor fields or only locally significant results. Despite using fields, your predictor would then be local.

Response: We appreciate this insight. As mentioned in our response to the 2nd major comment, we will include global significance testing and compare the global significant fields to the locally

significant gridpoints. We will keep the predictor local, but explicitly add this to our methodology description.

L186-191 Perhaps mention here that the four significant areas in the northern extra-tropics represent the Rossby wave propagation in addition to teleconnections (Arctic Oscillation-like?) with the Arctic versus (sub-)tropical areas of Sahel and Indian Ocean. It is quite nice to see that these physically meaningful areas show up also statistically.

Response: We thank the reviewer for suggesting to add the Rossby wave pattern and the AO here. We expanded the paragraph to include these large-scale patterns.

L223-224 "purely coincidental". Not a coincidence at all. There is a direct relationship between the correlation coefficient and the RMSE, i.e. when standardized observations and predictions are used as RMSE inputs (hence bias = 0). This means that the RMSE is a measure of the unexplained variation, which is inversely proportional to the explained variation, which is the square of the correlation coefficient (as can be seen in Fig. 4). Therefore, it is not purely coincidental that for both predictors the optimal sample sizes for RMSE and correlation are equal, but a consequence of the mathematical relationship between these two statistics. Please replace the sentence with the opposite statement.

Response: We apologize for the incorrect statement. The RMSE and correlation coefficient are indeed related. We will correct the mistake and replace the sentence, also referring to a description of this relationship in Barnston (1992).

Figure 5 Very good illustration and impressive result.

Response: We thank the reviewer.

L265-267 I generally like that test regarding the question what the best selection of the 25 members would be knowing the observed state. Here, you could've gone even further by evaluating the single best member per year out of 64 members relative to ERA5 for 1960-2017/18. That would be a prediction-system-specific optimum of the ensemble initialised in late autumn for DJF which could be compared to your "almost perfect test".

Response: We thank the reviewer for the suggestion to test the large-scale prediction skill of the single best member per year. The results are shown in Fig. R3 below. For the single best member, we identify an increase in ACC for MSLP in west of Norway and west of Iberia, as well as a slight decrease over the North Sea. For Z500, the ACC is increased northeast of Iceland, but reduced over the North Sea and the majority of Europe. For U200, we find a slight increase over the North Sea, but a stronger decrease from the Central Atlantic to the Alps, and also from Greenland into northern Scandinavia. Compared to Fig. 7 in the manuscript (i.e, the perfect test for 25 members), the ACC change is overall more negative, but similar features emerge, such as the increase of U200 over the North Sea, and the tripole-like structure of ACC changes for Z500 and MSLP. We speculate that a single member might not be sufficient to predict the same large-scale patterns in the same region correctly every year, which could then lead to lower correlation gains overall than for a 25-member ensemble mean.

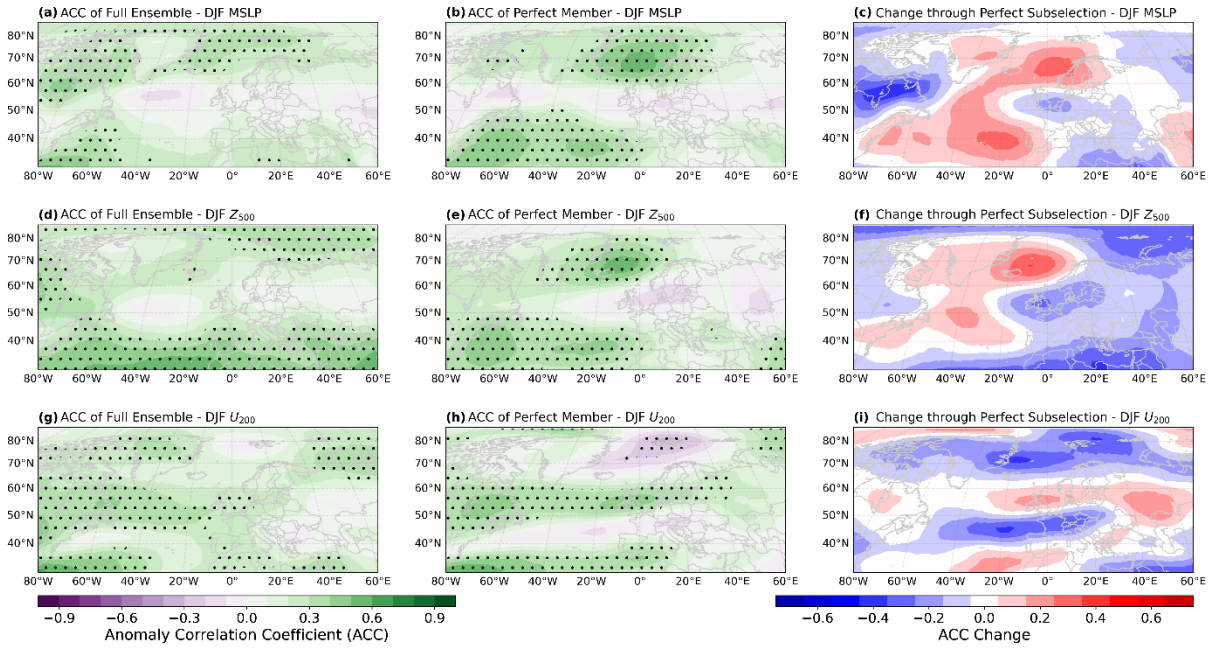

Fig. R3 Anomaly correlation coefficients (ACC) for ensemble mean predictions of the full 64-member ensemble (left column), the 1-perfect-member-subselection (middle column), and the change in ACC between the full ensemble and the perfect member (right column) for winter-mean (DJF) MSLP anomalies (first row), 500 hPa geopotential height anomalies (Z500, second row), and 200 hPa zonal wind anomalies (U200, third row). Winter-mean anomalies are calculated by averaging monthly anomalies from December, January, and February. Period 1960/61–2017/18. Stippling indicates statistical significance ($p \leq 0.05$) determined through 1000-fold bootstrapping

L282 Agree. I guess here you could mention the potential for machine learning methods.

Response: We thank the reviewer for this suggestion and will add a remark possible involvement of machine learning techniques.

Figure 8 How much do these composites differ from a first-year composite? Could the strong Antarctic difference be caused by a long-term trend in the model runs over time rather than highlighting differences from the composites?

Response: We appreciate this comment. The first-year composite (see Fig. R4 below) displays a similar behavior as the composite of all lead years (Fig. 8 in the manuscript), with a strong signal in the Southern Ocean and Antarctic regions and composite close to 0 K in the tropics, this time even slightly negative instead of slightly positive.

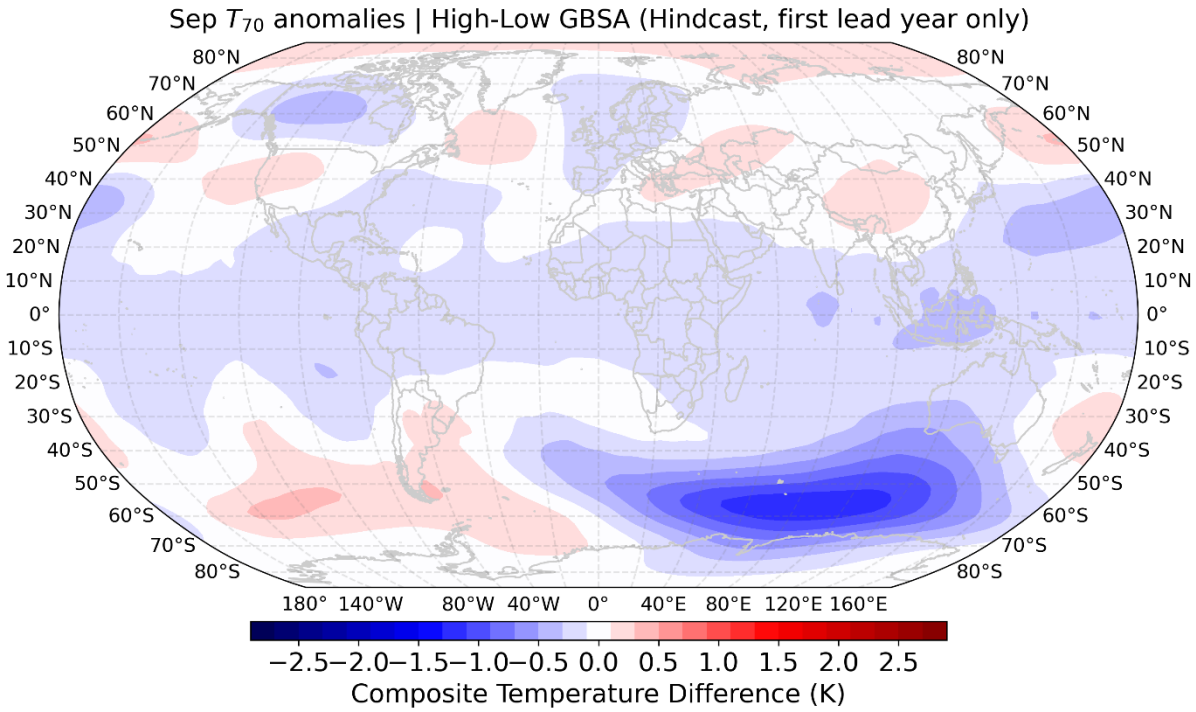

Fig. R2 Composite mean T70 of 100 model years with the highest subsequent DJF GBSA minus composite mean T70 of 100 model years with the lowest subsequent DJF GBSA in MPI-ESM-LR decadal hindcast runs. Data are taken from all initializations, all members, but only the first lead year after the initialization.

L299 The whole discussion section does not make any attempts to put results into context with other seasonal prediction studies (e.g., Kruschke et al. 2014; 2016 and many others). I see that some relevant studies were briefly discussed in Krieger et al. (2022) but not here. I suggest adding a paragraph or several sentences throughout chapter 4 where similarities and differences to other studies are discussed.

Response: We thank the reviewer for this comment and apologize for not mentioning more studies from the research field in the discussion section. We will expand both the introduction and discussion with relevant studies and discuss our results by putting them into context through comparison with the findings of similar studies.

L300-303 Although you're using a decadal prediction system, does that really differ from using a seasonal prediction system in your specific case? The initialisation in November to predict DJF is pretty much what a seasonal prediction system would do.

Response: We thank the reviewer for voicing this concern. In our case, that is, using the MPI-ESM prediction systems, the decadal and seasonal systems differ mostly in ensemble size and resolution. The decadal prediction system consists of 80 members, 64 of which provide 3-hourly resolution, while the seasonal system only contains 30 members at 6-hourly resolution. Secondly, the decadal system runs at low spatial resolution (T63 grid, 1.8 degree grid spacing), while the seasonal system runs at high spatial resolution (T127 grid, 0.9 degree grid spacing). The higher resolution in the seasonal system is also present in the number of vertical layers, which results in the seasonal system being able to maintain a QBO signal, whereas the lower vertical resolution in the decadal system is unable to properly maintain the QBO in the stratosphere. There are also minor differences in the nudging data of the respective assimilation runs, where in the seasonal system ERA5 is used for all hindcasts, while the decadal system uses ERA-Interim until 2015 and ERA5 from 2016 onwards. Other than that, the two systems are very similar.

L305 Most likely because the annual GBSA is dominated by the variation in winter (high correlation of high annual percentiles with high winter percentiles, i.e. same tail values)?

Response: This is likely the case, yes. Winter contributes most to the annual storm activity metric as the highest wind speeds occur during the winter months and, therefore, the upper percentiles of winter and the entire year show a high correlation. We will add an explanatory sentence to the discussion.

L308 "by two decades"

Response: We thank the reviewer for this correction.

L321 Regarding NAO, perhaps AO would be more appropriate as mentioned above?

Response: We appreciate this suggestion. We correlated the tropical September T70 predictor time series and the DJF Arctic Oscillation Index for the time period 1960/61-2017/18 and found a correlation coefficient of 0.11. To put this into perspective, the correlation between T70 and the DJF NAO index for the same time period is 0.27, while the correlation between T70 and DJF GBSA is 0.49. We will mention AO in this section alongside the NAO, as we already added the AO to the analysis of the correlation maps and see the need to refer back to it here.

References:

Barnston, A. G. (1992): Correspondence among the Correlation, RMSE, and Heidke Forecast Verification Measures; Refinement of the Heidke Score. *Weather and* Forecasting, 7 (4). DOI: 10.1175/1520-0434(1992)007<0699:CATCRA>2.0.CO;2