Reviewer 3

General Comments:

This manuscript by Mahmood et al. (egusphere-2022-98) explores an alternate way of decadal to multidecadal predictions by subsampling individual CMIP6 historical simulations that better matches observed SST at a given time (start dates) and by tracking the trajectories of the subsampled simulations for the next two decades (analogue method hereafter). The authors show that the added value of the analogue method over the uninitialized simulations is comparable to that of initialized decadal prediction simulation (DCPP). The manuscript is overall well organized and written. I have enjoyed reading the manuscript. I also acknowledge the thorough effort to test the sensitivity of the results to subsampling criteria. I think the manuscript has potential to draw attention from climate research community, as the proposed method can leverage existing simulations in the application of the prediction of climate, instead of running computationally expensive decadal prediction simulations. However, before I recommend accepting for publication, I have a couple of specific comments that I wish the authors further demonstrate or explain along with some minor suggestions.

Author Response: We thank the reviewer for providing these supportive and constructive comments, which we believe have helped to clarify some important aspects of our study.

Specific Comments:

1) The authors shows that the analogue method exhibits high skill in the Pacific Ocean, even higher than skill in the subpolar North Atlantic, which even lasts for FY11-20 . Such a long memory in the Pacific is surprising and in stark contrast to the current understanding that predictability in the Pacific Ocean is low on decadal timescales while very high in the subpolar North Atlantic. The low predictability in initialized decadal predictions may be related to initialization shock/drift, as the authors also discuss in the manuscript. However, the low predictability in the Pacific Ocean is also pervasive in “perfect model” experiments (e.g., Collins 2002; Pohlmann et al. 2004), which does not suffer from initialization shock/drift. Why the author’s analogue method shows such superior skill in the Pacific Ocean? Isn’t this high skill possibly related to the forced signal that is not completely removed by the method the authors used (Smith et al. 2019)? One way to verify this is to perform a bootstrapping method for the statistical test, rather than Student’s t-test. If the ACC from Best30 is found outside of the (eg., 2.5 to 97.5 percentile) distribution of the ACCs from randomly sampled 30 members, assuming that the uninitialized ensemble mean used in Smith et al.’s method is the total 212-member ensemble mean, the authors can say more confidently that the high Pacific skill is indeed not from the forced signal.

Author Response: We thank the reviewer for this comment, which is also echoed by some comments from the other reviewers.

We have implemented the bootstrapping method as suggested by the reviewer, and find the ACC of the Best30 ensemble in the Pacific is outside of the 2.5 to 97.5 percentile distribution of randomly sampled 30 members (Figure R4). In fact the areas with significant skill or significant skill differences are very similar compared to Figure 2 in the manuscript, suggesting our results are robust and not just an artifact from incomplete removal of the forced signal.

In addition, we have also compared global trend maps and regional time series, and these give

some indication that the long-term changes in Best30 are (slightly) more similar to observations than long-term changes in the full ensemble in areas that exhibit added skill. This suggests that the added skill may, at least in part, also arise from a more realistic estimate of long-term changes/trends as a consequence of initialisation. We have extended the discussion to also cover this aspect and added a new supplementary Figure S2. The added text reads as: "*The added skill in the constrained projections likely comes from an improved representation of long-term changes in response to forcing (as also found for decadal predictions, e.g. Doblas-Reyes et al., 2013), and also the representation of decadal-scale variations. Inspection of regional average time series in regions with added skill (e.g. in the Pacific, eastern Asia or the North Atlantic) indicates warming trends more similar to the observations in the constrained ensemble compared to the full CMIP6 ensemble in particular in the early parts of the hindcast period. These time series also show that the constrained ensemble better captures the observed variations in the warming rate, likely in relation to decadal-scale climate variability.*"

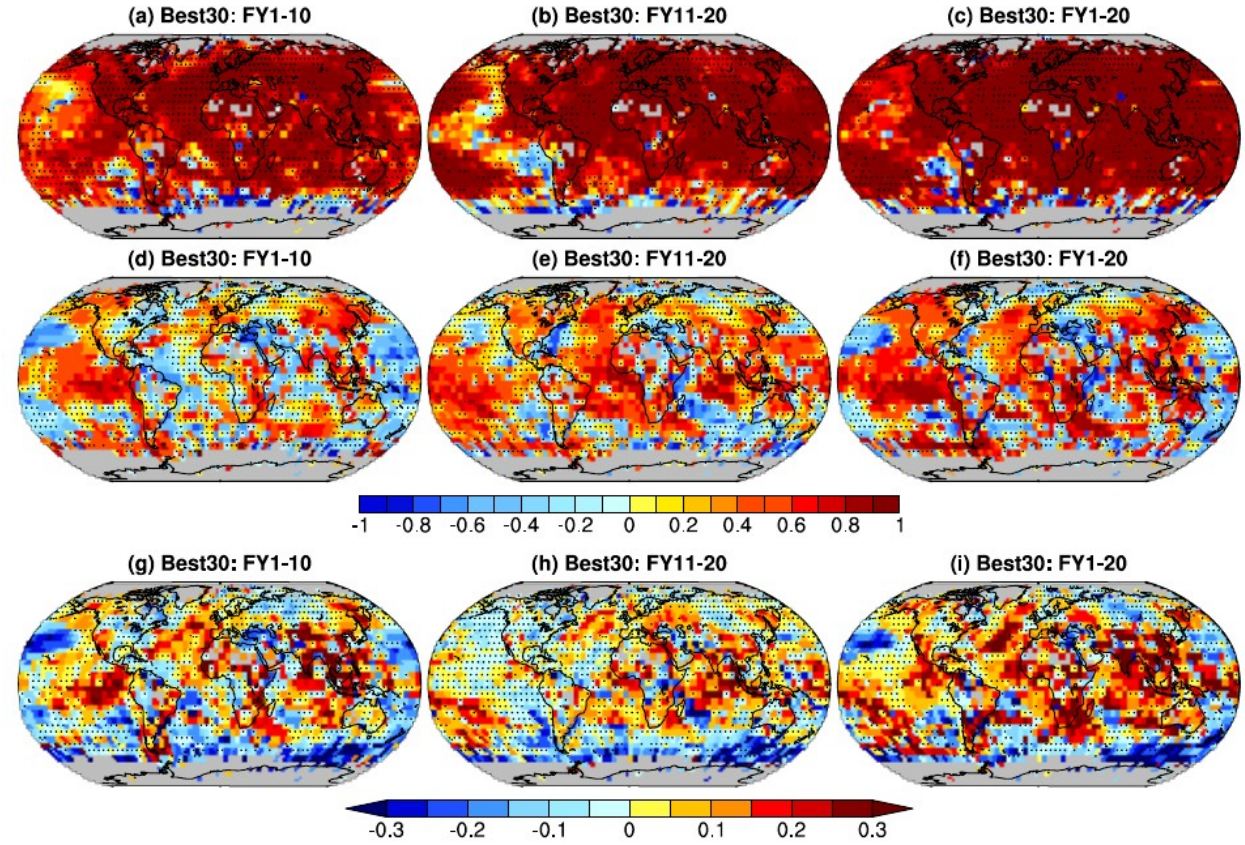

Figure R4: ACC (a-c), residual correlation (d-f) and RPSS (g-i) that are the same as shown Figure 2 in the manuscript but the stipplings here represents values that lie within 2.5th and 97.5th percentile range of the corresponding 1000 distributions obtained by by a bootstrapping method randomly selecting 30 members at each start date.

2) If that is the case, why the skill is so high in the Pacific? Since this would the most important finding of the study, in my opinion, as it is in contrast to the current understanding, I recommend that the authors further demonstrate the reasons for the high Pacific skill.

Author Response: Please see our response to the previous comment. A comparison of long-term changes e.g. in these Pacific regions indicates that the constrained ensemble seems to show more realistic trends in particular during the earlier part of the hindcast period, but also improved representation of decadal-scale variabiliy. These different aspects are now discussed in the text and we have added a new supplementary Figure S2.

3) The Atlantic skill is low for FY1-10, but picks up for FY11-20 (Fig. 2d-e). Why is this the case? I think the low skill for FY1-10 is because Best30 is dominated by the correlations in the Pacific and as demonstrated in the regional SST constraints, but it is hard to understand why there is an rebound in ACC skill.

Author Response: This is an interesting feature, which we are not able to fully explain at this point. Please note that in the analysis of pentadal forecast periods (new supplementary Figure S3) we do find added skill in the Northeast Atlantic during years 1-5 (similar to DCPP), which disappears however in the second pentad.

4) The authors introduce several statistical methods in section 2, without a description, just by referring to citations. I recommend adding a brief description for each method.

Author Response: We have added brief explanations for the different methods in Section 2.

Technical corrections:

l. 43: Remove “in” after phasing.

Author Response: We have reworded “phasing in” with “aligning the phases of”.

l. 88: …anomalies “relative to” the reference climatological period…

Author Response: Thank you, we reworded as suggested.