Comment on egusphere-2022-98
Anonymous Referee #2

Referee comment on "Constraining low-frequency variability in climate projections to predict climate on decadal to multi-decadal time scales - a 'poor-man' initialized prediction system" by Rashed Mahmood et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-98-RC2, 2022

The paper presents an interesting approach to providing climate predictions based on constraining non-initialised climate projections with observed climate variability. Only those ensemble members of climate projections that show the largest agreement with observed SST anomalies in years prior to the forecast start date are used to construct climate predictions. Instead of a full initialisation with the observed state as normally done in climate predictions (e.g. for seasonal and decadal prediction), a simplified ('poor-man') approach of aligning the phase of theses of simulated and observed SST variability is used. After applying the approach to hindcasts from 1961 onwards, the forecast quality of the predictions is evaluated and compared to both fully initialised decadal predictions and unconstrained climate projections. It was concluded that the constrained ensemble provides skillful predictions of near-surface temperature, sea-level pressure and precipitation in large areas of the globe. During the first decade of predictions the skill of the poor-man predictions is comparable to the initialised decadal predictions. Significant added value from the constrained approach was found in the second decade for which initialised predictions are not available. Sensitivities to certain choices like the past period and geographical regions of the constraint, ensemble size and skill metric, have been discussed.

I think the approach explored in this study is very interesting and certainly deserves to be published. In particular, I agree with the authors that the potential benefits of their approach over both initialised decadal predictions and unconstrained projections for providing seamless climate information could be big and important. However, I cannot recommend publication of the paper in its current form because it lacks several critical aspects that are discussed below.

Major comment
In my view the manuscript suffers substantially from the poor demonstration of the results. While the motivation and the methodological approach are nicely laid out, the analysis of the results and their graphical presentation do not provide enough evidence to the reader to be convinced of the benefits that the new prediction system might bring. With “enough evidence” I don’t mean the quantity of analysis or plots but rather the opposite: the authors have taken the approach to include into the manuscript and the supplementary material almost every possible plot one can think of for the quantities they have analysed. However and unfortunately, the large number of plots does not provide an equally large amount of useful information. I would suggest to critically review all plots and only show those which really help support the claims you are trying to make. It is the responsibility of the authors to make a meaningful selection of the diagnostics that help tell the story you wish to convey and should go into a publication. This critical selection should not be left to the reader alone. I have the following specific recommendations:

Fig 1: I think this could be cut short without loss of information by only showing one start date as a demonstrator and carefully describing the methodology in the text and figure caption, as already done.

Fig 2:

- I don’t find showing means over 10 or 20 years are helpful in the prediction context. The window is too long to provide useful information. It would be better to split the windows into smaller ones to identify those time ranges where the approach can improve either decadal predictions of projections. For example, if the added value over non-initialised projections kicks in after 10 years, it would be most interesting to know when this happens – it is just immediately after the 10 years or more towards the end of the 20-year period? Averages over 10 years smear out the impact, and means over 20 years can potentially even be misleading by implying the skill comes from the later years when most likely it is coming from the earlier years. I would suggest looking at 1-5, 6-10, 11-15 and 16-20 years. Or, if the results reveal interesting insight, even for finer forecast ranges. This recommendation applies to almost all plots in the paper and supplementary material.
- Please also show ACC of the unconstrained projections after 10 years to provide a reference to which to compare to. Fig 3 for SLP shows differences which is helpful but Fig 2 for surface temperature does not.
- It would also be interesting to show how a similarly constrained decadal prediction ensemble would perform, that is sub-sampling those ensemble members from DCPP that most closely resemble the past SST observation after e.g. forecast year 1. That would of course imply that the predictions are only useable after applying the constrain (e.g. after 1 year) but for the longer time scales this could still be useful.
- What is the reference forecast used in the RPSS computation? Please add this information in the figure caption.

Since showing too many global maps is not sustainable, I would recommend to condense the critical information either into 2D plots or bar charts (similar to what has already been
done in the Supplement but for finer forecast ranges). These could be good options for the various sensitivity studies. For example, global or key regional scores could be plotted in a 2D plot as a function of forecast year and selection period to replace Fig S2 etc. Such a condensation would make space to show a direct comparison (or differences) with the performance of the unconstrained ensemble or the decadal predictions.

I find some of the results are a bit over-interpreted and should be re-worded slightly more carefully. For example, on line 173 you say that added value is found over similar regions across different forecast times providing confidence in the robustness. However, the plots these lines refer to (Fig 2h-j) indicate for example some inconsistencies in the North Atlantic and the tropical East Pacific. Or for SLP in Fig 3, the highly skilful subtropical North Atlantic for FY11-20 (Fig 3b) is not showing during the first 10 years (Fig 3a). Why is this? Around lines 180, mention the problematic issues over the Indian Ocean.

The result that the constrained projections can outperform the initialised predictions is very interesting. I feel it would require some more discussion as to what the mechanisms are that can lead to this perhaps surprising skill. Discussing potential explanations would make the paper much stronger than simply describing it.

**Supplementary Information:**

It is not clear which variables have been analysed in Fig S1-S3 and S5-S6.

**Minor comments:**

Fig 4 caption: unclear what exactly is meant by added skill – please specify.

Switching between the ACC and the residual correlations introduces some inconsistencies in the manuscript. For the purpose of this paper, it might be sufficient to only show ACC. Fig 5 and 6 could go in the Supplement.

Why are the atmospheric fields computed on a 5x5 degree grid and not on a 3x3 degree grid as the SSTs?

Section 3.1: The text could be improved by introducing more paragraphs and reducing the use of brackets ( ).
Sensitivity to temporal averaging of SST anomalies, around lines 206-209: emphasise this finding more as very interesting.

The cited reference of Menary et al (2021) sounds very relevant – could you expand on your discussion of this paper in Section 4?

Sensitivity to ensemble members: Why have you stopped sub-sampling at 50 members? It would be interesting to see the convergence for the full ensemble. Would it be possible to show a plot showing this convergence for perhaps a global quantity? Is there an optimal ensemble size?

Figure 7: show comparison with global pattern (repeat from Fig 1). Also show comparison with unconstrained and decadal prediction. Makes interpretation of these results easier.

Fig 8: Can you speculate as to why the constraint over the North Atlantic makes the forecasts worse? Fig 8b is not needed.

*Supplementary Information:*

Figures S8 – S10: instead of repeating maps very similar to Fig 2—4, it would perhaps be more informative to show differences to these other maps.