Reviewer 2:

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 unconstraint climate projections. It was concluded that the constrained ensemble provides skilful predictions of near-surface temperature, lea-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.

Author Response: We thank the reviewer for the constructive comments that helped us to further clarify some aspects of the manuscript.

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:

Author Response: We thank the reviewer for pointing out this shortcoming that made it difficult to follow the clear line of the story in our submitted manuscript. In our revision we have critically considered the relevance of the different plots, and moved in particular those map plots related to sensitivity tests to the supplementary information. We have in their place included a new plot (new Figure 5) that summarizes key information of those sensitivity tests. Now the main text

figures are reduced to those more relevant to presenting the constraining method, and we hope the reviewer agrees that this improves the readability of the paper.

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.

Author Response: We have carefully considered the suggestion by the reviewer, but feel that it is useful to keep showing several start dates to also illustrate the point that the selection differs for each (start) year. When presenting the work on different occasions we have noticed some audiences misunderstood the method by missing the important information that another sub-ensemble is selected in each (start) year. We therefore think that showing several start dates in the figure illustrates the method more completely, despite making the figure a little more complex.

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.

Author Response: Please note that our constraining approach presented here is targeted at aligning low-frequency variability between the simulations and the observations, with the explicit goal to refine near-term climate information on decadal to multi-decadal time scales. We therefore prefer to keep the presentation of results with a focus on the different 10-year and 20-year periods, which are representative for near-term climate change as opposed to predicting inter-annual variations.

Following the suggestion by the reviewer, however, we have also added a new Supplementary Figure S3 which shows the results for the different 5-year periods of years 1-5, 6-10, 11-15 and 16-20. The skill patterns for these shorter periods are largely consistent with those of the decadal and multi-decadal averages. We have added a paragraph to discuss the skill for these pentadal forecast periods.

- 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.

Author Response: We added a Supplementary Figure to show ACC of the unconstrained projections (new Figure S1).

- 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.

Author Response: We agree that applying similar constraints also to the initialised decadal predictions is a very interesting and promising perspective. However, we don't see how this can be implemented in a consistent way to constraining the projections in this same paper (using different selection periods, which would introduce inconsistencies), and therefore suggest that such constraint of decadal predictions based on their agreement with (early) observations should be the topic of future research.

- What is the reference forecast used in the RPSS computation? Please add this information in the figure caption.

Author Response: The reference forecast for RPSS is the full, unconstrained, CMIP6 ensemble, which has been mentioned explicitly 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.

Author Response: Thanks for pointing this out, and we agree with your criticism. In the revised manuscript we now have moved all map plots related to the sensitivity tests (i.e. previous Figures 5, 6, 7) to the Supplementary information, and summarise some global information from these figures in a new (bar plot) figure (new Figure. 5), as suggested by the reviewer.

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.

Author Response: In our revision we reworded the sentence in line 173 that claimed robustness from similar regions with added value across different forecast times as follows: *"While many regions that exhibit added value from the constraint show up for the different forecast times shown, in other regions such as large parts of the Atlantic Ocean or the tropical Indian Ocean positive residual correlations emerge only in the second decade of the hindcasts."*

We also added a sentence mentioning the issues over the Indian Ocean after line 180: "*Some added skill also emerges only in the second forecast decade e.g. over parts of the subtropical Atlantic and the Indian Ocean (noting however that ACC over the Indian Ocean remains negative for all forecast periods shown).*"

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.

Author Response: We agree that it is intriguing to understand what leads to the higher skill in the constrained projections, as also noted by reviewer 1. Inspection of regional time series in regions with added skill shows that the constrained ensemble shows improved representation of both long-term trends and multi-decadal variability. This may be partly related to the constrained ensemble better representing the forced climate response, but also to a preferred selection of some models with a better representation of either long-term changes and variability. Another important point is that the constrained ensemble members all represent undisturbed model attractors under transient forcing, whereas the initialised decadal predictions suffer from initialisation shocks and the drifts inherent to initialization. We have added some discussion throughout the text, but think a deeper investigation of processes is beyond the scope of this paper as it would require retrieving a range of different variables from the >200 ensemble members.

Supplementary Information:

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

Author Response: We specified in the captions of Figs S1-S3 and S6 that the results are for near-surface temperature. Fig. S5 specified “same as Fig. 2”.

Minor comments:

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

Author Response: We understand this comment refers to Supplementary Figure S4 (previous number) and the results from that figure are now part of new Figure 5 in which we have clarified in the caption that added skill was “measured as residual correlation and RPSS against the unconstrained CMIP6 ensemble as reference”

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.

Author Response: To avoid this confusion of using different measures to identify added skill, we now show residual correlations for all variables (Figures 2, 3, 4).

Figures 5,6,7 are now moved to supplementary material, and their key conclusions summarised in the (new) Figure 5.

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

Author Response: We followed recommendations for decadal predictions (Goddard et al., 2013), with the rationale that remapping atmospheric fields to a coarser grid removed small-scale noise in the skill analysis. For the constraint based on SST data we give preference to a

slightly finer common grid (within what is possible given some models still have a relatively low resolution), to also capture effects of finer scale variability patterns.

Section 3.1: The text could be improved by introducing more paragraphs and reducing the use of brackets ().

Author Response: Thanks for pointing this out. We have slightly rewritten parts of Section 3.1, to split up overly long paragraphs and reduced brackets, to improve readability.

Sensitivity to temporal averaging of SST anomalies, around lines 206-209: emphasise this finding more as very interesting.

Author Response: We added a sentence to emphasise this point of shorter averaging periods being favourable for inter-annual predictions: "*This is plausible as shorter averaging periods will emphasise the signals related to inter-annual variability in the member selections, whereas longer averaging periods will emphasise signals related to lower frequency variability relevant for predicting variations on decadal time scales.*"

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

Author Response: We have expanded the discussion of the Menary (2021) reference, which we agree is very relevant in the context of our study.

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?

Author Response: As discussed in the paper, a larger ensemble size of sub-selected projection members means to include members that are less similar at initialisation, and therefore may deteriorate the skill. We therefore do not expect a "convergence" of the skill as in e.g. in DCPP initialised decadal predictions. The optimal ensemble size (similar as the other sensitivities we discuss) will depend on the region and variable of interest. We tried to highlight this in our discussion of the different sensitivities.

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.

Author Response: The maps in (previous) Figure 7 (now supplementary figure S4) are directly comparable to the maps based on global constraints in Figure 2. To avoid redundancy of results between figures we prefer not to repeat the "default" setting results for all of the figures showing sensitivity tests. However, in the revised manuscript we now included a new overview figure (new Figure 5), which compares the global areas of added skill for all settings, and this figure also includes the bars for the "default" shown in Figure 2.

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

Author Response: We also checked the skill/added skill for GMST over the entire hindcast period (i.e. not just the hiatus period shown here), and e.g. the residual correlation based on the

North Atlantic constraint is negative, meaning this constraint does not improve the GMST predictions over the full ensemble (whereas residual correlations for the global and Pacific constraint are positive). We prefer not to speculate about reasons in the manuscript, but added the information about the lack of added skill and that this suggests the North Atlantic not controlling variations in GMST:

"No reduced warming rate is found for the Best30 ensemble constrained based on North Atlantic SSTs, suggesting that the North Atlantic did not contribute to this early 2000s global warming slowdown. Note that, also considering the entire hindcast period, the North Atlantic constraint does not improve GMST predictions (indicated e.g. by negative residual correlations) compared to the full unconstrained ensemble. This suggests that, at least based on the models used, the North Atlantic does not seem to control variations in global mean temperatures".

Fig 8b has been removed, and we included the slope values it previously showed in the legend.

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.

Author Response: The purpose of these figures is to show that the main conclusions hold regardless of reference dataset used. We think that this message is better delivered by repeating the figures, as this allows us to confirm that the broad global patterns and features are robust. A difference map might be more useful to measure the magnitude of the uncertainties, but this was not the intended purpose of this analysis.