Reviewer 1:

In my opinion this is a very well written and clear paper which presents an interesting new way of combining a large multi-model dataset in order to constrain future projections. As such I have very few comments to make and I believe that this paper should be published with only a few corrections.

The results are impressive and warrant publication. However the one area I would like to see more explanation on is the cause of the skill found. In the summary (line 300) you claim that "these results indicate that there is significant multi-decadal predictability from internal variability". I am not convinced though that as they stand your results can justify this claim (although I'm not suggesting this is necessarily incorrect). By selecting the best 30 (or similar) simulations from the CMIP6 archive based on the 9 most recent years, could it not be possible that what you are doing is weighting your results to certain models (which is in itself a rich area of literature). This could be those models with the most realistic response to forcing, especially when outside the anomaly period at the start or end of the record. Or it could potentially be those with the most realistic modes of variability. Both of which, I would have thought could theoretically give an increase in skill over the next decades. This is especially pertinent, given that you state that the method gives a better constraint that a single model ensemble (although as you say this is expected given the increased number of simulations to sample from). I think that this point needs to be discussed, and preferably investigated. One simple test you could do, without the need for further analysis, would be to check if you are selecting simulations preferentially from one climate model or if the 30 members are selected from the full model range. In addition it would also be interesting to see if the skill varies through time.

Author Response:

We thank the reviewer for the encouragement and the constructive insights. The reviewer is right that the added skill can potentially also be due to selecting those models with a more realistic response to forcing, or representation of variability modes. We have extended our analysis to provide some insight into characteristics of the variations that may be related to the enhanced skill in the constrained ensemble, and added/adjusted some discussion in this regard.

First, we added some time series plots for the regions where the temperature hindcasts of the constrained ensemble show added skill over the full CMIP6 ensemble (e.g. tropical Pacific, north Atlantic and eastern Asia (supplementary Figure S2)). These figures indicate improved long-term behavior of the constrained ensemble, which may be related to the representation of trends but also multi-decadal variability. In particular, the constrained ensemble better captures the cold anomalies in these regions in the early part of the hindcast where both the unconstrained CMIP6 and DCPP ensembles are warmer than observations and the Best30 constrained ensemble. The constrained ensemble also better captures observed decadal-scale variations in the warming rate in these regions, whereas the warming rate in the all-member CMIP6 ensemble is more homogeneous in time. We have added some discussion on these characteristics: "*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".

In addition, we also checked statistics on how often the different models are selected, and this can in fact be fairly uneven for some start dates (but note that also in the full, unconstrained, ensemble there are substantial differences about the number of runs contributed by different models). Figure R1 (below) shows that the CanESM5 model (which provides 25 ensemble members) is selected most frequently for most start dates, and also e.g. MIROC6 (providing 50 members) and MIROC-ES2L (providing 30 members) are chosen more frequently than other models, however the selection frequency is overall not proportional to the number of ensemble members provided per model.

To further test if the skill is due to the over-proportional selection of some models, we also implemented the constraining method so it limits the number of runs that can be selected from any models to 5 or 3 members (see Figure R2). The pattern of skill for these selections is mostly similar to the results shown in the main text (although slightly lower), which suggests that the role of overly selecting just a few models may be rather limited, and points to an important part of the skill we find being indeed due to the phasing of internal variability.

Finally, the question about temporal variations of skill is indeed an interesting one. As we are focusing on decadal to multi-decadal time scales here, the degrees of freedom of the time series are already relatively low. It would be interesting to address this question in the future, as the 'initialisation' in comparison to observed SST patterns would in theory allow us to extend the set of hindcasts substantially back in time. However, in this study we prefer to start the hindcasts in 1961 to allow comparison with the DCPP-A decadal hindcasts.

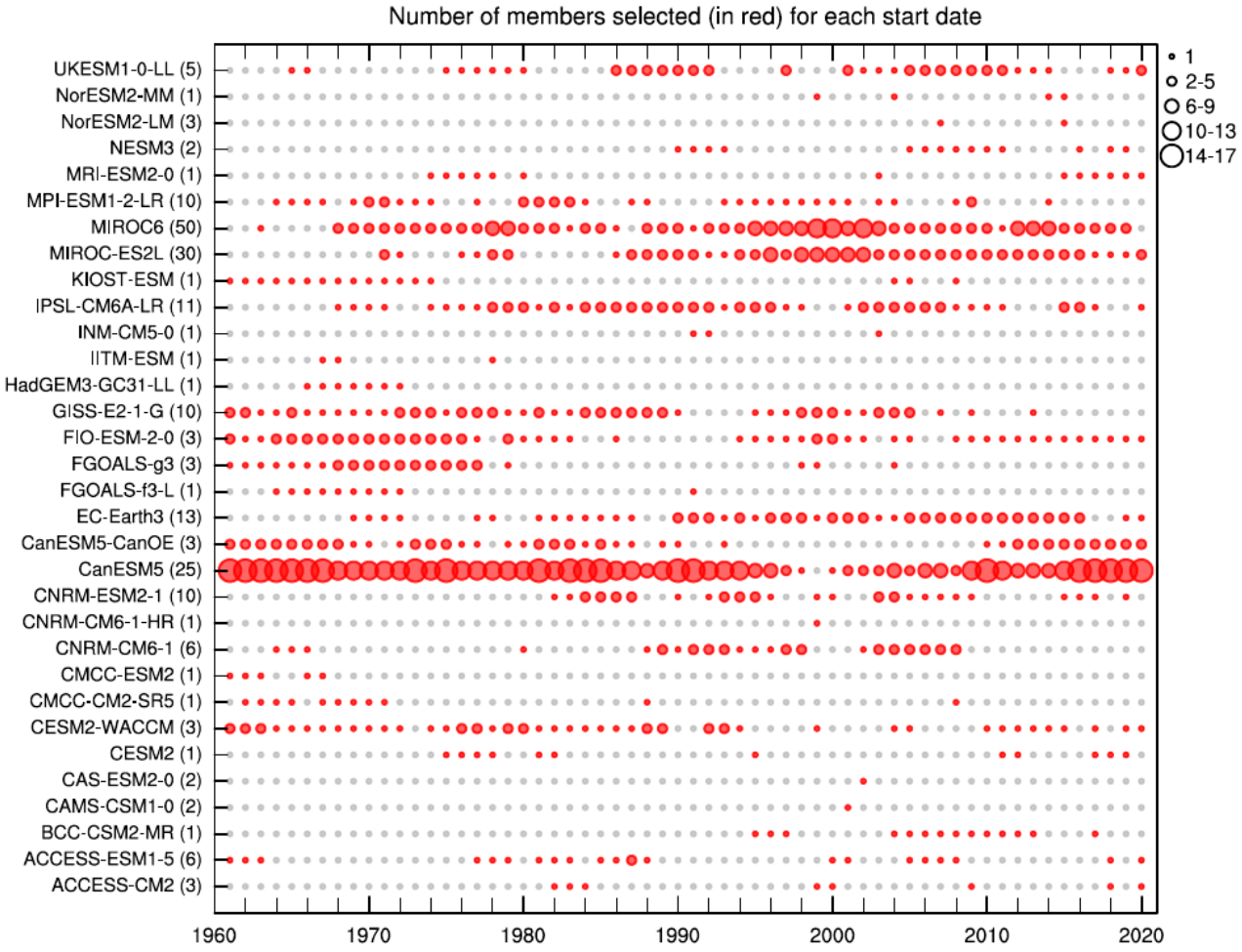

Figure R1: Count of how often each model is selected as part of the Best30 ensemble for each start date. The total number of ensemble members used are shown along with model names. The circle size indicates the count of how many ensemble members from each model are selected.

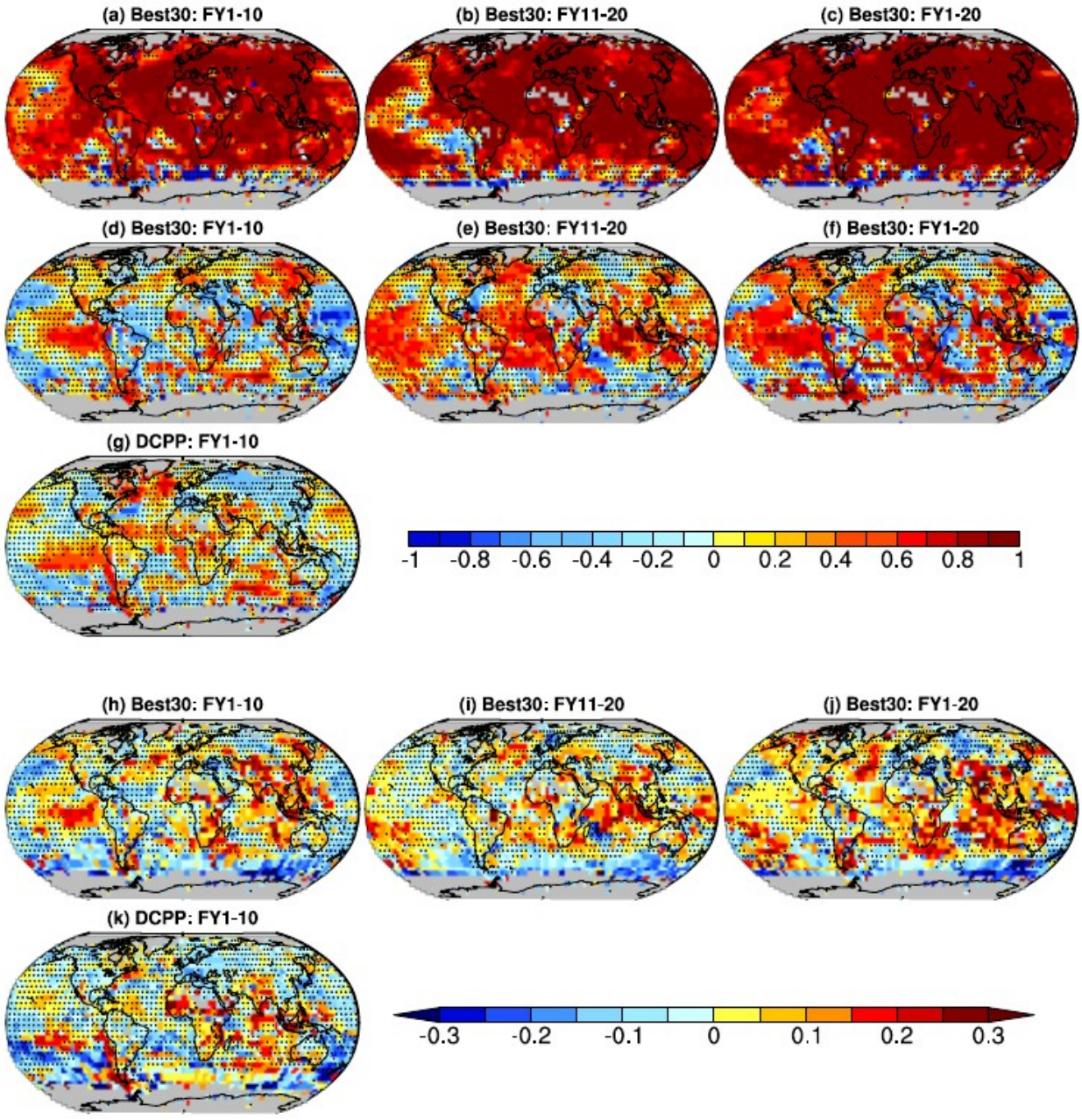

Figure R2: As Figure 2 of the manuscript, but limiting the number of members for a model to be selected as part of Best30 to a maximum of 5.

I am also somewhat surprised that you have not shown results for future projections, since that would seems to be the logical direction for this type of analysis. Although, perhaps this is being left for future work (which would be perfectly justifiable).

Author Response: We would like to understand some features of future projections in more detail, in particular as it also involves effects related to the different climate sensitivities of the models, and plan to present this in dedicated future work.

More minor points:

L 84. I think that adding the acronym for the Extended Reconstructed Sea Surface Temperature Version 5 dataset would be helpful.

Author Response: Thank you, we added in the revised version of the manuscript the acronym ERSSTv5.

L88. Why was a reference period of 1981-2010 chosen, and are the results sensitive to this?

Author Response: That chosen reference period lays roughly in the middle of the investigation period. However, as we are using anomalies, the choice of reference period would only affect the magnitude and sign of the local anomaly values, but not the patterns of where the high and the low values are (which is what is important for the constraint based on pattern correlations). We also repeated the analysis with the longer 1961-2010 reference period, and as expected the results regarding ACC and residual correlations are very similar to the results shown in Figure 2 (see Figure R3 below):

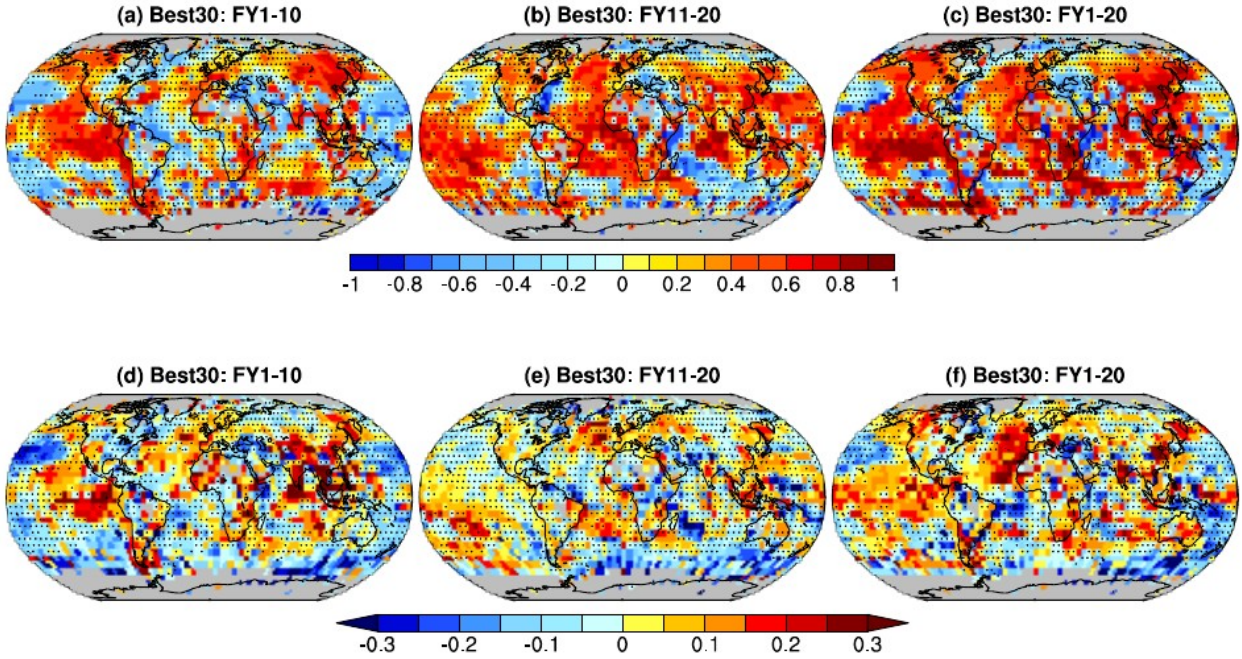

Figure R3: Residual correlations (a-c) and RPSS (d-f) for the Best30 ensemble means similar to the results shown in Figure 2 of the manuscript but based on anomalies computed with a reference climatology of 1961-2010 period.

L117. Was the forced signal that you removed the multi-model mean?

Author Response: Yes, we now clarified in the text that the forced signal was estimated from the ensemble mean of all 212 CMIP6 members.

L151. Why global warming and not external forcing in general? I would have thought that anthropogenic aerosols and volcanic eruptions might also have an impact.

Author Response: You are right, we have replaced ‘global warming’ by ‘external forcing’. Thank you.

L261. It would be useful to cite papers which have suggested that forcing (particularly natural) could contribute to the slow down, (see e.g. box 3.1 IPCC AR6 WG1)

Author Response: Thank you. We have added a brief note also mentioning studies that highlighted forcing (e.g. related to small volcanic eruptions) contributing to the slowdown in global warming during that period.