Reviewer 2

In my previous review I had criticised the quality of the presentation of the results. The enormous number of plots shown made it extremely difficult for the reader to filter out the relevant information. The authors have made an effort to improve the presentation by moving the maps of the sensitivity tests to the Supplement and instead replacing them with a new summary Fig.5.

While I do welcome this change towards improved readability, I don’t think the authors should have stopped there. Many of the relevant and important results are now included in the Supplement, specifically those that were added in response to the questions of the reviewers. Yes that makes the results in the main manuscript more accessible. However, it still leaves the reader with what I see as an unacceptable task of having to filter out the appropriate information from the immense wealth of plots in the Supplement. The Supplement now contains 13 figures of global maps with 9-12 subplots each, in total that's more than 130 global maps of results. I have spent a reasonable amount of time studying the manuscript and have found it very tricky to find those maps that I needed to convince myself of the results discussed in the paper.

As a reviewer, I see my role to suggest ways forward on how to improve the paper – improving the readability seems to me the most pressing issue with this manuscript now. If the authors would like to address a wider readership and not only the dedicated, specialist reader, I would urge them to seriously revise their material in terms of how it is presented. In my view the manuscript describes very interesting and novel work which clearly should be published. However, I am concerned that the paper in its current form might not pass the readability threshold for many potential readers – it is just not clear enough how you demonstrate in your graphics the main results. Instead of focussing on using the figures to support your findings and conclusions and pointing the reader to those, you have put an excessive amount of data analysis into plots and compiled them all together, leaving the reader to see subtleties of results in something like 10 subplots per page. Unless you think more rigorously about how your plots can demonstrate the main findings and underline your statements, I’m afraid I can’t recommend publication.

Author response:

We are grateful to the reviewer for carefully checking our revised manuscript and providing further comments. We are pleased to note that there seem to be no further scientific doubts, and the remaining concerns by the reviewer are primarily about how the results are presented.

We carefully considered the concerns and suggestions by the reviewer when revising the manuscript again, and questioned for each figure and result whether it was reasonably presented for the purpose it had within the study, and tried to optimise if not. But we also note that in some cases this enters the realms of personal taste. While we understand that the purpose of these comments is to optimise the readability of our manuscript (and we are grateful for the critical feedback), we also believe that some views need to be balanced against the other two reviewers who both also provided thoughtful reviews but did not raise such concerns.
Our logic when structuring the material between main text and supplement was that the main text figures contain the crucial information needed to follow the paper (i.e. the key results and a figure summarising the large amount of information from the sensitivity tests). And the Supplement now primarily contains those supporting results that may provide useful information for readers interested in specific details, but may not be strictly necessary to follow the story of the paper. We agree with the reviewer that this results in a large number of maps in particular (e.g. the sensitivity of each method choice is shown for 3 forecast times and 2 skill measures). But as argued above, these are intended to provide supporting information for readers interested in specific details. If removing some of these plots, readers may on the contrary also miss some information they would be interested in.

Having said this, in our revision we have (slightly) reduced the number of maps shown in the Supplement by removing those that were repeated but shown elsewhere (our initial intention when including such repeated figures was to facilitate easier comparison, but these could certainly be removed without losing information). If we were to further reduce the number of maps shown in the Supplement, one possibility could be to remove those figures from the sensitivity tests showing e.g. RPSS (e.g. S5, S6(f-j), S8, S11) and only keep those figures showing correlations. On the one hand, the conclusions of these figures are similar to what the corresponding residual correlation plots show. On the other hand, these maps underpin the bar plots shown in Figure 5b, and some readers may be interested to see these details as well. Also correlation (as a deterministic skill measure) and RPSS (as a probabilistic skill measure) are complementary. We would therefore favour keeping the map plots for both measures, but we are happy to follow the guidance by the Editor on this question in case he sees further need to reduce the number of maps shown in the Supplement.

Furthermore, in response to the issues raised by the reviewer, we applied changes to almost every figure (see list below). We hope the reviewer agrees that these changes improve the readability by making the information easier to grasp and in some cases even added essential information that was missing before.

- We have reduced the total number of maps shown in the Supplement by removing repetitions of the DCPP map (as shown in Fig. 2) from the Supplementary Figures S9 and S13, and by removing repetitions of the Best30 results from Figures S7 and S8.
- We have added colour bar titles to identify which measure is shown directly on the Figure: in Figures 2, 3, 4, S1, S3-S16.
- We have added axis labels where they have been previously missing (Figs 5, 6, S2), which also contributes to easier readability of the figures.
- Following the reviewer’s specific suggestion we have put some common information to the outer frame in front of rows of map plots.
Specific comments:

In addition to reducing the number of overall maps shown, better labelling of the subplots can also help improve the readability of the figures. In figures like Fig 2-4 and similar ones in the Supplement, the titles of the many subplots are often identical for different performance measures and experiments. Instead, adding descriptions to the outer frame of the figure (rather than for individual subplots) can help make the plots easier to digest. I would also find it helpful to have the information about the performance measures and experiments included in the figure in addition to the caption.

Author Response:
We have improved the labeling in the different figures to improve readability, following the suggestions by the reviewer. In particular, we have added descriptions to the outer frames of the figures, and added titles to the colourbars to specify the measures they show, added axis labels where missing, etc. We hope the reviewer agrees that this polishing has helped to improve the readability of the figures. Please also see our response to the general comments above.

In Fig 2S, it is not obvious to me why the observed data curves for the two different 10-year forecast year windows are different. Naively I would expect the observations to be independent of the forecast year. That is, why are the observed timeseries from 1960 to 2000 different for FY1-10 and FY11-20? In case of the North Atlantic, why are the late 1990s temperatures warmer than the early 1990s in the plot for FY1-10 but the same late 1990s temperatures are colder for the plot of FY11-20? If this is possibly an effect of some sampling uncertainty, what are the implications of such drastic changes?

Author Response:
We thank the reviewer for pointing out this confusion. In the previous version of the figure, each data point was assigned to the first year after applying the constraint, which led to a shift of 10 years between the columns for FY1-10 and FY11-20. We have corrected this in our revision, now assigning each data point to the first year of the time window shown. This results in the axis labels for FY11-20 shifted by 10 years.