We thank the editor for the additional comments on the manuscript; our response to each comment, and details of further changes made, are below.

Overall the paper is much improved and the authors have addressed reviewer comments.

However, there is a significant clarification that needs to be made. The authors have clarified in separate correspondence that: "the method is always applied on each model separately. So the pattern for a given scenario (or set of scenarios) is found for each model separately, and then this model-specific pattern is multiplied by the GMST for the target scenario from that same model, to get a model-specific emulation. Then in the plots the multi model means are taken - and for uncertainty analysis this is compared to the inter-model deviation."

Please make sure the methods section clarifies this, this should be stated in more than one place so that this point is not missed. In particular provide more detail on the methodology in terms of what steps were taken in what order, since this will presumably make a difference in the results.

This does raise some questions. The paper, therefore, seems to be focusing on decomposing pattern scaling of the multi model mean pattern. This, therefore, seems to result in the paper focusing on areas where models agree in the pattern of change. Where models disagree, the mean change will be smaller and less signifiant. In this sense, the current conclusions in the paper seem to be somewhat limited - they don't apply to GCM patterns in general, but pertain more to those areas of the globe where there is robust agreement between GCMs.

This is contrast to evaluating how the pattern scaling method works for individual climate models, and then presenting analysis of how well the method works across models.This is likely a more common application of pattern scaling. For example, most impact analyses use not mean patterns of change across models, but patterns for specific models (usually repeating the analysis for multiple models). Can the results be analyzed to also determine if the conclusions of this work are similar for individual GCM patterns in terms of the decomposition of results into the two components? Are the conclusions robust across different GCMs? Can the analysis and results be used to address this point?

The lack of clarity in our draft manuscript has led to some confusion on this point.

We apply all the analysis on individual models, including calculating pattern scaling errors. We only take multi model means when we are plotting, and then use the intermodel standard deviation of the results being plotted to give a sense of the level of agreement amongst models.

For example Figure 4 shows the multi model mean pattern scaling errors - i.e. first we do the pattern scaling for each model, then we calculate each model's error by subtracting the ESM data from each emulation separately, and then we average the error across models. And in Figure 6, the time-series of the "Timeseries" and "Pattern" errors are calculated for each model separately, and then the multimodel mean is plotted in the solid line, with the shading showing 1 standard deviation across the models, showing the results are robust across models.

We decompose the errors separately for each model before comparing. We do analyse where geographically the models agree/disagree on the pattern itself, which we feel is an important question to ask, but we don't average across models until the very end.

To address this lack of clarity we have edited the methods section to better present the methodology. In particular we have added a paragraph at the end of section 2.2 to clarify the application to individual models. We have also added to the conclusions to emphasise this point.

Specific comments

Section "2.2 Pattern Scaling Methodology". Current version is largely one large paragraph, which is difficult to read. Break this up (and add the material mentioned above).

We have split this large paragraph into several smaller ones to aid reading, as part of the editing process above.

Line 141

"Inter-model results are averaged first over each model ensemble, with this model average then compared, to avoid weighting by the ensemble size of each model."

This implies that for models with more ensembles, patterns are likely more robust, with inter-ensemble noise averaging out across ensemble members. While for models with fewer ensembles, patterns are less robust (larger regression uncertainty, all else being equal.) How does this impact the results?

This is an interesting question; we did some analysis of the sensitivity of the pattern to the ensemble size, and found that >=3 members was generally enough to give a robust pattern over most areas. We might look to include this analysis in a future paper and explore this further, but we don't think it justifies inclusion in this paper. Most of our ensembles were larger than 3 so this variability likely wouldn't drive a huge variation in our results, but some ensembles had only one member so there will be an effect. We have added a sentence to the discussion noting the range in ensemble sizes and suggesting this might be useful to investigate further in additional work.

This leads to another question, it appears that regression fitting uncertainty is not used in this analysis? Or is it?

We don't analyse the regression uncertainty in the presented results here; since pattern scaling assumes linearity, we test the effect of this nonlinearity error by comparing to the ESM data, rather than analysing the error in the regression itself.

Line 160: "For calculation of the pattern in a model with multiple ensemble members, MESMER applies the regression across the data from all the members simultaneously."

clarify what this means. Is the regression performed for each ensemble member and then somehow averaged? (Are the coefficients of the regression equation averages of the coefficients for each ensemble member? or something else? How are non-zero intercepts treated.)

We have clarified this section – the regression is applied once, on the concatenated members. Any intercept is added to the emulation, noted on lines 160 and 197.

Line 199

Here

"The timeseries error, in row two of Figure 1, ..."

Should this be column 2? (e.g. isn't every panel in column 2 the time series error?).

Not exactly; what we term the "Timeseries error" is shown in the 2nd row in Figure 1; row 3 shows the "Pattern error".

Please list in the SI the exact scenario data members used (either by DOI or other identifier), along with the data of access of the cmip6-ng database. This is necessary to assure your results can be, in principle, replicated. While the CMIP6 model data is relatively stable at this point, changes and new submissions may occur, so it is necessary to document the exact data used.

We have replaced Tables S1 and S2 with a single Table S1 which lists all ESMs, the scenarios for which they are used, and the individual members used for these, using the ripf codes.

Line 220 this "Figures 2a and b shows the multi-model mean hist-aer and hist-GHG response pattern based on regression across the whole period (1850-2020)". is unclear.

Define how the multi-model mean regression is derived in the method section.

We hope this has been clarified by our additions to the methodology. The pattern is calculated for each model separately (i.e. the regression is applied to 1850-2020 for each model separately), and the multi model mean of these patterns is shown.

line 495 -typo

-> that of the scenario being emulated

Many thanks for spotting this – we have corrected this typo.