Reply to reviewer 1

Thank you for your thorough read, and useful and constructive comments. We have considered them all and made adjustments, clarifications or additions accordingly, all of which we feel strengthen the manuscript. Below we answer the concrete comments one by one in greater detail.

“As currently presented, the paper features mainly the validation of METEOR against the CMIP6 MMM, rather than validations of individual CMIP6 models. This is not made clear in the abstract or most of the text, where the reader gets the impression that METEOR, in its current calibration, is a useful emulator of individual ESMs. That might be the case, but it is not shown. In other words, the paper is not clearly framed as being limited to emulating only the multi-model mean. If the authors wish to present METEOR as an individual-GCM/ESM emulator, then the paper needs to test the appropriateness of CMIP6 model-by-model responses. Only small in-sample goodness-of-fit metrics are shown (e.g., Figure A1 panels a and d present the RMSE for the GHG response in the in-sample abrupt-4×CO_2). I therefore strongly encourage the authors to show more model-by-model validation—for example, by including absolute-error maps of 20-year means for individual model SSP5-8.5 or SSP1-2.6 out-of-sample temperature and precipitation fields for 2080–2100. Tables of RMSE and MAE values by model and scenario would be useful in an Appendix, allowing comparison to alternative emulation techniques. Similarly, Figures 8 and 9 could be extended to include maps of the best and worst CMIP6 model fits, rather than showing only MMM differences.”

We agree that the framing of the figures is skewed towards the multi-model mean, but we do mean to convey the applicability of METEOR as an individual model emulator. The appendix has many individual model plots, and we will update these and make the individual model results clearer. We will also upload spatial fit plots for all the models that we ran METEOR with in a Zenodo repository. However, the reason why we focus more on the multi-model mean, is that the choices of models would be somewhat arbitrary, and there are really a lot of models to include. Defending the choice of one particular model over the other can then be tricky. However, noting your comment, we have chosen now to include plots for all the models for which we had data for ssp5-3.4-over in plots individually in the main text, as this was a limited number of models, illustrating fits for an out of sample scenario, and they include overshoot - the modelling of which is a particular motivation for METEOR. We fully agree with the reviewer that error metric tables for all models and scenarios are useful in the appendix and have added them accordingly.

“At present, Figures 5–7 and B13–B20 show useful comparison plots for global-mean temperature and precipitation responses. That is reassuring (and a great result), but for an emulator of regional climate responses, more regional comparisons are needed. The global-mean response can be obtained much more simply—e.g., as an extension of the C-SCM with a few lines of code and these calibration parameters. I suggest replacing (or extending) Figures B13–B20 with figures that show the worst- and best-performing regions (using either custom definitions or IPCC AR6 regions). Regional responses could also be shown as maps—you already include CMIP6 MMM comparison maps in your Figures 8 and 9.”

We agree with this criticism and hope that the addition of the new main-text figures is useful in that regard. They show regional model versus emulation scatter plots per region for 9 different regions, and regionally separated plots of change in precipitation versus temperature change per model for both direct output and emulation in the over-shoot scenario for each of 8 regions, to both show better how well the model can fit both per model and per region. For ssp5-3.4-over, we also show maps of fits per model in the appendix. We also hope that the above-mentioned spatial maps for every model uploaded to a Zenodo repository will help with this.

“The utility of these results for impact emulators depends on each emulator’s needs. METEOR v1.0 is limited to annual-mean projections of best-estimate warming and precipitation changes, and does not yet include variability, compound-event modeling, climate-oscillation modes, distribution tails, etc. Although some of these caveats are mentioned in the conclusion, an explicit upfront statement of the current emulator’s scope (and its limitations) would be helpful.”

We will make the limitations of METEOR v1.0 clearer also earlier. We would like to point out though that, although it has not been validated outside of the mean annual variables considered in the paper, the setup is a bit less restrictive than this, as METEOR can in principle model other annual-mean projected variables for which there is data, and for which the underlaying assumption of forcing driven timescale patterns holds, that of course does not include any of the implications you list here.

“Looking at the GHG and “residual” response patterns, one wonders whether they are intended purely as statistical fits (in which case they need not be physically interpretable, as long as applications stay within the training spectrum), or whether they represent physically meaningful patterns. If the latter, one could apply the emulator beyond 2100 to

2300 with more confidence. Since the authors do not clearly state that these are statistical fits—and some discussion refers to physical interpretation of short- and long-term responses—I suggest the following:

1. **Equilibrium response aggregate pattern:** Add a fourth column to Figures 3 and 4, as well as B1–B12, that sums the short-, medium-, and long-term response patterns. This should yield the equilibrium response pattern, which readers can then evaluate for physical plausibility. If the equilibrium response is not physically plausible (and some of the patterns seem hard to interpret), then these components should be framed explicitly as purely statistical fits valid up to 2100 for the shown validations. Alternatively, you might introduce training constraints—for example, requiring that the sum of the three timescales falls within a physically plausible range. You could also discuss whether the land-ocean warming ratio evolves plausibly from short-term through equilibrium response.
2. **Full colorbar:** Many patterns appear clipped by the chosen colorbar limits, making it hard to see true minima and maxima. Please include a full colorbar for these figures and choose its range to include extreme values (possibly on a logarithmic scale) so that readers can see tail-end values. For example, in Figure B6 the MIROC-ES2L long-term GHG precipitation response is unclear; likewise CanESM5's short-term precipitation response in Figure B5 and UKESM1-0-LL's medium-term temperature response in Figure B3.

"

We thank the reviewer for this and agree that some more discussion on the interpretability is in order. The model does in principle only provide statistical fits, but they should contain physical information, the first point here also has led us to reevaluate and redo our figures 3, 4 and B1-B16. Though it is in principle true that the sum of the three patterns is the equilibrium response of the model, this is not really a meaningfully constrained quantity from the model and nor is the display of the three patterns in the way shown in these figures. For the shortest timescale, the equilibrium pattern response makes sense, but as the timescales increase, there is an ambiguity between the amplitude of the response and the timescale of equilibration – such that extrapolating beyond the timescale of the training data to an equilibration in hundreds or thousands of years is highly under-constrained.

The fit and training data fit less and less to the equilibrium response, and more and more to the linear or only the few first terms in the Taylor expansion of equations 3 and 10 for the time. In the equilibrium case the fit would be fitting the underlaying pattern, BGHG and Baer, but with only 150 years of simulations (or less), the fit for the longest timescale only

ever has training data to fit something like 1/tau_K *BGHG (i.e. linear regime), so the relative strength of the pattern is dampened significantly. As the training data does not go over this, we also do not expect the model to yield reliable results for runs that are significantly prolonged in time. I.e. we have much more trust in out of sample results for different forcing/emissions-pathways than in the extensions of the runs to very much longer timescales (for that longer training datasets would be needed). In effect, the time-evolving response of the slowest modes to a step change in forcing appears to be a straight line in the ~100 year training data considered here – there is simply no information on how that mode will equilibrate. For applications on the ~100 year timescale, this is not a problem – but the model is not suited to an extrapolation to equilibrium.

Exactly where the trustworthiness of the model ends is a topic which should be explored in the future, and we will mention this. What this also means is that the pattern comparison in these figures as they stand are not very meaningful for two reasons: 1. The three patterns are not scaled in a meaningful way so the relative strengths between them are not really reflective of the relative strengths between them in any part of the model which has any validity, and 2. The mean between models also doesn't make much sense here, as models with larger tau_k values will have larger relative weight, particularly for the long timescale patterns, making the summation and mean between them not particularly instructive, and hence the physical interpretation of them even less so. To amend this, we have now reframed the plots to show the contribution of different timescales to the total warming response 100 years after an abrupt change. We think this is a more meaningful illustration, given it shows the relatively minor contribution of the slow timescales to the total response in year 100 – but does not imply any confidence in extrapolation to longer time scales

This is achieved by scaling the patterns to their mean value in years 80-120 in the reconstruction of abrupt-4xCO2 for the GHG patterns and 1980-2020 in the reconstruction of historical-ssp245 for the aerosol patterns. This way we can also add them together to produce a meaningful summed pattern, and the multi-model-mean will be fairly weighted between models, showing results within the valid range for the emulator. These updates also solve the colorbar issue as the very large amplitude of the longest timescale pattern is appropriately dampened by its temporal coefficient.

"Since METEOR emulates both variables, it would be useful to examine their regional co-evolution. For instance, map percent precipitation change per degree of warming—some regions should show ~2–5 % $°C^{-1}$, moisture-saturated regions near Clausius–Clapeyron (~7 % $°C^{-1}$), etc. This would provide a physics-based check on the emulator's joint behavior."

We thank the reviewer for this suggestion and hope the new ssp5-3.4-over plots that show the relationship between temperature and precipitation change per model and model emulation globally and for 8 different regions can show the degree to which METEOR is able to capture the joint behaviour.

“The reported skill metrics (Pearson, RMSE) need context. Consider benchmarking against the ClimateBench test (doi:10.1029/2021MS002954) using NorESM2 output, or comparing to other published emulators. You might also compare each model’s emulation error to the inter-model spread in response patterns, to assess whether emulator errors are small relative to GCM diversity.”

We agree with this assessment and have therefore added a comparison to the ClimateBench test. We also include similar per model results in a new supplementary table. The skill metrics previously included are also comparable to skill metrics provided by the PRIME emulator, which we have pointed out in the text. We have also made spatial RMSE map plots for all scenarios.

“Lines 11–12, Abstract: You state that the emulation system can “accurately predict gridded responses to out-of-sample scenarios.” That is too broad, since you demonstrate accuracy only for the MMM, annual means, and expected values. Please qualify.”

We have now qualified this statement, however, we believe that with updated figures we also demonstrate accuracy at model level and for gridded responses, so the qualification is not as strong as suggested in this comment.

“Line 47: Do you mean that ClimateBench data are not widely available? They are provided via Zenodo—please clarify.”

We understand that the statement could be misunderstood. What we mean here is that the ESM output available to train on for ClimateBench is not available for emulation across the majority of CMIP6 (and upcoming CMIP7) models. In this paper, we chose to train exclusively on output experiments which have CMIP6 outputs for all ESMs. We agree that some of the experiments used in ClimateBench could possibly yield better and more physically informative fits to aerosol forcing specifically, even for METEOR, however, the point here is that we think a setup that can be run for *any* CMIP6 ESM model is an advantage, and showcasing the model and it’s performance on this dataset is therefore our priority. Both the *ssp370-lowNTCF* and the DAMIP experiments have been run by considerably fewer of the ESMs. We have clarified this now in the text. METEOR can be run from lightly processed widely available CMIP6 data, and also comes with a zarrstore data download capability, which can be used to directly download and process the CMIP6 data

available there, with the user not having to figure out any downloading or processing of CMIP data. We feel this greatly adds to the model's usability.

Producing an independent calibration of METEOR for a subset of models with a larger array of simulations (such as from climatebench) would certainly be valuable, but would imply a significantly different calibration pipeline. We feel this is beyond the scope of the current study.

"**Line 105:** "most impactful non-GHG forcer." Perhaps note that this is currently true but may differ under low-emission scenarios."

We agree with this point and have added a caveat in the text accordingly.

"Line 134: When you subtract the piControl "climatology," do you mean a 20- or 30-year rolling mean, a trend, or a non-parametric low-pass filter? Please specify."

We are subtracting the mean of the of the full piControl annual mean timeseries. We have added specification for this in the text.

"Line 137: Clarify whether you use cos(lat) for area weighting or each model's native areacella."

We use cos(lat), this is now specified.

"Figures 3 & 4: Much of the long-term precipitation response lies outside the colorbar range—consider widening it or otherwise showing pattern extrema."

The updated figure versions don't have this issue anymore.

"**Tropospheric ozone response:** Where is this captured? I assume in the residual (aerosol-scaled) response—please state."

We have not been clear enough on this point, but the aerosol-scaled residual response is *exclusively* mapped to sulphate aerosol related forcing. In our current setup, sulphate aerosol scales both direct sulphate aerosol forcing, and the totality of the aerosol-cloud interaction, but all other forcing responses, including tropospheric and stratospheric ozone, BC- and OC- aerosol direct forcing, and stratospheric water vapour forcing are mapped using the GHG-response patterns and timescale. Of course, as we are identifying the sulphate aerosol responses with a residual signal, any and all other forcers (and other uncertainties or inaccuracies in the GHG response modelling), will also feed into these, but they are not directly included. We have tried to make this clearer in the text now, by adding a sentence to the Methods introduction section stating this.

“Residual scaling bias: Using sulfate as the scaler for residual response may bias low-emission scenarios, since nitrate aerosols could dominate forcing by century’s end. Discuss this potential bias.”

We have added some discussion on this point. Nitrate aerosol bias is, however, a wider problem as the forcing from nitrate aerosol is both not well-constrained and highly dependent on both emission location, height and sector.