Author comments in response to Reviewer 1's comments on "UKCM2-LL: a new low-resolution GC5 configuration with constrained climate sensitivity – methodology and development" by Rostron et al.

The manuscript documents the process by which a new configuration of the Met Office Unified Model was assembled in order to have an effective climate sensitivity (EffCS) within the IPCC's likely range of values. The authors utilize a series of steps relying on amip-like simulations to form a PPE, which is run in pairs of present-day and plus-4K warming conditions to create a relationship between errors in variable fields of interest and the feedback parameter. Next, they use an emulator to probe the parameter space for suitable candidate configurations to test in coupled mode. After they narrow these down, an expert-based fine-tuning procedure is done to arrive at the final solution.

I appreciate the authors taking the time to put together a manuscript like this that provides a description of their procedure. As more modeling centers embark on auto-tuning or other tuning methods for purpose-driven applications, it will be interesting to see how the community builds upon and adapts these efforts over time. I particularly enjoyed the discussion in section 5 related to the authors' impressions of their own methods, as well as the deeper questions facing the community. That said, there are some minor areas where this manuscript could be improved, which I have outlined below.

We thank the reviewer for their careful review of the manuscript and for their constructive comments and suggestions. These have helped us to clarify several aspects of the paper and to improve the discussion. We have addressed each of the reviewer's comments below. (The reviewer's comments are in black; our responses are in blue.)

Comment 1

Lines 182-186: the authors mention using the modal and median parameter values. Looking at Rostron et al., 2025, it seems to rely on the prior parameter distributions, but how are these defined? The "flat top" nature is presumably expert judgement weighing in on the likely range as a subset of the total range, I imagine, but this isn't clear here. Is there a citation that can be given and/or a quick synopsis on how these prior distributions are defined? For clarity, it would be helpful to report what the median and modal values being tested are.

Thank you for picking up on this. The reviewer is correct that the modal and median parameter values rely on the prior parameter distributions defined through expert elicitation, but we agree this was not clear in the manuscript.

To address this, we will expand the description of the elicitation process, including how it built upon earlier GA7 and GA8 PPE elicitation exercises, rather than relying on the citations to Sexton et al. (2019, 2021) and Rostron et al. (2025).

We will also add the modal and median values for the 37 most influential parameters in a new table in the Appendix (which will be added in response to Comment 5).

Comment 2

Related to the variables in Table A1: how was it decided which variables should be assessed annually and which should be separated into DJF and JJA? For example, both SWCRE and LWCRE should follow monsoon precipitation, but only SWCRE and precipitation are assessed seasonally. Perhaps this is planned analyses for the companion paper by Bodas-Salcedo et al., but was there anything learned about potential structural errors in the model's ability to simulate these fields from this work?

Thank you for this comment. We acknowledge that the manuscript currently provides little detail on the choice of metrics in Table A1 (used for E_{amip}). We chose metrics that reflected variables typically used in the standard GC model development (e.g., Willett et al., 2026, Figure 8), and that we felt would capture a range of dynamic and thermodynamic atmospheric processes, whilst attempting to limit redundancy between the metrics.

DJF and JJA means were included for variables where capturing seasonal variability was considered particularly important e.g. mid-latitude seasonality in SWCRE. We felt that monsoon variability would largely be captured by precipitation and SWCRE, and that including seasonal LWCRE metrics would add little extra information, due to the strong correlation between LWCRE and precipitation in monsoon regions (and the tropics more generally). Without a universally accepted standard set of metrics, however, there is always going to be some level of expert judgement in these choices. We will update Section 2.5.1 to provide more detail on our motivation for choosing the metrics in Table A1.

Whilst we did highlight the structural behaviour between E_{amip} and λ_{amip}, we did not investigate structural model errors using the component metrics of E_{amip} (Table A1), nor is this planned for the companion paper by Bodas-Salcedo et al. (in prep.). However, this is a very good suggestion, and we will expand on the first paragraph in Section 5.2 to include this as an explicit example of how PPEs can be used to aid future model development.

Comment 3

The solar bug fix is described as having "been tested previously" and "found to have minimal impacts on simulation results," but Fig. B1-B3 show large trends during the adjustment period for that change. Are those trends meant to be interpreted as continuations of the trends from the previous changes?

Thank you for raising this. The trends highlighted by the reviewer are indeed continuations of trends from previous changes.

We demonstrate this in Fig. R1 below, which shows continuous timeseries to highlight the impact of the solar bug fix change. The black line is proto-UKCM2, the green line includes all of the changes up to 'ps_natl_dms_emiss=1.5' (see e.g. Fig. 4 in the original manuscript) and the yellow line additionally includes the solar bug fix. The impact on the incoming SW TOA flux is a clear but small reduction of 0.1 Wm^{-2} in the global mean. The timeseries for the other variables show that the impact is negligible compared to the interannual variability, and that the trends highlighted by the reviewer are continuations of trends from previous changes.

We acknowledge that the way Figs. 4 and B1-B3 are presented as non-continuous timeseries makes it difficult to determine whether trends seen for each change are continuations of previous trends, or new ones driven by the most recent change. We also acknowledge that our description of these figures did not make it clear that many of the changes, though applied sequentially, have simulations that start from the same initial conditions (e.g. those between proto-UKCM2 and 'ps_natl_dms_emiss=1.5'). We will revise the Fig. 4 caption to clarify that the plotted periods are not intended to be interpreted as a continuous timeseries, and that several of the simulations, corresponding to successive parameter changes, were initialised from the same starting state. However, we feel that presenting the data in this way remains useful because it highlights the impact of parameter changes where these are clear relative to internal variability and model drift, which was one of the primary purposes of these plots. We have therefore kept the figures as they were in the submitted manuscript.

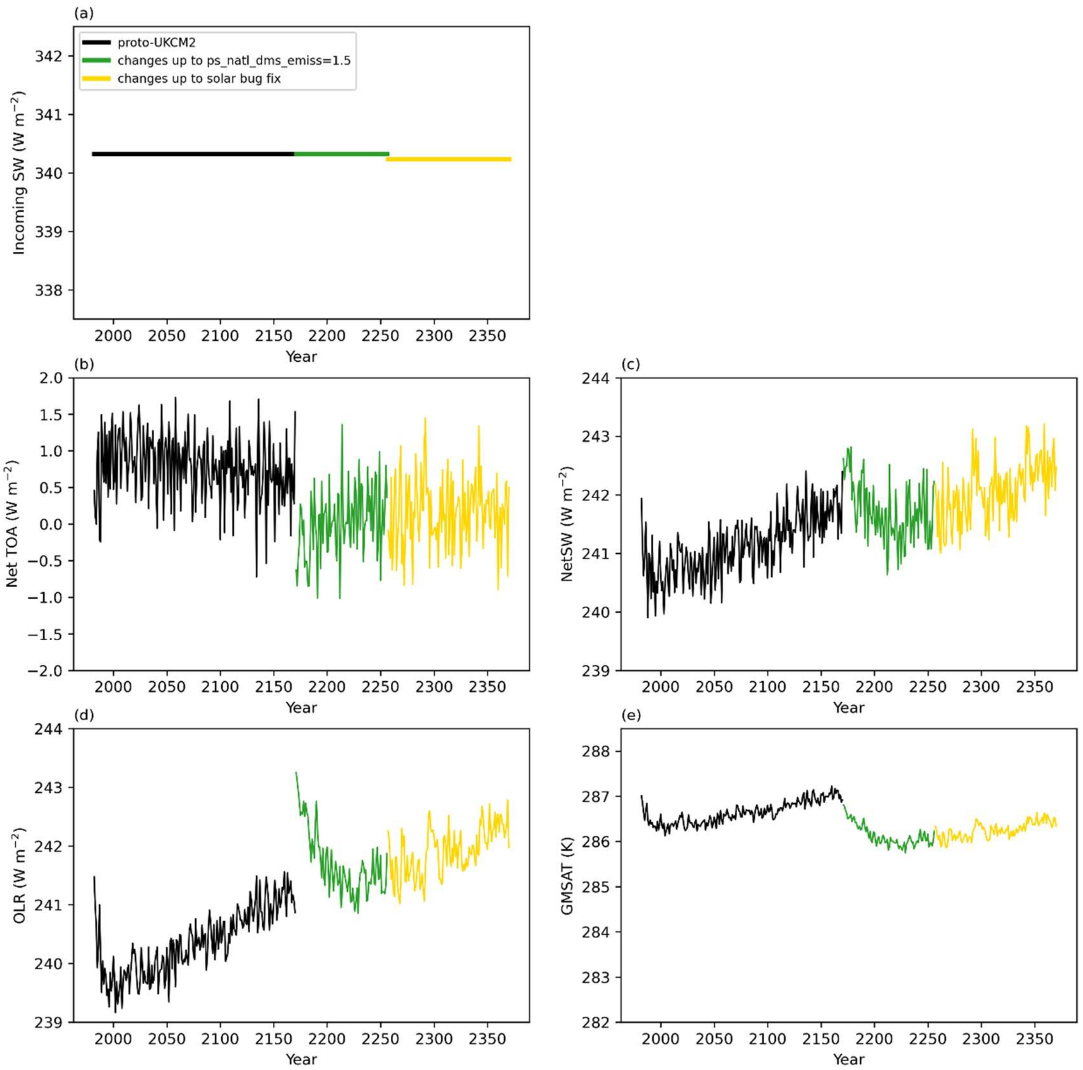

Figure R1 Timeseries for proto-UKCM2 (black), for a configuration including all changes up to ps_natl_dms_emiss=1.5 (green) and for a configuration additionally including the solar bug fix (yellow). The green and yellow simulations were initialised from the end point of the displayed black and green simulations, respectively (creating a continuous timeseries). Variables are global-mean annual-means for (a) incoming SW TOA flux, (b) Net TOA, (c) Net SW TOA, (d) outgoing LW TOA, and (e) surface-air temperature.

Comment 4

Why not include UKCM2 on Figure 1? Why not provide the RMSE values for the metrics going into Eamip for GC5 and UKCM2? It seems like it would be valuable to see how these criteria that were selected for the initial PPE perform after all of the fine tuning results. I imagine the follow-up paper will cover many of these features, but it would be valuable to at least get a hint at how this model performs relative to its starting point. Especially if claims like "its performance is competitive with CMIP6 models across a wide range of variables," are being made (L680-681). We can piece together which dot is the proto-UKCM2 by comparing Figs 1 and 6, so we know that Eamip is 1.4x that of GC5. Maybe just the global RMSE values for each variable would be sufficient here, leaving the follow-up work to focus on the spatial patterns.

Thank you for this comment. We agree that the manuscript should give more detail on how UKCM2 performs relative to its starting point (GC5). We will address this in two ways. First, we will modify Fig. 1 to highlight proto-UKCM2, to give a clearer indication of the degradation in performance, without readers having to infer which point corresponds to proto-UKCM2 by comparison with Fig. 6. Whilst we agree with the reviewer that including UKCM2 would be a useful addition, we have found that re-running the exact 5-year AMIP experiments required would be technically challenging due to an upgrade to our HPC and to the software used to run the experiments, and so we have decided not to pursue this.

Secondly, we will add text to Section 5 ('Summary and Discussion') to provide more detailed information on the degradation in climatological performance analysed in Bodas-Salcedo et al. (in prep.). While we do not provide the E_{amip} component RMSE values for UKCM2 (due to the issues described above) the additional text will provide quantitative information on its performance (relative to GC5) evaluated in coupled historical experiments in Bodas-Salcedo et al. (in prep.). This shows an average degradation (relative to GC5) of around 30% for global-mean, annual-mean RMSEs for 16 climate variables, averaged over 1981-2000. Compared to GC5, performance is approximately maintained for upper-level temperature and relative humidity, and surface temperature and pressure, whilst the largest degradations (of around 60%) occur for SW TOA flux and SW cloud radiative forcing. (Potential relationships between these degradations and the feedback changes – which we have shown are largely SW cloud feedback changes – will be investigated in more detail in follow-up work.)

Comment 5

It would be valuable to future research efforts to give a complete list of the 37 "most influential" parameters used in the Monte Carlo sampling, if possible. Knowing which parameters the feedback parameter is sensitive to would be valuable for designing future PPEs.

We agree with the reviewer that a table including the parameters which were optimised will be useful for future research. We will include this in the Appendix in the updated manuscript, including modal and median parameter values (in response to the Comment 1 above).

Comment 6

Line 235: "with at least of 174" -> "with at least 174"

Thank you – this has been corrected.

Comment 7

Line 313: please specify your method of computing λamip

A description of the method for computing λ_{amip} was given in Lines 307-310, but we appreciate that an explicit equation may help to clarify this. We will therefore add the corresponding equation to the revised manuscript: $\lambda_{amip} = \Delta N/\Delta T$ (where ΔN is the change in global-mean net TOA flux and ΔT is the change in global-mean surface air temperature).

Comment 8

Line 527: "less stringent criteria…" can you provide examples?

Thank you for picking up on this – we did not make this clear in the original manuscript. The selection of the additional 8 candidates was based on the emulated λ_{amip} value – from lowest to highest - amongst the candidates which had not already been selected. We will update the manuscript at line 527 to explicitly state these criteria.

In hindsight it would have been better to base the selection on the simulated values – either using λ_{amip} values, or a combination of λ_{amip} and E_{amip} as was done for the selection of the initial fourteen candidates. This would be a simple improvement to our methodology, and we will add this to the Discussion Section 5.1 ('Improvements to the methodology').

References

Bodas-Salcedo, A., Rostron, J. W., and Sexton, D. M. H.: UKCM2-LL: a new low-resolution GC5 configuration with constrained climate sensitivity – assessment of CMIP7 DECK simulations, in prep.

Rostron, J. W., Sexton, D. M. H., Furtado, K., and Tsushima, Y.: A clearer view of systematic errors in model development: two practical approaches using perturbed parameter ensembles, Clim. Dyn., 63, 354, https://doi.org/10.1007/s00382-025-07717-5, 2025.

Sexton, D. M. H., Karmalkar, A. V., Murphy, J. M., Williams, K. D., Boutle, I. A., Morcrette, C. J., Stirling, A. J., and Vosper, S. B.: Finding plausible and diverse variants of a climate model. Part 1: establishing the relationship between errors at weather and climate time scales, Clim. Dyn., 53, 989–1022, https://doi.org/10.1007/s00382-019-04625-3, 2019.

Sexton, D. M. H., McSweeney, C. F., Rostron, J. W., Yamazaki, K., Booth, B. B. B., Murphy, J. M., Regayre, L., Johnson, J. S., and Karmalkar, A. V.: A perturbed parameter ensemble of HadGEM3-GC3.05 coupled model projections: part 1: selecting the parameter combinations, Clim. Dyn., 56, 3395–3436, https://doi.org/10.1007/s00382-021-05709-9, 2021.

Willett, M., Brooks, M., Bushell, A., Earnshaw, P., Smith, S., Tomassini, L., Best, M., Boutle, I., Brooke, J., Edwards, J. M., Elvidge, A. D., Furtado, K., Hardacre, C., Hartley, A. J., Hewitt, A. J., Johnson, B., Lock, A., Malcolm, A., Mulcahy, J., Müller, E., Renfrew, I. A., Rumbold, H., Rooney, G. G., Sellar, A., Ujiie, M., van Niekerk, A., Wiltshire, A., and Whitall, M.: The Met Office Unified Model Global Atmosphere 8.0 and JULES Global Land 9.0 configurations, Geosci. Model Dev., 19, 1473–1517, https://doi.org/10.5194/gmd-19-1473-2026, 2026.