Author comments in response to Reviewer 2's comments on "UKCM2-LL: a new low-resolution GC5 configuration with constrained climate sensitivity – methodology and development" by Rostron et al.

This manuscript documents the development of UKCM2-LL, a new low-resolution GC5-based coupled climate model configuration intended for the UK's CMIP7 contribution. The main motivation is that the lower-resolution version of GC5 has an excessively high effective climate sensitivity, EffCS = 6.7 K, and an overly rapid recent historical warming trend, which limits its suitability for long-timescale climate applications, decadal prediction, and Earth-system modelling. The authors develop UKCM2-LL by tuning GC5 parameters to reduce EffCS into the IPCC AR6 "very likely" range while attempting to retain acceptable climatological performance. The final configuration achieves EffCS = 3.6 K, although with some degradation in climatological performance relative to GC5. It uses a large 503-member atmosphere-only perturbed-parameter ensemble, statistical emulators, AMIP and AMIP-future-4K experiments, and a smaller set of coupled piControl and abrupt-4×CO_2 simulations to identify and refine candidate configurations. The study is valuable because it makes explicit the trade-off between reducing climate sensitivity and preserving present-day climatological performance in this model family. However, several aspects of the methodology require clearer justification, especially the relationship between the low-resolution and higher-resolution/operational configurations, the adequacy and transparency of the evaluation metrics, and the very short coupled spin-up used during candidate screening.

We thank the reviewer for their careful review of the manuscript and for their helpful comments and suggestions. These have led to a number of clarifications and improvements to the paper. We have addressed each of the reviewer's comments below. (The reviewer's comments are in black; our responses are in blue.)

Comment 1

Consistency between high and low-resolution configurations: A central issue is that the manuscript focuses exclusively on the low-resolution N96ORCA1 configuration, while the broader GC5 model development is also linked to operational/NWP improvements and higher-resolution applications. The paper states that both this study and the companion evaluation focus only on lower-resolution simulations.

If the low-resolution configuration is intended only as an independent CMIP7 low-resolution model, this should be stated more explicitly. However, if the low-resolution model is intended to

serve as a computationally efficient surrogate for the higher-resolution or operational version of GC5, then the manuscript needs to demonstrate that the two configurations behave similarly in the relevant aspects.

We thank the reviewer for this comment. The low-resolution version (UKCM2-LL) was developed first because we intended it to be our physical model submission to the CMIP7 fast track. Consequently, the focus of both this paper and the companion evaluation paper (Bodas-Salcedo et al., in prep.) has been on documenting and assessing UKCM2-LL in its own right.

However, UKCM2-LL is also intended to inform other configurations, as the reviewer has rightly suggested. For example, a medium-resolution version (UKCM2-MM) has been developed for use in decadal forecasting, which incorporates the same physical changes as UKCM2-LL (relative to GC5); and UKCM2-LL is expected to form the basis of the physical component of the next-generation UK Earth System Model (UKESM2). Both of these are expected to be used in aspects of CMIP7 MIPs, e.g. DCPP for UKCM2-MM, but not for the fast track.

However, we view these as subsequent developments. UKCM2-LL is intended to act as a reference configuration, with further parameter tuning and/or configuration changes potentially being introduced where they are required to satisfy the requirements of those applications. Consequently, the documentation and evaluation for these other configurations will be described in dedicated papers for each configuration.

We agree with the reviewer that the original manuscript was not clear enough on these intended purposes. We will therefore include a summary of these intended purposes at the beginning of Section 5 (‘Summary and Discussion’) in the updated manuscript.

Comment 2

Evaluation metrics: There are literally tons of evaluation metrics. At the first glance, metrics used in this study looks reasonable. However, I'd like to see more justification of why these particular metrics are sufficient.

Thank you for raising this point. We agree that the original manuscript did not provide sufficient justification for the choice of metrics.

We used different sets of metrics for the two development stages. For the AMIP-based optimisation, our aim was to identify candidates with low λ_{amip} while optimising climatological performance (as far as possible). We therefore chose metrics that sampled a broad range of atmospheric processes, including thermodynamic and dynamic processes in the upper and lower troposphere; radiative fluxes including cloud radiative effects, along with an implied ocean heat transport. We also sought to minimise redundancy between metrics, so that no process was disproportionately weighted. Additionally, we used constraints on Net TOA and water fluxes because we wanted candidates that would stand a good chance of performing well in coupled experiments.

For the coupled-model development stage, the objective was different. Here, we chose a set of atmosphere, ocean, and sea-ice metrics that captured a range of key large-scale aspects of the climate system. In general, these metrics were selected either because they represent fundamental drivers of the climate system (e.g. TOA radiative fluxes), or because they provide integrated measures of large-scale climate behaviour (e.g. GMSAT, sea ice extent, AMOC, and ACC). They were used to assess the large-scale performance and stability of candidates in order to progressively narrow the set of viable candidates. The ultimate aim was not to provide an exhaustive assessment of model performance, but rather to produce a configuration with a reasonable climatology that would be suitable for more detailed evaluation, as presented in Bodas-Salcedo et al. (in prep.).

We acknowledge that in both stages these choices involved expert judgement and alternative choices could have been made, and that the development may have benefited from additional metrics (for example metrics for ENSO, monsoons or aspects of the hydrological cycle). However, increasing the number of metrics also potentially increases redundancy and, during the coupled-development stage, would make the expert-led tuning process increasingly difficult. Our choices therefore tried to strike a balance between achieving acceptable climatological performance and maintaining a tractable development process.

We will revise Sections 2.5.1 and 2.5.2 to provide more detail on the rationale behind the selection of the metrics - particularly for the coupled-assessment, where we will emphasise that these were designed to provide a broad but not exhaustive assessment of model performance.

Comment 3

44-yr spinup: It is rather short. If it is just extension of the high-res model, it is okay. This issue is connected to #1 regarding how one has to view this low-res version. Also, it seems to have a bit of drift as well. Not sure whether this can impact interepretation of the result.

We agree with the reviewer that 44 years is rather short for a spin-up in its usual capacity of allowing a simulation to approach quasi-equilibrium prior to a *picontrol* integration. However, we were using the 44-year spin-ups for a different purpose: to provide an initial screening of candidates that we felt had the least chance of fulfilling the project's aims, even with further tuning. We chose to retain the '*picontrol-spinup*' (and '*picontrol*') naming because it reflects the experimental setup (i.e. with constant pre-industrial forcing). We will update the text in Section 2.3.2 to emphasise the point that these experiments should not be interpreted as conventional *picontrol-spinup* or *picontrol* integrations intended to demonstrate near-equilibrium behaviour, but rather as short screening experiments.

In this context, 44 years was sufficient to identify candidates that were unlikely to meet the project's aims, while recognising that drifts may still be present in each simulation. To some extent, this was constrained by the resource demand of running 22 coupled simulations, and this is an area where future development exercises like this could be improved. We discuss these potential improvements in Section 5.1 (e.g. with a faster model; faster spin-up techniques; better initial selection of candidates that are closer to equilibrium from the start).

Regarding drift, it is difficult to separate drift from variability in the 200-year timeseries shown for UKCM2-LL in Fig. 4 and B1-B3 (noting that this has undergone 274 years of integration since the beginning of the proto-UKCM2 simulation) A more thorough analysis of trends over a 400-year period (starting 400 years after the end of the timeseries show in Fig. 4) has been conducted in Bodas-Salcedo et al. (in prep.). For UKCM2-LL under CMIP6 forcing, this shows moderate trends in the TOA flux components (-0.24 W m^{-2} $century^{-1}$ and 0.26 W m^{-2} $century^{-1}$ for outgoing SW and LW, respectively), suggesting the model is still spinning up. Under CMIP7 forcing (and after more integration time) these trends are smaller: -0.03 W m^{-2} $century^{-1}$ and 0.04 W m^{-2} $century^{-1}$, respectively. While these indicate that a quasi-equilibrium state has not quite yet been reached, they are comparable to values found amongst CMIP6 models, and so did not motivate any further tuning of UKCM2-LL.

Consequently, while these results suggest that a true equilibrium state had not yet been reached, we do not believe the remaining drifts are sufficiently large to alter the conclusion that we succeeded in producing a configuration that was considered suitable for submission to CMIP7.

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.