Dear editor, thank you for guiding our paper through peer-review. In the following we give a point-by-point reply to all comments from the two reviewers. The original reviewers' comments are given in regular, our answers in bold. All line numbers refer to the revised manuscript.

<u>**Reviewer #2**</u>

The authors have substantially revised the paper and I am largely satisfied with the changes. I have a few remaining comments:

L114-116: This does not make sense. It is exactly because coupled models simulate their own distinct transient evolution that a given time period will be characterized by a climate state different from each other and from observations. To do the most apples-to-apples comparison of cloud properties between model groupings and between models and observations, one should use AMIP simulations from the overlapping period. That way, differences are unambiguously attributable to model physics, not the state of the atmosphere (notably, SST pattern differences and aerosol loading differences). I am not asking you to do this, but I think this statement should be rephrased for accuracy.

The point of the reviewer is well taken. We compare, however, on purpose the historical simulations with the CERES-EBAF dataset as it is this model configuration that is being used for projections of future climate. For this reason, we are interested in the performance of the coupled system, not the atmosphere-only models as relevant errors might also be introduced by biases in the simulated ocean surface properties. Time periods not fully matching are a challenge for this kind of comparison. For multi-model climatologies, however, the exact time period is found to make very little difference. We emphasized this point by adding "when comparing long-term climatologies" to the corresponding sentence (l. 93/94):

"[…] that this difference in the time periods has very little impact on the multi-year group averages when comparing multi-year climatologies of cloud parameters."

L 167: Consider citing Zelinka et al (2022) here, as they recently investigated this connection. https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021JD035198

Thanks, we added the reference (l. 142).

L169-171: Here it is shown that high ECS models tend to have higher IWP in the midlatitudes. This is stated to be consistent with the hypothesis that improved representation of supercooled liquid leads to higher ECS. This statement is not right on two levels. Historically GCMs have had too much ice and not enough liquid, so higher IWP would typically not be considered as an improvement. Secondly, models with larger mean-state ice tend to have a larger negative phase feedback as that ice transitions to liquid with warming. In isolation, this would mean high IWP models would tend to have lower ECS.

We removed this statement.

L214-215: Is it really necessary to reproduce this statement via direct quote?

We rephrased the sentence with no direct quote (l. 149/150).

L343-345: What does this statement have to do with the results shown here? There is no evaluation of whether the models over- or under-estimate cloud feedback in stratocumulus regions. Moreover, inter-model differences in cloud feedback strength is governed by inter-model differences in representation of cloud physics, not in pattern effects (especially in response to a large warming signal like analyzed here). Suggest deleting this statement.

As suggested, we deleted this statement.

<u>Reviewer #3</u>

The authors have taken care of most of my concerns in this new version. Some minor comments remain, but should be fairly easy to address. I'm listing them below

A follow up comment to one of the authors' response:
"The increase in ECS between CMIP generations opened up a highly discussed topic about possible reasons and how realistic this is. This motivated us to look in more detail into the models. But not all CMIP6 models have higher ECS values than the CMIP5 models, there are also CMIP5 models in the high ECS group. As we were interested in whether there are systematic differences in projected cloud properties among the models contributing to differences in ECS, we sorted the CMIP models in three ECS groups. Most of the present-day evaluation with satellite data has been removed (see detailed comments above). The focus is now on projected changes in cloud properties in the models and the differences in present-day fields among the model groups that are a starting point for the following discussion of the projections."
I understand your point of view now that you have elaborated. However, you didn't change the conclusion opening and the idea that you convey in the above statement doesn't really appear in this opening.
The direct link between change in cloud properties with respect to climate change may have contributed to ECS is not done in the conclusion. I would recommend implementing some form of this sentence in it "As we were interested in whether there are systematic differences in projected cloud properties among the models contributing to differences in ECS, we sorted the CMIP models in three ECS groups."

As suggested, we rephrased parts of the first section "summary and conclusions" as follows (l. 311-313):

"Of particular interest was whether there are systematic differences in projected cloud properties among the models contributing to differences in ECS. We therefore sorted 51 CMIP5 and CMIP6 models providing the required output in three equally sized ECS groups."

Other minor comments (line numbers of the highlighted version):
L55-56: Not sure this sentence still makes sense now that the comparison with observations has been removed.

We removed the sentence.

L60: I would remove the comma after both.

Thanks for spotting this, we deleted the comma.

Cloud feedback computations: The method used to compute cloud feedbacks works well for the SW component between 60˚S/N, but it is not accurate in the polar regions and everywhere for the LW component (Shell et al., 2008). Therefore, it should be clearly mentioned that this cloud feedbacks are not adjusted for the effect of non-cloud feedbacks.

We added the following two sentences to the description of the cloud feedback calculations (l. 110-114):

"While this method is commonly used to calculated cloud feedbacks, we would like to note that the results using this method are not exactly the same as those calculated with a more accurate offline radiative transfer method (Soden et al., 2004). Particularly the shortwave CRE can be affected in regions with high surface albedos such as polar latitudes if the surface albedo changes between the two model simulations e.g. because of melting sea ice (Shell et al., 2008)."

L264-269: I don't quite follow the logic. The high ECS models produce larger IWP values over the SO and more clouds, therefore it proves that more realistic supercooled clouds lead to less negative cloud phase feedback? If anything, I would think the opposite. Higher amount of ice in the high-ECS model would suggest a potential for more negative cloud phase feedback, since the ice reservoir to be turn into liquid clouds in response to warming is larger than in the other model groups.

We removed this statement.

L275-280: Interesting to see that there is no increase in LWP over the Sc decks although there is an increase in their cloud fraction as shown by the authors, which seems to suggest that the clouds of the group are probably less bright (also consistent with Cesana et al., 2023).

We added the following two sentences (l. 177-180): "In addition to the higher cloud cover in the high ECS group in these regions the clouds seem to be less bright in comparison the two other groups. This indicates an improvement of the representation of stratocumulus clouds in the high ECS group, which is consistent with the findings of Cesana et al. (2023)."

L308-312: Could you please quantify and share the results in the supplementary?

Comparing the three group averaged climatologies for the netCRE from the historical experiments with the ones from the AMIP experiments gives the following results:
Global mean bias: 0.3 – 0.6 W/m^2
Global mean RMSE: 2.7 – 3.3 W/m^2
Pattern correlation: 0.97 – 0.98
The differences between the historical and AMIP annual mean netCRE are below 5 W/m^2 (about 5-10% relative difference) throughout most of the globe. In the ITCZ, differencescan reach up to about 10 W/m^2 (up to 10% W/m^2).

We added the following sentences to the text (l. 203-206):

"For the net cloud radiative effect, differences between the annual mean climatologies from the historical and AMIP simulations are below 5 W m^{-2} throughout most of the globe but differences in the ITCZ and tropical Atlantic can reach up to about 10 W m^{-2}. Globally averaged, the mean bias for the three group averages ranges between 0.3 and 0.6 W m^{-2}, RMSE between 2.7 and 3.3 W m^{-2} and pattern correlations between 0.97 and 0.98."

As the differences between the historical and the AMIP simulations are rather small, we do not think it is really worth putting this figure into a (then to be newly created) supplementary information.

L450-452: “in the high ECS group”, not the low, unless I don’t understand Fig. 3.

Thanks for spotting. We corrected this to “high” (l. 320).

References

Cesana, G. v., Ackerman, A. S., Črnivec, N., Pincus, R., & Chepfer, H. (2023). An observation-based method to assess tropical stratocumulus and shallow cumulus clouds and feedbacks in CMIP6 and CMIP5 models. Environmental Research Communications, 5(4). https://doi.org/10.1088/2515-7620/acc78a

Shell, K. M., Kiehl, J. T., & Shields, C. A. (2008). Using the radiative kernel technique to calculate climate feedbacks in NCAR’s Community Atmospheric Model. Journal of Climate, 21(10), 2269–2282. https://doi.org/10.1175/2007JCLI2044.1

Soden, B. J., A. J. Broccoli, and R. S. Hemler, 2004: On the Use of Cloud Forcing to Estimate Cloud Feedback. *J. Climate*, **17**, 3661–3665, https://doi.org/10.1175/1520-0442(2004)017<3661:OTUOCF>2.0.CO;2.