Dear editor, thank you for giving us the opportunity to address the reviewers' concerns. As recommended by reviewer #2, we deleted all comparisons of model results with observational data for cloud fraction, ice water path and liquid water path from the Section 3.2 (now renamed as "Present-day cloud fields") and Figures 4, 5, 6, and 9. Tables 3 and 4 have been adjusted accordingly. Table 2 and Figure 3 have been removed. In order to compare the differences in present-day cloud properties from the three ECS model groups with each other, we combined the model results (with no observations) from Figures 3, 4 and 5 into a new Figure 3 in Section 3.2.

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

In response to my concerns, the authors have added caveats to the paper that the model evaluation of cloud properties is qualitative rather than quantitative. However, they largely leave Section 3.2 (Evaluation of Cloud Properties) unchanged. In this section, numerical values of regional averages are compared between models and observations, global mean values are compared between models and observations, root mean square differences between models and observations are computed, and spatial pattern correlations between models and observations are computed, in all cases for the three groups of models. Hence Section 3.2 remains almost entirely a quantitative evaluation of modeled cloud properties against observations.

We thank Reviewer #2 for helping us to improve the manuscript. We addressed all comments in the revised version and in our point-by-point answers below (given in bold). If not otherwise noted, all line numbers refer to the revised manuscript.

As recommended, we removed all comparisons of model results with observational estimates for cloud fraction, liquid water path and ice water path in Section 3.2 and throughout the rest of manuscript including all figures and tables (see answer above).

The authors acknowledge "This study is therefore not possible with the available CMIP model output generated by satellite simulators." I agree that rigorously evaluating models against observations is not possible across the whole suite of CMIP5 and CMIP6 models as attempted here because most models lack COSP output that makes for a meaningful and reliable comparison. This is why I originally raised this concern, and I remain confused as to why the authors continued down this path. Moreover, I disagree that one can "qualitatively" evaluate models against observations while relying on comparing fundamentally different quantities. This could be misleading at best and simply wrong at worst. Therefore I don't believe the authors can confidently say make statements like "we found that models with a high climate sensitivity typically have a better representation of observed cloud properties than models with a low or medium ECS." I recommend removing Section 3.2 entirely, or focusing only on fields for which model-observations differences in the definition of the field can be minimized.

We deleted now all comparisons of the cloud properties cloud fraction and ice and liquid water path with satellite observations in Section 3.2 and throughout the rest of the manuscript. Only a comparison of TOA cloud radiative effects with CERES-EBAF remains as this dataset has been "developed to be compared directly with climate model results without the need for simulators or other sampling strategies" (CERES-EBAF Expert Developer Guidance). Our statement on a better representation of cloud properties by high ECS models has been deleted from the abstract and

strictly restricted to results from the comparison with CERES-EBAF data in Section 3.2. A new figure 3 in Section 3.2 now compares the ECS group mean climatologies for these cloud properties with each other only (no satellite observations) to assess systematic differences among the three model groups in today's climate before looking into projected changes.

My concern about comparing models and observations from different time periods remains, but is secondary to the more fundamental concern above. Arguably the least ambiguous way of comparing models and observations would be to use COSP output from atmosphere-only AMIP simulations in which the observed radiative forcings, SSTs, and sea ice concentrations are prescribed, and choosing the period of perfect temporal overlap between models and observations.

We see the point of the reviewer. Since we are comparing only multi-year climatologies, however, the exact time period as well as whether SSTs and sea ice concentrations are prescribed (AMIP runs) or coupled online to the model (historical runs) have very little influence. This is demonstrated in Figure 1 below showing the AMIP output for the same time period (1985-2004) as the historical runs of most models considered in the paper (AMIP output was not available from all models). The differences between these two climatologies are typically quite small. This confirms findings of previous studies (e.g. Lauer et al., 2023) that in general, the skill of CMIP5 and CMIP6 AMIP multi-model means in reproducing the observed cloud climatologies in terms of global means, biases, pattern correlations and RMSDs does not systematically differ from the ones obtained from the historical simulations. As it is the coupled model configuration used in the historical runs that is used for the climate projections analysed in this paper, we decided to keep the analysis of the historical model runs.

The time period for the models (1985-2004) has been chosen to maximize the overlap of the 20-year periods from different generations of models (CMIP5 and CMIP6). While this choice of model years is somewhat arbitrary, we found that it has very little impact on the multi-year multi-model averages. This is not surprising as ESMs are not expected to reproduce the exact observed phase of climate modes largely controlling present-day variability of clouds but rather their statistical properties.

For clarification, we added the following paragraph to Section 3.2 (l. 196-200):

"Biases in simulated sea surface temperatures (SSTs) can affect simulated cloud properties. We therefore also analyzed some results from AMIP simulations that use the atmosphere components of the CMIP models and for which SSTs and sea ice concentrations from observations are prescribed (not shown). Similar to previous studies (e.g. Lauer and Hamilton, 2013; Lauer et al., 2023), we found rather little differences in the multi-year climatologies of cloud properties from the models between the historical and AMIP runs analyzed here."

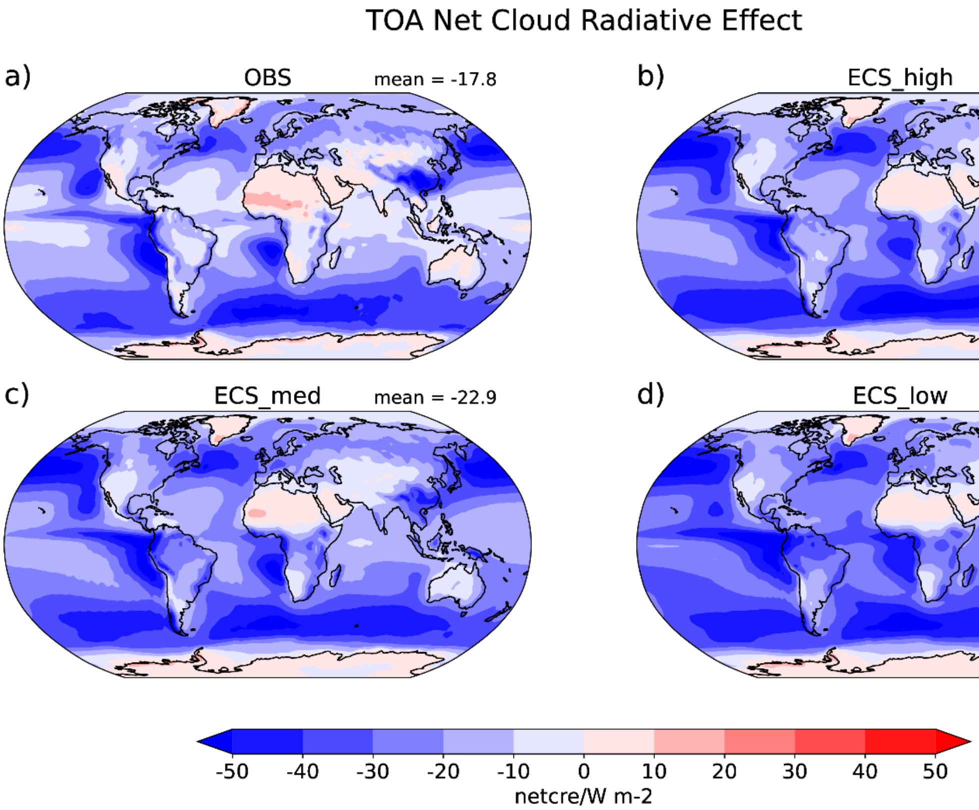

Fig. 1: Geographical map of the multi-year annual mean net cloud radiative effect from (a) CERES EBAF Ed4.2 (OBS) and (b,c,d) group means of AMIP simulations (1985-2004) from all three ECS groups.

I do not have a problem with Sections 3.1 or 3.3, though I am still not sure we learn much from these sections that are not already well established. If Section 3.2 were removed or trimmed to focus only on fields that can rigorously be compared between models and observations, I could support publication of this paper.

We significantly shortened Section 3.2 by deleting all comparisons of model results for cloud fraction, liquid water path and ice water path with observational data. We only kept a comparison with TOA radiative fluxes from CERES-EBAF Ed. 4.2, that have been specifically tailored for climate model evaluation as suggested in the first round of reviews.

<u>Reviewer #3</u>

Review of “Cloud properties and their projected changes in CMIP models with

low/medium/high climate sensitivity” by L. Bock and A. Lauer.

In this study, the authors classify outputs of 51 CMIP5 and CMIP6 models into low, medium and high ECS groups and then compare them with observations. They further look at the change in cloud properties between historical and 4xCO2 simulations weighted by global mean surface warming. They find that the models from the high ECS group better represent the observed climatology of cloud-related variables and have different sensitivities to warming than the low- and medium-ECS groups.

The topic of this paper aligns well with the scope of the journal. While I recognize the value of investigating the response of clouds to warming and the amount of work it takes in terms of data processing, I find that the study suffers from two major flaws: a non-consistent direct comparison of model cloud-related outputs with observations and the use of coupled historical simulations in the model evaluation. Also, it seems that the authors have already evaluated these models in a separate paper, so I see little value in doing this again. However, I find the analysis of the cloud response to climate change very interesting. More details are given below.

We also thank Reviewer #3 for the constructive comments that helped improving the manuscript. We addressed all comments in the revised version and in our point-by-point answers below (given in bold). If not otherwise noted, all line numbers refer to the revised manuscript.

Main comments:

My biggest concern is the comparison of cloud-related fields with observations, which doesn't account for observational uncertainties and inherent limitations of the satellite instruments. The LWP products suffer from large uncertainties (sometimes several times greater than the observed value itself, Lebsock and Su, 2014; Elsaesser et al., 2017) and cannot be used to assess models on a global scale. IWP products seem to more reliable but there is still the question of whether precipitation is accounted for or not (e.g., Li et al., 2014). The cloud fraction also cannot be compared directly to observations because of the instrument limitations and the difference in cloud definitions between models and observations. I'm attaching a figure showing the impact of using ISCCP (basically AVHRR), MODIS and CALIPSO simulators on the original output of the model for 3 CMIP6 models. The differences are very large, region dependent and model dependent…

Following the suggestion of reviewer #2, we deleted all comparisons of model results with satellite data for total cloud fraction and ice and liquid water path throughout the manuscript including all figures and tables (see details above). Section 3.2 has been shortened and now focuses on a discussion of the differences in present-day cloud properties between the three ECS model groups.

The second main concern is the comparison of historical simulations with present-day observations that have not the same surface forcings. The SST pattern and magnitude, which have strong impact on all the variables that are studied here, are not well reproduced by the coupled models as shown in the literature (e.g., Seager et al., 2019). Therefore, it is not a fair comparison. Instead, the authors should use AMIP type simulations to assess the models.

We agree with the reviewer that biases in simulated sea surface temperatures and sea ice concentrations can (and do) affect simulated cloud properties. Regarding multi-year multi-model annual mean climatologies, however, the differences between historical and AMIP simulations are found to be rather small. The same is true for the exact time period chosen for the models that we found to have very little impact on the multi-year group averages. For details and the extension of the text please see our answer to reviewer #2 and Figure 1.

Another main comment, which could be easily fixed, is the conclusion. Except for the last paragraph, which is very insightful, the conclusion is far too long (2 pages) and does not summarize the results but rather re-state them without any apparent structure.

As recommended, we condensed the section summary and conclusions by shortening the summary part considerably.

Minor comments:

Almost no information about the observations used is given and including potential uncertainties, which are raised here and there in the manuscript but without being formally quantified. As is, it looks like the authors have very little knowledge about the observations they're using.

In the revised version of the manuscript, the only observational dataset used is from CERES-EBAF. We extended the description of CERES-EBAF in Sect. 2.2 providing more details on the dataset including uncertainty estimates for the net cloud radiative effect used in this study.

I couldn't find a clear definition of how the feedbacks are computed.

We extended the description of the cloud feedback calculation in the paper as follows: "The net cloud feedback is defined as change in the sum of shortwave and longwave cloud radiative effects at the top of the atmosphere (TOA) per degree of surface warming (2-m temperature) calculated as the difference between abrupt-4×CO2 simulations and the corresponding piControl simulations. The TOA shortwave and longwave cloud radiative effects are calculated as the differences between the respective TOA all-sky and clear-sky radiative fluxes." (l. 108-111)

The introduction is not doing the best job at motivating the ECS discrimination. If the idea is that larger cloud feedback could be related to mean state cloud properties, then I would classify the models by the global mean feedback rather than ECS, because the ECS is affected by other feedbacks than those from clouds. The way it is presented in the paper is even slightly backward in my opinion.

In this study is intended as a contribution to the question whether there are systematic differences among CMIP models with different ECS. As cloud feedbacks are an important contribution to the modeled ECS, we focus on projected changes in cloud properties in the models. Sorting the models by simulated cloud feedbacks would make any conclusions of the differences found for ECS more indirect and more difficult to assess. We made this point clearer by adding the following sentence to the introduction (l. 59-62):

"A number of different feedbacks are relevant to ECS with cloud feedbacks being an important contribution. In order to assess whether there are systematic differences in simulated cloud properties among model with different ECS, we compare the simulated cloud properties from three groups of models sorted by their ECS values and quantify how the projected changes in cloud properties and cloud radiative effects differ."

I question the usefulness of having 2 to 3 versions of a model with the same atmospheric component especially when it comes to evaluating atmospheric quantities. They have disproportional impact on the mean. This question is not specific to that study though.

We agree that there are several models with a similar or even the same atmospheric component which might skew the multi-model means. To our knowledge, however, there is no established general way of considering model inter-dependence in calculating multi-model means, which is probably the reason why multi-model studies typically do not consider a weighing of individual models. Our study is no exception here.

L129-130: then why is there no distinction between CMIP5 and CMIP6 models?

The differences between the ECS groups are typically larger than the differences between CMIP5 and CMIP6. A comparison of cloud properties from CMIP5 and CMIP6 is part of the Lauer et al. (2023) paper and thus not repeated here.

L136-139: I'm not sure what this means. There can be ice clouds at all latitudes in the high levels. This latitude loosely corresponds to where clouds can be mixed-phase cloud almost year-round. These clouds show different feedbacks from the warm clouds.

We actually meant the presence of ice in the clouds in present-day climate, for which the cloud phase feedback leads to an overall negative net cloud feedback. We clarified this in the revised text (l. 126-129):

"[…] corresponds to the latitude region where a change from clouds with an ice component in the piControl simulations to clouds consisting almost entirely of liquid droplets in the abrupt-4xCO2 experiment (cloud phase feedback) starts to contribute significantly to the total cloud feedback (Ceppi et al., 2017)."

L158: simulated is written two times in a row and I don't think that Z20 is saying or showing such a conclusion in their study.

Thanks for spotting this. We deleted one "simulated" and replaced "mean state" with "representation of clouds" and "correlated" with "related" to match exactly the statement of Zelinka et al. (2020): "The representation of cloud properties in ESMs is related to the simulated cloud feedback (Zelinka et al., 2020)." (l. 139).

Fig. 3: Aside from the non-consistent model-to-observations comparison, the spread and SD appear to be very similar between the group, so the differences are not significant and there is no clear systematic behavior among the groups.

We removed this figure as we deleted the comparisons with observations for the cloud properties as suggested by reviewer #2.

L198-199: I don't understand this sentence.

We removed this sentence as we deleted the comparisons of the model results with observations.

L243: not the number on the figure, so I suppose the authors decided to switch the default dataset to CERES during the first round of review, but failed to revise the text.

Thanks for spotting this. This has been corrected in the revised version of the manuscript.

I don't see the added value of Fig. 10 and 11 compared to Fig. 9. I think that Fig. 9 and the analysis associated to it is the best part of this manuscript and should be the focus of this study.

We think that the different cloud regimes shown in Figures 10 and 11 do have an additional value compared to the zonal means shown in Figure 9 as different cloud types react differently to warming. This behaviour is easily masked in the zonal means averaging over different cloud types. Additionally, the model spread and the uncertainties help with assessing which differences are significant.

Each main Sc decks is singled out yet no there is no motivation for doing this, I'm also not a fan of picking fixed regions to study cloud response to warming since these decks can evolve in terms of location.

As the qualitative differences between the different stratocumulus decks are rather small, we combined the three stratocumulus regions in Figure 11 into a single panel that is now included in the new Figure 7. The simplification of using fixed regions seemed fine to us given that the spread among models in e.g. the stratocumulus cloud cover is rather large and thus the exact region is probably of secondary importance.

The beginning of the conclusion is confusing. The authors argue that the increase of ECS between CMIP generation motivated them to investigate the response of clouds to climate change, yet most of the study is based on present-day climate evaluation and they do not segregate between CMIP5 and CMIP6.

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.

Line 382: Z20 do not say this, instead they argue that this is a possibility that should be investigated.

Statement has been removed.

References:

Elsaesser, G.S., C.W. O'Dell, M.D. Lebsock, R. Bennartz, and T.J. Greenwald, 2017: The Multi-Sensor Advanced Climatology of Liquid Water Path (MAC-LWP). J. Climate, 30, no. 24, 10193-10210, doi:10.1175/JCLI-D-16-0902.1.

Lebsock, M., and Su, H. (2014), Application of active spaceborne remote sensing for understanding biases between passive cloud water path retrievals, J. Geophys. Res. Atmos., 119, 8962–8979, doi:10.1002/2014JD021568.

Li, J.-L. F., Forbes, R. M., Waliser, D. E., Stephens, G., and Lee, S. (2014), Characterizing the radiative impacts of precipitating snow in the ECMWF Integrated Forecast System global model, J. Geophys. Res. Atmos., 119, 9626–9637, doi:10.1002/2014JD021450.