Review of "Machine learning of cloud types shows higher climate sensitivity is associated with lower cloud biases"
by Kuma et al
MS No.: acp-2022-184

Summary
The authors train a supervised deep convolutional artificial neural network to predict the frequency of occurrence of four human-observed cloud types based on the CERES-measured top of atmosphere longwave and shortwave radiation fields over a large 4000 km x 4000 km region. After validating its ability to reproduce the observed cloud types on data withheld during training and comparing their results with an independent cloud categorization analysis, they apply the algorithm to climate model output, thereby allowing them to evaluate the fidelity with which models simulate the various cloud type occurrence frequencies. Model skill in simulating these current-climate cloud occurrences is assessed in light of the model climate sensitivities (ECS and TCR) and cloud feedbacks, and it is found that more sensitive models tend to have smaller mean-state cloud errors, although the cloud feedback shows little relationship with mean-state cloud errors. The authors argue that the most likely explanation for their results is that high ECS is plausible, in contrast to recent expert assessments (Sherwood et al. 2020; Masson-Delmotte et al. 2021)

I find the paper's overall goal to be interesting and worthwhile, but I had substantial difficultly following it in several sections and I believe the authors need to be a little more circumspect with the interpretation of the results. I recommend major revisions, as detailed below.

Comments

1. I had a very hard time following what was done in setting up, training, and validating the ANN. This includes both understanding it at the conceptual level and in many of the details. I think the authors need to begin Section 2.2 with a "30,000 foot" view of what they are trying to do, namely, predict the frequency of occurrence of 4 WMO cloud types within a 4000 km x 4000 km box based on the (spatial pattern of?) TOA radiative fluxes observed within that box. I think some interpretation of what information the ANN is learning from is needed. Is it the spatial pattern / orientation of SW and LW radiation within the region, the regional-average values, or something else entirely that provides the needed information? Are the SW and LW information equally important or does one band provide most of the information? Why is a deep convolutional artificial neural network needed in the first place; what is it providing that simpler methods would fail to yield? Table 2 and Algorithm 1 are utterly incomprehensible to me, and probably to a majority of readers of this journal. These details should probably go in a supplementary materials or an appendix, and they should be replaced with something more like schematics that give a sense of the basic workflow, how the ANN is set-up, how the data is split among training and testing, etc. All of these details are hard to extract from the paper.
2. Is it wise to use random selection for splitting the training and testing subsets? I would assume that there could be some autocorrelation in the data that would cause such an approach to overstate the skill relative to a situation in which, say, the (chronologically) first 80% of the data is used as training and the last 20% of the data is used for testing, etc.
3. What does it mean that the ANN explains 47% of the variance? Variance in what? And where does this number come from; is it in a figure? On line 407, it is stated that this number is "relative to an uninformative predictor" but elsewhere it is not stated relative to anything. How do I interpret this?
4. Figures 4 and 5: "Stratiform" is misspelled

5. Section 3.3.: How did you calculate these joint histograms? I can't find any details on how this is done or what data is used in the methods section.
6. Section 3.4: I am completely lost in this section on comparing to MODIS and ISCCP cloud clusters. How is the ANN now being generated on a 5x5 degree grid when previously the highest resolution is 4000 km x 4000 km? What exactly are you showing in Figure 7? I assume the reader has to be familiar with Schuddeboom et al. (2018) and McDonald and Parsons (2018) to understand this, but how many readers will be? I read this several times and I simply cannot wrap my head around what is being done here, other than a vague sanity check that what the ANN calls "high" , "stratiform", etc. is consistent with independent cloud clustering methods. I recommend a complete re-write of this section keeping in mind that the average reader is not familiar with these other studies.
7. Section 3.5: I do not see any justification for regressing the observed cloud types on global mean surface temperature during the brief CERES record that (1) it is likely dominated by internal variability that is not directly relevant to the long-term cloud feedback and (2) likely includes effects related to changes in aerosols and other non-CO2 forcings. Placing these results side-by-side with the abrupt-4xCO2 cloud changes is misleading and not a robust evaluation of models. You already note that this is not a "reliable observational reference", so I wonder why you did it. A more minor point: the abrupt-4xCO2 simulation does not occur during a particular time in the historical record (noted here as 1850-1949) but rather to an arbitrary 150-year period whenever the modeling center decided to branch from its piControl simulation. So I think you meant to say simply that you used data from the 150-year experiment.
8. Bayes factors: Maybe I am just ignorant, but this is the first time I had ever seen these numbers for significance testing. Perhaps other readers will also be clueless about what these numbers mean. Please provide some brief explanation when these first appear in the text, and discuss in more detail their meaning in the appendix.
9. Lines 321-323: The choice to report the ECS values for RMSE values below an arbitrary threshold (2.4%) is a little egregious, given that if the threshold for what is considered "low" RMSE is relaxed only slightly (to say 5%), the entire range of ECS values is now supported. Ditto for the "high" RMSE values: If you take all models with higher-than-average or even probably the models in the top 10 percentile of RMSE, you will include the high-ECS CanESM5 model.
10. Figure 10b: I wonder how large the Bayes factor would be if models from the same modeling center were averaged together before computing significance. Would the relationships derived in the paper between cloud occurrence RMSE and ECS remain so strong once the 3 UKMO models, 2 CNRM models, 2 INMCM models, 3 IPSL models, and 3 MPI models are combined? This would represent a substantial decrease in sample size from 18 to 10.
11. Lines 337-340: I find this reasoning for why simulating good mean-state clouds should translate to simulating good clouds in a future warmed state to be dubious. It is likely that clouds will inhabit an environment with different conditions in the future (e.g., one with higher SSTs, stronger inversions, and a sharper moisture contrast between the boundary layer and free-troposphere) – refer to the cloud controlling factor literature (Bretherton 2015; Klein et al. 2017).
12. Line 347: "lower ECS" is a little misleading, as I think you mean lower than the highest ECS values in CMIP models, but not lower than, say, the canonical IPCC range.
13. Lines 353-354: In my opinion the simplest / most likely explanation is neither of these, but rather that you are looking at a very small sample size of models (especially once you combine closely related models from the same center) and spurious correlations can occur. Perhaps more importantly, I am led to doubt the robustness of the correlation with ECS because the correlation with cloud feedback is poor (Figure 10d). How are we to believe that accurately simulating mean-state clouds translates to a better representation of ECS if the most obvious intermediary (cloud feedback) shows no relationship with mean-state cloud quality?

14. Lines 400-404: I don't understand what is being suggested here or how it could be used in concert with the techniques employed in this study.

15. Lines 432-435: My read of Zelinka et al (2022) is that the quality of present-day cloud representation has very little bearing on the quality of its cloud feedback (see their Figure 4b). Seems worth mentioning this, rather than the weaker statement that it is an open question.

16. Lines 435-440: Somewhere around here it may bear mentioning the notion that emergent constraints based on mean-state climatological observables (like the occurrence frequency of the 4 cloud types in this study) are generally less useful or robust than those that narrow in on processes relevant to the climate change phenomenon of interest (Klein and Hall 2015; Hall et al. 2019)

References

Bretherton, C. S., 2015: Insights into low-latitude cloud feedbacks from high-resolution models. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*, **373**, https://doi.org/10.1098/rsta.2014.0415.

Hall, A., P. Cox, C. Huntingford, and S. Klein, 2019: Progressing emergent constraints on future climate change. *Nature Climate Change*, **9**, 269–278, https://doi.org/10.1038/s41558-019-0436-6.

Klein, S. A., and A. Hall, 2015: Emergent Constraints for Cloud Feedbacks. *Current Climate Change Reports*, **1**, 276–287, https://doi.org/10.1007/s40641-015-0027-1.

——, ——, J. R. Norris, and R. Pincus, 2017: Low-Cloud Feedbacks from Cloud-Controlling Factors: A Review. *Surveys in Geophysics*, https://doi.org/10.1007/s10712-017-9433-3.

Masson-Delmotte, V., and Coauthors, eds., 2021: *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press,.

Sherwood, S. C., and Coauthors, 2020: An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence. *Reviews of Geophysics*, **58**, e2019RG000678, https://doi.org/10.1029/2019RG000678.