**Review of "Machine learning of cloud types in satellite observations and climate models"**
**by Kuma et al**
**MS No.: acp-2022-184**

**Summary**
The authors have revised the manuscript to address the reviewers' previous comments, which has generally improved the paper. This includes dialing back the conclusions regarding implications for climate sensitivity, adding much more detail to help readers understand the methodology, and performing the analysis at a pixel-by-pixel level. However, there remain several things that need to be addressed before the paper is suitable for publication.

**Major Comments**

- I am still quite unclear on how the ANN works. The input TOA flux data are used to predict the probability of the 4 cloud types at each pixel in each 16x16 pixel domain, but the ground truth labels in most cases will only occupy a small portion of the domain where the IDD stations are. How can the pixels with no ground truth learn anything from the CERES TOA fluxes?

- I don't understand what is meant by the "application phase." Does this refer to the phase when you deploy the trained ANN to make predictions on unseen data? If so, then why does it only use 20 random samples per day rather than all of the TOA radiation data?

- Why are there no figures demonstrating the skill of the ANN in predicting unseen data? I see that the ANN is trained on CERES and IDD data in years 2004, 2005, 2007 and 2009–2017, with years 2007, 2012 and 2017 used as a validation dataset. My understanding of how ML studies are typically done is that the data is split into three categories: training, validation, and testing. It appears as though here you have used all of the data for training and validation, but did not reserve some data for doing out-of-sample testing. How can we be sure that the ANN works well on unseen data and has not over-fit to the training data?

- All of the analysis is basically in frequency of occurrence space rather than in within-regime cloud property space. But surely the latter should be a large part of the story. A model could for example get the frequency of occurrence of each regime perfectly right but the cloud properties (cloud fraction, albedo, altitude) within the 4 regimes could be biased. Is there a reason that within-regime cloud properties are not evaluated as well as frequencies of occurrence?

**Minor Comments**

L8: delete "a" before "top"

L55: there are issues with subject-verb agreement ("they…is…has…")

L99: I don't understand this statement about grouping together multiple cloud genera, since throughout the paper the results for 10- and 27-type classifications are also shown.

Section 2: I dislike the organization here. It goes from Methods description (Section 2.1) to Data used (Section 2.2) back to Methods description (Section 2.3). Why not put the Data section first?

Figure 2: Suggest calling the radiation fields what they actually are (normalized CRE) rather than "reflected TOA radiation" (colorbar) or "shortwave and longwave radiation (caption). Do the 4 cloud type maps have to sum to 100%? It doesn't look like this is the case. Is there a clear-sky probability?

L291: "histograms" should be singular

L333-334: Why are you reporting the years like this? It doesn't make any sense, as I noted in my previous review.

Figures 18-19: These are completely ineffective and uninformative figures that should be removed. Are responses of the individual 27 cloud types really trustworthy? Even if they are, is examining responses with this level of granularity bringing any new any insights? I doubt it.

L345: "the the"

L358: What is the P value for high clouds? Which value of P marks the transition from "statistically identifiable" to "not statistically identifiable"? In the previous paper, the Bayes factor (ratio of the two probabilities) was reported, but now just the probability of the null hypothesis is reported. Is there a reason for this?

L397-414: I don't see the value in this discussion. All of the options discussed would still not allow for an unambiguous estimate of the response of clouds to global warming relevant for ECS. This is because of spurious trends in the datasets, the influence of factors other than just global warming (aerosols, most notably) on the trends, the fact that the observed warming pattern over the satellite period is very different from that expected in response to CO2, and other things. I am not aware of any TOA flux measurements on MISR, MODIS, CloudSat, or Calipso, so I'm not sure why those would be used instead of a radiometer like CERES. The idea of running a COSP simulator in a model to generate fields to be run through an ANN to tell you about cloud types seems really bizarre since COSP is already providing detailed information about cloud types. I suggest deleting this paragraph.

L428: I'm confused. Zelinka et al call what an open question? The previous sentence just looks like a statement of fact – that better present-day cloud properties is associated with larger cloud feedback, similar to what is found in this study. What is the question? In the next line, is it really necessary to directly quote that paper (Zelinka et al) for a fairly mundane statement summarizing the results from another paper (Tsushima et al)? Usually quotes would be reserved for something where the exact phrasing is vital or compelling.

L438: I believe the correct citation is "Jiménez-de-la-Cuesta" and Mauritsen (2019)

L343-445: This paragraph seems to be all over the place and it is not clear what point you are trying to make.

Appendix A: suggest telling the reader what values of P or Bayes factor represent statistical significance (e.g., something analogous to p values being less than 0.05 for a statistically significant result at 95% confidence)

Figure B1: I still don't really know what I am looking at here. Is there a way of showing the reader what "perfect" validation looks like? I have no idea what "right" or "wrong" looks like.