Atmospheric Chemistry and Physics Discussions

Atmos. Chem. Phys. Discuss., referee comment RC1
https://doi.org/10.5194/acp-2022-184-RC1, 2022

Comment on acp-2022-184

Steven Sherwood (Referee)

Referee comment on "Machine learning of cloud types in satellite observations and climate models" by Peter Kuma et al., Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2022-184-RC1, 2022

General Comments

This paper uses an artificial neural network (ANN) to learn cloud populations (via four cloud types) from top-of-atmosphere LW and SW cloud-radiative effect, demonstrating moderate skill. It then uses this as a metric for comparing climate model cloud fields to observations, by computing RMS error between the satellite-derived and model-derived distributions of the cloud types. This is in principle a very interesting idea, because it provides an arguably more objective way of identifying observational metrics. The authors report a decreasing trend in RMS error with increasing climate sensitivity, which they argue implies that ECS is high in spite of recent assessments finding this to be very unlikely.

I believe this work requires major revisions before it should be published, and I'm not sure it should be published at all unless some of these concerns can be satisfactorily overcome (or I have something wrong that just needs explaining better), which hopefully they can. The main problems are elaborated below.

Specific Comments

- The conclusions drawn are not convincing given the limited number of independent models and the modest strength of relationships seen in Fig. 10. For example if you took out the three MPI models, or the three UKMO models, in either case the correlation would become pretty weak. The authors are using a Cauchy error model which is more forgiving of outliers (of which there are several) than the more common Gaussian model, and as they do not specify otherwise I assume they are taking the models to be independent (which is not a good assumption since, as the authors themselves discuss, various models by the same centre often behave similarly). This

problem could be addressed if the authors were able to include more models (and I am surprised more cannot be used). Certainly, given that several models disobey the fit, this leaves a reasonable chance that the real world would also disobey the fit, whereas the authors seem to be tacitly assuming in their discussion that the only way ECS could be anything but very high is for the whole relationship to be perversely wrong. Having said this I do think it is useful and interesting to show that the more accurate models have higher ECS (although that has been shown by other studies using different metrics), since even if this doens't prove ECS is high, it does identify a conundrum that the modeling community needs to solve.

Also the authors should be aware of and probably cite papers such as Zhu et al. 2022 doi:10.1029/2021MS002776, who found that the NCAR model could be improved by making a change to the cloud scheme which also reduced the ECS of the model, i.e., a counterexample to the claimed relationship, or Zhao et al. 2016 (10.1175/JCLI-D-15-0191.1) who found that the ECS of the GFDL model could be changed substantially without affecting the latitudinal distribution of cloud radiative effect. Other authors (Klein, Hall, Caldwell) have pointed out that emergent constraints should be treated with much caution unless there is a mechanism linking the observable to the feedback; simply pointing out that clouds are being observed and are involved in the feedback is not a mechanism.

Finally, while the authors are entitled to their opinion on how much credibility to give their analysis vs. that of the IPCC and Sherwood et al. (which they should not call an "expert judgment" study since that implies it was based on an expert elicitation rather than an analysis of evidence), I would say it is unfairly dismissive given that those assessments quantitatively incorporated many independent lines of evidence including other emergent constraint studies not dissimilar to this one, for example Volodin et al. which is arguably based on a similar argument and dataset to the current paper (and still has some limited skill — see Schlund et al. 2020). The constraint on ECS offered by the authors is much more indirect and model-dependent than the multiple additional constraints used by the other assessments.

- I am not convinced that the ANN is behaving as expected. First, the authors have not shown any spatial maps of their verification data to compare with the maps of predicted cloud types. Second, the optical-depth/cloud-height histograms (Fig. 6) don't make any sense—they show that high clouds have the same distribution as the overall mean, but two different low-cloud types differ in opposite directions. But the high-cloud composite should show more, er, high cloud. I don't think this can be right.
- The cloud dataset is not adequately described. Are cloud amounts of each type given in oktas? Or are the clouds seen at each time simply assigned to one of the 27 categories? Or can multiple categories be assigned in a single synoptic observation, i.e., each category assigned a one or zero at each observing time? What exactly is the ANN going to predict? This needs to be given in Section 2.1.4.
- The way the ANN is employed is also not adequately explained, I'm not sure I fully understand what the authors have done, and what I do think I understand, I mostly had to piece together based on strands in the discussion of results. Also the motivation for the experiment design is not explained — what do we expect this approach to gain that was not gained by all the other efforts to classify or divide cloud scenes into different categories? You need a new section before "Results" where you explain the methodology properly. In 3.1 you don't say enough. For example you need to explain exactly what the features are and what you are trying to predict (see above comment, we don't even know the cloud states are represented via numbers). Are you (as I suspect) presenting each ~50x50 grid as single training instances? In which case the output of the ANN is a same-size grid of some measure of cloud state (Predicted cloud type, amount of each cloud type—don't even know if this is a categorical (classifier) or a real-valued (regression) target variable). Or are you prediction *how many* grid points will be assigned to each cloud type? Are any other variables used as training features,

for example surface temperature (which you said earlier you were using but I don't see where it comes in)? Or the only predictors are the grid of normalised SW and LW CREs? As I understand it, according to Fig. 1, most of these grid points will have no verification value available, only those observed by an IDD station. I assume you then retain the predictions only at those locations? Help!

- If I understand correctly, I don't think the authors are doing this in an optimal way. Based on my inferences, I believe each training instance is a tile of roughly 50x50 grid points; that the features are the gridded maps of SW and LW (normalised) CRE; and that the target variable predicted is the amount of each cloud type in the tile. This means the ANN predicts only four numbers for each tile, thus the authors can make only extremely smooth estimates (e.g. Fig. 4) with no detail at or below the tile dimension. Yet one should be able to predict high, mid and low cloud quite effectively on a *ocal* basis just from local SW and LW CRE: high LW CRE means high clouds, high SW and low LW CRE means low clouds; etc. Indeed this is routinely done and is the basic for e.g. the ISCCP cloud classification and other similar ones. It is not clear whether, or how this study has used any of the nominally available spatial information in the tile. If it is not being used then there is no point in starting with tiles, why not use each grid point (where you have a verification datum) as a training instance? If the authors do want to use spatial information (i.e. texture in the cloud field), then I would expect the authors to train on the entire global dataset using a convolutional neural network or other image processing approach that can produce detailed localised predictions rather than only producing populations accumulated over a large region. Or indeed they could use many more (and probably smaller) tiles in a standard NN and predict only the cloud properties at the centre of the tile. This would enable much more incisive testing of both the algorithm and the climate models.
- I don't find the comparison to traditional classification (Fig. 7) to be useful because of the way the data are being so severely coarse-grained (see point 4). The traditional measures are all local. The comparisons seem to suggest that their classifications have little to do with the traditional ones, which again is highly suspicious since the latter are based on robust physical arguments and should work well at least for high/med/low cloud distinction.

Technical suggestions

Section 2.1.2: I'm surprised that so few models are able to be included, since nearly all will have done the two required experiments, and yet you have fewer than half of the models. I assume that most models did not provide all of the desired radiation variables. Can you specify what was the main thing missing from the other models?

Fig. 1: caption states that panels show spectra, but a spectrum is a graph of intensity vs. wavelength. These panels show images not spectra. Also, is the GCM image also from a single day? What day? how was it chosen? It looks fairly close to the observed one so I am guessing you searched somehow for a day that was close—if so this needs to be explained.

158: I don't understand what is being assumed multivariate normal here, or even what the three dimensions are (lon/lat/time?). Why do we assume a normal distribution in time? I would have thought we were just grabbing data from each day, at uniformly (not

normally) sampled tile locations.

159: please also specify that these are TOA upward values

170: please clarify if you discarded these points both for SW and LW training, or only for SW. I don't see any reason to discard them for LW, unless it would make your ML approach inefficient to have unpaired LW values. Clouds in the polar night will have nontrivial LW forcing effects.

Fig. 3: Do I correctly interpret from this figure that the ANN is performing better on the evaluation partition than the training partition of the data? That is not usual. Maybe there is something I am missing here?

199: when you say "Cloud types in … reanalyses", I assume you mean the cloud types inferred by running the TOA radiation data through the ANN, not the actual clouds in the reanalysis? Please clarify.

201-4: This is not a complete sentence (no verb)

205: I don't think you can dismiss the error that easily. There are strat-cu decks that would not have any other cloud type, and would be large enough to be resolved. I do see them, smoothed out, but to tell if they are fully represented or not we'd need a comparison truth plot.

Fig. 4: Why doesn't this figure include a set of panels for the target (IDD) observations? Only then can we see whether the algorithm is working, no? (see point #2 above)

224: wrong citation format

Fig. 7: we need to be told what the three rows and five columns represent. No good to just cite a paper we have to go look at to find out. More generally, I am not sure this whole analysis adds much to the paper anyway (see point #6 above).

Fig. 8: The caption doesn't fully clarify what is different between the second and third rows. Is the third row the histogram of 5x5 degree averages? The others are histograms at the original scale (1 degree?)