12 April 2024

To: Andrew Sayer, AMT Editor

Dear Sir,

We have reviewed the helpful comments provided by Reviewer #2 and have made a number of changes where appropriate. Those changes are noted in our response to the reviewer and can be seen in the revised manuscript, which retains the tracked changes. We have also made several zip files of the output of our analyses so that anyone interested in reproducing the results will have a reference. The link to the zip files is given along with raw data and Mathworks in the relevant section after the conclusions and summary.

We contend that we have gone far and above the efforts of previous studies of multilayer clouds to assess the results and provide a realistic picture of what the algorithm actually does. To our knowledge, no previous paper has explicitly accounted for the what the false detections actually mean in terms of the number of multilayered clouds that are detected by a given algorithm. In other words, instead of ignoring the impact of the false detections, we have hung out our dirty laundry with the NGA parameter and the revised plots in Figs. 3 and 4. We hope that these examples will set a new standard for reporting of multilayer cloud algorithm results in future publications.

We hope that you find the responses and changes satisfactory for publication. Thanks for your consideration of our revised manuscript.

Sincerely,

Sunny Sun-Mack

Response to Review of first revision. Blue indicates our written response to each topic. Red indicates the text we have placed in the new revision.

> We appreciate the reviewer's suffering through the slog and providing us with some very good
> recommendations. We hope that the changes are satisfactory and make the trek a bit smoother.

The changes introduced are all for the better, and the trek is, just as advertised, a bit smoother.

The revised manuscript provides a rich description of successful modeling efforts that will allow departures from the single-layer, plane-parallel cloud morphology assumption to improve cloud property retrievals critical to CERES and other applications. The authors have not made a serious effort to quantify the nature of the improvement against the previous version of MLANN, but this is only one part of their report. The general performance assessment, comparisons against related works, and investigation of full-swath retrievals reflect mostly sound analyses and support the stated conclusions.

Evidence for improvement against the previous version of the MLANN is only quantitatively described by total accuracy. This reviewer has raised concerns about the utility of the accuracy metric for unbalanced classes. The authors the say that "much" of the change in accuracy is due to the new dataset, which gained increased imbalance due to the shorter CALIPSO horizontal averaging. A null model also has better accuracy if you increase the imablance in the dataset. Rather than a qualitative attribution of model improvement to the class being easier to guess blindly, this reviewer would like to know if the model is better because of improved architecture or the increased subsetting (from 2 trained models to 8 trained models).

Assuming the reviewer did not read Minnis et al. (2019), as was indicated in the first review, the reviewer must have missed the point of that formulation, which was applied only to SF surfaces and had separate training for ice and water clouds (4 models). The only change here was the addition of snow-covered surfaces, which added 4 more models to the mix. We have to admit, however, that he/she was led astray by our statement in the last paragraph of the Introduction, where we errantly noted,

"*In addition to the **separate day/night training used in previous versions**, the MLANN herein is trained separately for both snow-free and snow/ice-covered surfaces using an entire year of data.*"

That was a leftover from a cut and paste that we apparently overlooked. The separation of ice and water was a step up from the formulation of Sun-Mack et al. (2017), which only had 2 models for SF surfaces, day and night. Thus, Minnis et al. (2019) increased the MLANN accuracy with the use of the separate cloud phases. To make the progress of the MCANN, nee MLANN, clearer, we changed the last two paragraphs of the Intro to read:

"To improve the CERES ML detection, Sun-Mack et al. (2017) developed a multi-layer cloud detection ANN (MLANN) to distinguish between SL and ML clouds using MODIS radiance data matched to CALIPSO and CloudSat vertical profiles of clouds over snow-free surfaces. The MLANN was trained separately for day and night data. Minnis et al. (2019) enhanced the MLANN by including more input parameters and additional output variables such as upper layer CTH, COD, and cloud-base height (CBH). They also used only high-confidence CloudSat and CALIPSO data for training and further trained the MLANN separately for clouds identified as either ice or water phase by the CERES CM4 algorithms. Using one month of data, they found that, for nonpolar clouds, the MLANN correctly identified ML and SL clouds together 80.4% and 77.1% of the time during day and night, respectively, using CALIPSO data averaged over an 80-km distance. Those values are 5% greater in absolute terms than their earlier counterparts. While the accuracies are quite encouraging, the approach needs further refinement and complete seasonal and global coverage.

This paper reports on continued development of the MLANN to detect ML clouds. Revisions to the previous training are made using newer versions of CALIPSO and CloudSat products with constrained horizontal resolution. To be more representative of its use and avoid confusion with other machine-learning terms, the acronym for this revision is changed from MLANN to the Multilayer Cloud-detection Artificial Neural Network, MCANN. To expand coverage to the entire globe, the MCANN is trained separately for CERES ice and water cloud pixels separately for snow-free and snow/ice-covered surfaces using an entire year of data. Input to the MCANN is also enhanced with some new variables. Finally, because the MCANN is trained with near-nadir data, its utility for full swath MODIS data is examined."

The authors suggest that the F1 score handles imbalance, which could be true depending on details of scoring, but do not apply even that in their comparison. I encourage the authors to better quantify the improvement in the MLANN, and consider the subtleties of the F1 score (in particular, consider the differences between what scikit implements as "binary", "macro", and "weighted" averaging [1]).

We have now included the F1 score as well as other parameters to further bolster our claim that the new version is more accurate than the previous one. The beginning of the Discussion Section now reads:

"These results represent a significant improvement over the MLANN of Minnis et al. (2019), which only attained accuracies of 80.4% and 77.1% during the day and night, respectively over SF surfaces **using a single month of data. Over SF surfaces, PR, RC, and F1 from the MCANN are 74% (72%), 57% (52%), and 64% (60%) for daytime (nighttime), respectively. In relative terms, all of those values exceed their MLANN counterparts by 1% to 20%. The MLANN NGA values are slightly higher.** Much of the increased accuracy **of the MCANN relative to the MLANN** is due…"

The slightly higher NGA values arise simply because there is a much greater population of ML clouds than available in the current study. We believe the numbers we have listed in the latest revision are sufficient. We have also provided the relevant confusion matrices in Minnis et al. (2019) and in this paper for anyone to make whatever calculations they wish in order to further assess the relative accuracies of these two methods.

The MLANN validation results (Table 4, Figures 3-4) emphasize two metrics, the accuracy and the multi-layer cloud (MLC) fraction. You tire of me complaining about accuracy, so I will only raise issue with the MLC fraction. The MLC fraction makes a lot of sense in Figure 14, where the CC labels are not available for a pixel by pixel comparison. When the correct classes are available, you actually know how much the MLC fraction is pumped up by false positives. Consider your statement (line 308) that MLANN underestimates the MLC fraction over snow-free surfaces by 4.8% during the day; actually, a whopping 4.1% of the classifications going into the fact were false positives so the correct MLC fraction was worse underestimate. The MLC fraction based only on correct classifications could be determined and presented in Figures 3-4, or a different aggregation I've not thought of would be better.

We are glad that the reviewer is seeing the point of NGA. What is net gain? All previous published maps and zonal means use the total number of ML detections as we do in Figures 3b, 3d, and 4. Plotting what is asked is unprecedented, but certainly logical, given that we have made a fuss about NGA. Therefore, we agree and have added two additional maps to Figure 3 and new points to Figure 3 that provide only the correct fraction of ML detections. We have accompanied those changes with relevant discussion. We added the following to the penultimate paragraph in the Results section.

"Figures 3c and 3d show the ML fractions determined from all of the positive ML detections from the MCANN, whether they are correct or not. Figures 3e and 3f show the corresponding distributions of ML fractions based only on pixels that are actually correct according to the matched CC data. As expected from Table 4, the magnitudes for both

day and night are significantly reduced compared to those for all of the MCANN results. The relative distributions, of the correct values are similar to their all-MCANN counterparts."

We added the following to the end of the last paragraph of the results section.

"When only the correct MCANN values are considered (open squares), the zonal means drop further below the CC averages. The correct-MCANN differences relative to the CC values, shown at the bottom of Fig. 4, vary zonally much like their all-MCANN counterparts for both day (orange dotted line) and night (blue dotted line), but are 0.02 to 0.07 lower. The sources of these differences are discussed further below."

Table 4 is said to be based on the "training" data: I assume, but it's not explicitly clear, that this is the 20% test split.

Table 4 is based on the 2009 independent dataset. Perhaps, Table 3 was meant here. If that is the case, then we have clarified it in the text and table caption.
For the former, the starting line of the relevant paragraph now reads, "*... The results in Table 3, based on the entire 2008 training data, include the confusion matrices...*".
For the latter, the Table 3 caption reads: "*... layer identification from the entire 2008 CloudSat-CALIPSO training set...*".

The authors acknowledge that comparison against related works is contextual, and not any kind of statistical ranking, because of the differences in the datasets. I agree, and also agree it is necessary to do. The final summary of this comparison (line 676 - 678) does not reflect the fuzziness of the comparison and should be more circumspect.

To overly emphasize this concern, we have inserted the following paragraph between the first and second paragraphs of the Conclusions section:

"Direct comparisons of the MCANN to other multilayer cloud detection methods are not possible due to differences in ML cloud definitions, input satellite data, reported accuracy parameters, sampling, and cloud optical depth constraints. Nevertheless, the MCANN results were evaluated here against published results based on other techniques. Attempts were made to minimize the characteristic differences among the various results as much as possible. All conclusions drawn from those comparisons are limited by the unknown effects of the remaining differences among the methods."

I again encourage the authors to re-consider their choices of acronyms. Calling the

model the multi-layer artificial neural network (MLANN) is vague, because the name doesn't include clouds, and may mislead readers already familiar with multi-layer artificial neural networks.

In the above reviewer comments, there is much concern that we did not elaborate on the change in the accuracy from our previous version of the MLANN. In so doing, it recognizes our previous use of the acronym, MLANN. That would suggest that we maintain the MLANN for continuity. However, the argument used in this second review for a change in acronym is much better than that from the first review. “Cloud” is nowhere to be found. Thus, we agree to change the acronym to MCANN, Multilayer Cloud-detection ANN. The statement noting that change is included in the answer to one of the questions above. MLANN is retained when referring to the previous versions, because they are already published.

Funny thing about the concern previously raised with respect to the SS, MS, MM, and SM acronyms: I mistakenly referred to "recall" when I meant "precision" in my complaint because I mixed up the unfamiliar MS and SM, and the authors also mixed up MS and SM in their reply regarding the NGA. I doubt either of us would have made mistakes with the usual TN, FP, TP, and FN.

The use of SM in the reply was a typo, not a confusion of the MS and SM. The revised paper stated it correctly. Given that we have defined our variables in Table 2 and our previous response, it should be quite simple for anyone to understand what we have done.

[1]: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html