Reviewer 2

The submitted manuscript describes research towards an operational retrieval of vertical cloud structure from space-borne, imaging spectroradiometers. It continues development, by the same research team, of the MLANN algorithm: the present iteration 1) further subsets the algorithm (which already separately models day and night observations) to separately model observations from snow-or-ice covered and snow-or-ice free surfaces, and 2) uses updated labeled datasets. For the first time, the authors also consider the application of MLANN (which is trained on near-nadir viewing angles) to full swaths (i.e. including far-from-nadir viewing angles).

The manuscript is lengthy and often poorly organized, making a comprehensive list of revisions, sufficient to merit publication, too time-consuming to generate. I've listed several of my key concerns with the manuscript below, along with several examples of unnecessary additions to the manuscript's "cognitive load" which, if reduced, would allow this reviewer to complete a comprehensive list of recommendations.

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.

1. The discussion begins with an unsupported statement: "These results represent a significant improvement over the previous MLANN formulation." It's important that a revised manuscript include evidence supporting this conclusion. If this iteration of the MLANN has not shown improvements, then an entirely different manuscript crafted around a negative result is needed. While I'm sure that is not the case, where is the comparison? Section 5.2 and Table 5 have neglected that critical comparison.

The accuracies of the 2019 MLANN were originally reported in the penultimate paragraph of the Introduction. Apparently, the connection between Discussion statement and an implied comparison of the accuracies in Table 3 with those for 2019 in the Introduction was too subtle. To reduce cognitive distress, the discussion now begins with a more explicit supported version of that statement, as follows.

"These results represent a significant improvement over the previous MLANN formulation, which only attained accuracies of 80.4% and 77.1% during the day and night, respectively over SF surfaces."

2. Sections 2 and 3 are not laid out in a way that helps the reader. ANNs are data-driven, so presenting the model before the data demands the reader maintain abstractions "x" and "y" until the next section. I recommend reversing the presentation, so the reader knows the inputs and outputs. The "Input and Output Layers" subsection of 3 includes aspects of methodology (data splits, number of models trained) and feature definition (cloud layer classification) that don't have anything to do with the subsection heading.

Thanks for the comment. We have rearranged and rewritten the data and methodology sections.

3. **The accuracy metric featured foremost in the abstract and conclusion is sensitive to imbalance in the test data, which is nearly 80:20 in this manuscript. The null model for an 80:20 split also already has 68% accuracy (not the 50% naively assumed for a binary classification problem). You might consider using instead the Mathews correlation coefficient. You definitely do not want to compare, as in Table 5, the accuracy of models with different degrees of imbalance in the test data.**

We have added in the F1 score, which is also valuable for unbalanced binary data. It was also employed by Tan et al. (2022) and others (e.g., Haynes et al., 2022), and is easily computed from the data of Desmons et al. (2019). We believe it is a necessary exercise to perform the comparisons in Table 5 to put the MLANN results in context. The differences among the methods and datasets are clearly outlined in the text. We prefer to leave it to the reader to gain a sense of how well the MLANN compares to other published methods.

4. Terminology and acronyms are abundant and sometimes more confusing than helpful. One example: does "multi-layer" in MLANN refer to neural network layers or cloud layers? MLNN is a widely used acronym for neural networks with multiple layers of nodes, so most readers will associate the ML in MLANN with nodes not clouds.

As to the choice of the acronym, we have to disagree. The acronym is clearly defined and contains “A”, a distinct difference from MLNN. We expect that most readers of this paper are focused on the multilayer (ML) cloud component; multilayer clouds are in the title. We doubt there are that many readers of AMT familiar with MLNN as a reference to neural network layers. As a test of that contention, we looked at the 30 most recent AMT papers containing neural network in the title and none of them contained the acronym MLNN. We have removed the parameter, fraction correct or FC, to reduce the acronym load a bit. Its inclusion with its equivalent, ACC, was meant to make easy reference to the previous studies, but only added to the load.

A second example: the MS, SM, MM, SS acronyms for the confusion matrix are unnecessary departures from the true positive (TP), false positive (FP), etc. terminology widely used for binary classification problems. You could eliminate a whole Table by ditching MS, SM, MM, and SS in favor of conventional terminology.

Having slogged through much of the literature on using machine-learning methods for clouds ourselves, we encountered a variety of variable names employed for the confusion matrix. It is very confusing. We found that the MS, SM, MM, and SS set of Tan et al. (2022) was straightforward and explained at a glance what each number meant. We thought it would be suitable for multilayer detection. Apparently it is not suitable for everyone. Having gone back and forth on this, we opt to retain this terminology and Table 1 to ensure clarity. We have clearly defined what each variable is and find that it no more of burden than trying to remember if TP refers to ML or SL agreement..

A third example: what is the purpose of inventing "net gain of accuracy" when recall (of the multi-layer class) already quantifies MM with respect to MS and is both familiar and normalized?

Good question. The goal of this algorithm development is to detect as many multilayered clouds as possible while minimizing the number of false ML clouds as we stated in the text. Why? Certainly, not for the sake of doing it. It is simply the first step in characterizing the clouds within a column in terms of their macrophysical (vertical structure: height, thickness, layering), and microphysical and optical (phase, particle size, optical depth) properties. Having to concoct ML clouds and their properties out of whole cloth for a false ML pixel, when there is only one layer present, is certainly as much of an error as identifying a ML cloud as a SL cloud, the default case.
Because recall is a normalized parameter, it can yield the same value for different levels of multilayer cloud detection. Similarly, its counterpart, precision, is also normalized. As such, it is the same whether MM = 3 and MS =1 or MM = 30 and SM = 10. It does not tell us, to use an American football analogy, whether we moved the ball 2 yards or 20 yards toward the goal. We could not find an appropriate straightforward metric from among the standard parameters used in confusion matrix analysis, so we defined one, net gain in accuracy, NGA, for the purposes of this analysis. It represents the net effect of the effort. Additionally, it serves as a means of comparison with some other methods that may use different unbalanced test data. It is the only new parameter that we have introduced.

A fourth example: CoS is a new term and acronym for precision of the single-layer class. (At least, it is if I'm correct that the "1 - " in equation 6 is a mistake.) Every new term or acronym, especially ones close to but deviating from familiar ones, increases the cognitive load the reader must maintain while interpreting the study.

You are correct, the “1 -“ was a typo. CoS is not our term, but was used by Desmond et al. (2017). It was employed to facilitate comparisons with other algorithms as stated in the text and seen in Table 5. It is much more descriptive of the quantity than one of the standard confusion matrix terms.

5. The presentation of the algorithm spends too long reviewing general theory and barely touches useful specifics. I appreciate brevity allowed by references to prior papers documenting the MLANN, although I did not take time to read them too. Nevertheless, specifics that either differ, are essential for reproducibility, or simply quick to convey ought to be included (e.g. exact network shape for each model, the loss function). The description also has to be self-inconsistent, which it is not in describing 1) how "testing" and "validation" data are (both?) used to terminate training, and 2) describing y as a probability in the text but as any real number in Figure 1. One thing the cited papers may explain, but I would like to (also) see here, is the reasoning or optimization that 1) limited the ANN to a single hidden layer, and 2) allowed collinear variables as inputs (in the brightness temperatures along with brightness temperature differences).

We have rewritten the methodology section to eliminate generalities and provide specifics. Please see the revision.

The reasoning for using a single hidden layer is included in the revision:

"Only one hidden layer is used for this shallow neural network. It was found that a second layer yielded no significant increase in accuracy, but greatly increased processing time."

The reasoning for including the collinear variables is simple. Their inclusion increases the detection accuracy. Others using machine learning methods have come to the same conclusions (e.g., Tan et al., 2022; Hakansson et al. 2018). As a measure of that improvement, we have added the following paragraph to the methodology section.

"The input variables in Table 2 were selected by adding in parameters suspected of enhancing ML detectability and computing accuracies for each run. If no gain in accuracy occurred, the parameter was not used. Each predictive parameter's influence on the final MLANN formulation was assessed by computing the relative decrease in recall when a given parameter was removed from the training. Recall is the fraction of the CC ML pixels that were detected by the neural network.The decrease for each parameter was divided by the sum of all of the values to produce a relative ranking of importance. The ranks ranged from 0.038 for BT_{11} to 0.082 for the relative humidity profiles, which were treated as a single input for these purposes. The second highest ranked parameter is latitude, followed in the daytime by SZA and $\rho_{1.38}$. In general, the brightness temperatures were ranked lower than the BTDs, similar to the rankings reported by Tan et al. (2023) for their random forest method."

6. The optimization appears to take not precautions against local minima or overfitting. Widely used software for neural network optimization use some form of stochastic gradient descent, whereas this study uses Levenberg-Marquette. How are local minima avoided? The large numbers of parameters in neural network optimization can lead to overfitting, but the authors do not indicate any steps taken to avoid overfitting or indicate that none was observed.

Thanks for pointing out our omissions. In doing so, you enabled us to catch an error. We used the Levenberg-Marquardt function in the previous MLANN version and forgot to replace it in the current text with the scaled conjugate function. We now explain our approach to avoiding local minima and and overfitting. Because we did not use a deep-learning network, we did not employ the SGD. Instead, we took a "manual" approach, as described in the new paragraph:

"To avoid local minima in the neural network, the training runs were repeated many times using different samplings of the dataset (e.g., every 3rd pixel or every 5th pixel); different random initial weights; and various percentages for training, testing, and validation. Local minima were identified when the training convergence time was abnormally short or long. Overfitting was avoided by using a very large dataset (typically more than a million datapoints), which forces the net to generalize. It was also avoided by using a minimal number of neurons. Additionally, unreasonable data, such as fill values, were filtered out to minimize the noise. A set of range limits was used to eliminate any obviously errant data. Leaving such data in the input set prevents the training from generalizing. Unnecessary input parameters were also removed by trial and error to streamline the training. Finally, similar performances of the MLANN with the 2008 training and 2009 independent validation datasets ensured that the trained network was producing global minima without overfitting."

7. The function "g" in Figure 1 is a shifted and scaled sigmoid activation function, but neither the shifting nor scaling will have any effect given the free weights and biases. This unnecessary "tweak" is another addition to the cognitive load that makes reading this manuscript a real slog.

Figure 1 appears to have caused nothing but confusion. To ease the cognitive overload, it and all references to it have been removed from the revision.

8. The research ought to be reproducible. Normally I would say "more easily reproducible", but this work has not reached the bar of reproducible at all. The software and data ought to be provided, preferably coupled with a clear pipeline for training and evaluating the model. Software the authors didn't write must be cited (what software evaluated and optimized the neural network?), and any software the authors wrote ought to be included as a supplement. If a pipeline includes downloading and processing of raw data from some CERES Ordering Tool API, then the link-to-the-data provided in the manuscript is sufficient; if it does not, the processed data ought to be published.

The Mathworks software that we used is now referenced and linked. With that Matlab package and the C3M data linked at the end, it should be possible to reproduce any of the training statistics.

9. Minnis et al. 2019 is not in the references.

It has been added. It could also have been made available upon request.

10. The subscript on f in Figure 1 is not needed. In fact, it is a misdirection because the subscript indicate the existence of hidden layers which do not exist. The nodes currently labelled f_1 should be labelled u_1, u_2, u_3, etc. What is the difference in Figure 1 between x_1, x_2, x_3, etc. and the first layer of nodes with no label?

Figure 1 is no longer included.

11. Line 207-209 is an incomplete sentence.

Corrected

12. The studies included in table 5 do not use the same test data, so comparing the metrics is okay but not entirely quantitative. Indicating the "best" result with a bold-face font pushes the comparison too far.

Bolding removed.

13. The similarity of the CC and MLANN cloud fractions when stratified over space (Fig 14) or time (Fig 15) could be quantified with a correlation coefficient.

We have added the following to the paragraph describing day Fig. 14 results:

“Linear regression between the daytime CC and the MLANN regional means yields R^2 values of 0.80 and 0.66 for the near-nadir and full-swath results, respectively. The smaller value for the full-swath data is not surprising given its greater sampling. For the matched near-nadir and full-swath means, $R^2 = 0.81$.”

We added the following to the paragraph describing the nighttime Fig. 14 results:

“The correlation coefficients are 0.71 and 0.64 for the nocturnal CC regional means matched with their respective MLANN near-nadir and full-swath counterparts, while R^2 is 0.89 for the matched near-nadir and full-swath averages.”

We added the following to the paragraph describing the Fig. 15 results:

“The values of R^2 between the CC and MLANN monthly means are 0.92 and 0.90 for day and night, respectively.”