We'd like to thank the editor for handling our manuscript, as well as reviewer #1 for reading our manuscript and providing numerous helpful suggestions for improvement.

We have carefully read through all the comments and questions and revised the manuscript accordingly. Please find our point-by-point response to reviewer #1 below. Here, the reviewer's general remarks, as well as the specific questions/comments, are formatted to be left-aligned text in bold font. Our responses are indented and formatted in regular font.

Here is a summary of the major changes in the revised manuscript:

1) We defined an independent test data set that is not included in the determination of hyperparameters or the training of the ANNs. The new splits are 70%/20%/10% for the training/validation/test data set. We also excluded the training data set from the evaluation of the model performance. This means that, e.g., the histograms in Figure 5 or the global maps in Figure 6 are generated by profiles the ANN has not been trained on. This allows for a fairer evaluation of ANN performance.
2) Since we removed profiles for the test data set and introduced new splits between training and validation data, we needed to repeat the k-fold cross validation and training of the ANNs. This turned out to be a necessary step, as we were able to fix three bugs in our algorithm setup: (i) We had not considered the number of hidden layers to be a hyperparameter. Tests revealed that models with only one hidden layer slightly outperformed those with two layers for the cloud classification scheme (the cloud top pressure models still use two layers). (ii) We had shuffled the training and validation data twice. While this had (obviously) no effect on training performance, it affected the correct recording of the respective profile indices. In other words, we did not correctly track the profiles in the training and validation data sets. This, in turn, means that the validation statistics presented in Figures 4 and 10 were inaccurate, as the presented "validation data" was actually comprised of random profiles from both the training and validation data set. Note that in the original manuscript version, model performance was evaluated for the combined training and validation data set (e.g., Figures 5 and 7), which means this mistake had no effect (i.e., the evaluation was based on a combined data set). (iii) The wrong control file provided the cloud top pressure model in the original manuscript version. That specific model had no weight decay (i.e., the model could learn training data very well, to the detriment of generalization) and early-stopping was turned off. This, together with the wrong recording of training and validation indices, resulted in the unrealistic correlation coefficients of 0.99. The model in the revised manuscript exhibits a much more reasonable correlation coefficient of 0.82.
3) We added more detailed explanations of machine learning terminology and descriptions of the considered hyperparameters.
4) We replaced one of the example scenes over South East Asia. In the original manuscript, the two scenes in Figure 9 looked very similar. Instead, we decided to present a more complex cloud field, which nicely illustrates the performance of the cloud classification model, while highlighting instances where the cloud top pressure prediction struggles.
5) We extended the analysis of the cloud top pressure ANN performance considerably. That section now includes additional statistical analysis of the difference between predictions and observations, as well as the model's ability to detect clouds <400, 350, and 300 hPa.

We also added global maps of the model performance, as well as comparisons between MODIS, ANN, and v4.2x data (similar to the cloud classification analysis). Example maps now contain the same scenes as for the cloud classification part.

General comments

This paper nicely illustrates that the implementation of machine learning to MLS cloud classification leads to an impressive improvement in MLS cloud detection, compared to current operational techniques.

The paper is concise, well written, and discusses well-selected calculations. The discussion of both global statistics, and individual cases, is very appealing. The Summary and Conclusions section is very well written.

The discussion of the machine learning methodology is very concise, but could benefit by briefly defining some of the machine learning terms which may not be familiar to the atmospheric science research community.

Specific comments

The use of machine learning techniques and terminology is likely unfamiliar to many in the atmospheric sciences. There are several places in the text in which a few additional words / sentences could help the reader understand better what is being done by the authors. There are some terms which need to be defined. Please discuss, for example, what is meant by "feedforward" on line 121. Other terms that should be defined (briefly discussed) are "imbalanced classes", "learning rate", "Nesterov momentum value", and "weight *decay".*

We added the following descriptions to the manuscript.

"Here, we constructed and trained a multilayer perceptron, which is a subcategory of feedforward ANNs that sequentially connects neurons between different layers. In a feedforward ANN information only gets propagated forward through the different model layers and is not directed back to affect previous layers."

And:
"Generally, *F1* assigns more relevance to false predictions and is more suitable for imbalanced classes, where the respective data sizes vary significantly."

And:
"The hyperparameters to be determined are (i) the number of hidden layers, (ii) the number of neurons per hidden layer, (iii) the optimizer for the cloud classification, (iv) the mini-batch size, (v) the learning rate, and (vi) the value for the weight decay (i.e., the L2 regularization parameter). The number of hidden layers and neurons impact the complexity of the model. The choice of optimizer controls how fast and accurately the minimum of the loss function in Eq. (8) is determined, based on different feature sets and

minimization techniques. During each iteration the model computes an error gradient and updates the model weights accordingly. Instead of determining the error gradient from the full training data set, our models only use a random subset of the training data (called a mini-batch) during each iteration. This not only speeds up the training process, but also introduces noise in the estimates of the error gradient, which improves generalization of the models. The learning rate controls how quickly the weights are updated along the error gradient. Thus, the size of the learning rate affects the speed of convergence (higher is better) and ability to detect local minima in the loss function (lower is better). Meanwhile, L2 regularization is one method to specify the regularization term R in Eq. (8), where the sum of the squared weights is multiplied with the L2 parameter:

$$R = L2 \cdot \sum \omega^2 + \varpi^2 + \Omega^2 \ (9)$$

Note that for clarity we omitted the indices for the weights in Eq (8). The amount of regularization is directly proportional to the value of the L2 weight decay parameter. Regularization usually improves generalization of the models. More information about ANN hyperparameters and their impact on the reliability of model predictions can be found in, e.g., Reed and Marks (1999) and Goodfellow et al. (2016)."

Note that due to changes in setting up the models, as well as the performance evaluation, we found that the Adam optimizer slightly outperforms the stochastic gradient one. We changed the description accordingly.

Technical comments

Line 21 the phrase "cloud amount" is vague. Please be more specific.

We changed the wording to "cloud cover".

Line 46, add commas, revising to e.g. "radiances, from lower in the atmosphere, and smaller downwelling radiances from above, into the MLS raypath" to improve readability. In my first reading of the sentence I had a hard time making sense of the sentence.

We changed the sentence following the reviewer's recommendation: "a mix of large upwelling radiances, from lower in the atmosphere, and smaller downwelling radiances, from above…".

Line 55, what is meant by "discount them" ?

We meant to say that these radiances are discarded when the observation vector for the optimal estimation is constructed. We changed the wording to "discard".

Line 89, please specify Figures in Waters et al 2006 or other papers that illustrate the spectral sampling details of the AURA MLS experiment, so the reader can obtain a fuller understanding of the MLS experiment.

We added "; see Table 4 in Waters et al. (2006) and Figure 2.1.1 in Livesey et al. (2020)." to the revised manuscript.

Line 130. It would be helpful to point out that Figure 1 is presented for illustrative purposes, since line 253 later points out that each hidden layer has 851 neurons (instead of 2 neurons). "Figure 1 illustrates the general setup of a simplified multilayer perceptron that contains four layers, and is instructional. The full model setup is discussed in Section 3.4"

We changed the sentence as follows: "Figure 1 illustrates the general setup of a simplified multilayer perceptron that contains four layers, and is purely instructional. The complete model setup is more complex and is discussed in sections 3.2–3.4."

Line 168. Is the MLS aggregation at 1°x1° because the MLS data sampling is (line 100) near 165 km?

We spent quite some time thinking about this detail. Indeed, the half and full distance between adjacent MLS profiles is close to 0.75° and 1.5°, respectively. At the same time, the typical horizontal scales of clouds that can potentially impact MLS observations (i.e., optically thick mid-level cloud fields and high-reaching cumulonimbus) are in the range of 50-200 km (Guillaume et al. 2018). This gives us a range of ~0.5°-2.0°.

We tested the aggregation for different scales 0.5°-3.0° in increments of 0.5° to get an idea about the importance of the aggregation perimeter. We noticed no significant difference in performance for scales between 0.5° and 2.0° (variability in Matthew's Correlation Coefficient of <0.01). However, performance got gradually worse for 2.5° and 3.0°.

In the end we decided on 1°x1°, which is (i) close to half of the distance between adjacent MLS profiles, and (ii) in the middle of the relevant horizontal cloud scales.

We added some extra information to the manuscript at the end of the third paragraph of section 3.2: "Note that no significant decrease in classification performance is observed for varying aggregation scales between 0.5°x0.5° and 2°x2°."

Line 173 are the 5,000 samples MODIS, MLS, or MODIS-MLS samples?

This number refers to MODIS-MLS samples. We changed the sentence accordingly: "While not every grid box contains the same number of profiles, each area contains at least 2,100 MLS-MODIS samples. A maximum in sample frequency is observed over the regions with denser MLS coverage around the poles."

Note that we have changed the horizontal resolution from 60°x60° to 15°x15° in response to a comment from referee #2. We also added two separate maps, one for the statistics of the total MLS-MODIS data set, and one for the statistics of the clear and cloudy cases (as defined in section 3.2).

Line 262. Approximately how many epochs are calculated?

When we started to test different setups, we ran each model with a fixed number of 10,000 epochs. However, we quickly noticed that each model starts to converge to a solution (i.e., the validation loss does not decrease any longer) much earlier. This number is comparatively low; indeed, more complex regression simulations performed by the

MLS group require 10-100 times more epochs. This means that the 2-class binning performed by the cloud classification models in this study is computationally inexpensive and only takes about 1 day.

We added this information to the manuscript: “Note that the lowest validation loss usually occurred after ~2,000-3,000 epochs for both the cloud classification and p_{CT} prediction.”

Line 318 clarify what is meant by “classification going forward”.

We meant to say that in this study, we use the model with the highest Matthew’s Correlation Coefficient (*Mcc*), out of the 100 we trained. One could think of other approaches, e.g., picking the one with the median *Mcc*, or another binary metric. However, since the validation scores are so close to each other, there really isn’t a practical difference between each of the models.

We changed the sentence to:
“Given the statistical robustness of the results, the model with the highest *Mcc* and lowest RMSD provide the ANN weights for cloud classification and p_{CT} prediction in this study, respectively.”

Line 549. If the current MLS data version is V5, why not include the new ANN capability in the V5 product instead of “future versions of the v4.2x” product?

The way we phrased the outlook was confusing. We compared the ANN cloud flag to the operational v4.2x cloud flag, as v5.x data was still being processed at the time of writing. The MLS radiances and cloud detection code are identical between the two versions, however, revisions to the atmospheric composition retrieval algorithms yield some subtle differences in the cloud status flags. These differences have no impact on the conclusions reported in this manuscript. Since the ANN cloud classification scheme only uses MLS radiances as input, it is independent of the MLS L2 algorithm version.

We plan to continue to provide both v4.2x and v5.x data products for the foreseeable future. In the revised manuscript we changed this sentence to:
“This new cloud classification scheme, which will be included in future versions of the MLS dataset, provides the means to reliably identify profiles with potential mid- to high-level cloud influence. Note that MLS radiances are not affected by the change from v4.2x to v5.0x.”

We also added a clarifying statement to section 2:
“Note that the sampled radiances are identical between the two versions, while revisions to the atmospheric composition retrieval algorithms yield subtle differences in the derived cloudiness flags.”

References:

Guillaume, A., Kahn, B. H., Yue, Q., Fetzer, E. J., Wong, S., Manipon, G. J., Hua, H., & Wilson, B. D. (2018). Horizontal and Vertical Scaling of Cloud Geometry Inferred from CloudSat Data, Journal of the Atmospheric Sciences, 75(7), 2187-2197. Retrieved Aug 27, 2021.