Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., referee comment RC2
https://doi.org/10.5194/amt-2021-146-RC2, 2021

Comment on amt-2021-146

Carlos Jimenez (Referee)

Referee comment on "Improved cloud detection for the Aura Microwave Limb Sounder (MLS): training an artificial neural network on colocated MLS and Aqua MODIS data" by Frank Werner et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2021-146-RC2, 2021

General comments

The paper presents an ANN algorithm to detect clouds from MLS radiances. The previous algorithm was based on radiative transfer modeling of clear sky radiances, and a threshold on the differences with the observed radiances to decide on cloudiness. The new algorithm does not rely on radiative transfer calculations, but on approximating the statistical relationship between the MLS radiances and a MODIS cloud dataset with an ANN.

The paper is well written, and includes adequate analyses and examples to convincingly demonstrate that the proposed cloud flag outperforms the previous algorithm. Nevertheless, I would like to offer some suggestions regarding how the work has been carried out and how it is presented to the potential readers of the paper.

Overall, in my view there are three aspects of the paper that could have been addressed differently:

- Dependance of the statistical inversion model on a given cloud properties dataset. I would say that this has not been properly emphasized in the paper. Retrieving cloud properties is not straightforward, and relatively large differences are found when comparing products. If the statistical model approximating MLS radiances and cloud properties would have been trained with, e.g., Calipso cloud properties, or even a previous version of the MODIS dataset, that would have resulted in a different model. This is not a criticism of the choice of the MODIS product, which is perfectly justified, but a reminder that this is an important part of the algorithm. For instance, biases on the cloud product are likely to be learned by the statistical model, and not be shown when evaluating the model performance with the same dataset used to train the model. Sentences like "This algorithm is designed to classify clear and cloudy conditions for individual MLS profiles, based purely on the sampled MLS radiances" do not help to convey this message.

- Description of the ANN. In my view, it seems a bit unbalanced in terms of the elements that are, and are not, described. On the one hand, most readers are likely to be familiar with the description of the architecture of a standard multi-layer perceptron, so Section 3.1 could have been considerably shortened. On the other hand, terms like L2 regularization, mini-batches, Nesterov momentum, less arguably familiar concepts, are used without any further explanations. I know that it is difficult to strike the right balance there, as depends on the reader, but I feel that this could have been done in a more appropriate way.

- Retrieval of cloud top pressure. The paper gives the impression that the retrieval of cloud top pressure is not integrated in the paper from the beginning. While the cloud flag takes more of the paper, with different sections and comprehensive analyses, the cloud top pressure retrieval is not mentioned in the introduction (only mentioned at the end when describing the paper contents), the work takes just one section at the end of the paper, and seems much less detailed in terms of modeling and analyses. This may be intentional, as the authors may want to give more value to the cloud detection than to the cloud top pressure retrieval. But it feels a bit awkward, as if the cloud top pressure was an afterthought. The paper may have looked more consistent if, for instance, the ANN for the cloud top pressure would have also been described in the algorithms section, mentioning that these are 2 different inversion problems requiring different ANN setups (a binary classification, and a continuous mapping between 2 finite spaces), or if the qualitative assessments would have used the same selected scenes with a bit more of a joint discussion.

Some more specific comments are given below. There are mainly about the statistical inversion, as the MLS team knows very well its instrument, and the discussion concerning the instrument capabilities, and what it can be retrieved from its radiances, is already very solid.

Specific comments

L58. "However, the reliance on estimated clear sky radiances and the use of predefined thresholds induces uncertainties in the current algorithm". This can be interpreted as a lack of confidence on the clear sky radiative transfer modelling, which is the basis of most of MLS retrieval work. I would say that the problem is more defining universal thresholds that can reliably identify the clouds, as illustrated later in the paper.

L70. "This algorithm is designed to classify clear and cloudy conditions for individual MLS profiles, based purely on the sampled MLS radiances." I find this sentence misleading, as you always need something else than the radiances to do an inversion. The relationship between the radiances and your parameter of interest needs to be established, e.g, by a radiative transfer model (original cloud flag), or by a statistical model (proposed flag). For this specific work, the statistical algorithm depends on the sampled MLS radiances and their relationship to the MODIS retrievals.

L72. "both high and mid-level clouds (e.g., stratocumulus and altostratus)". I know that the mentioned clouds are just examples, but they seem to coincide with mid-level clouds. It may give the impression that "cirro" clouds are not targeted, e.g., because of limitations of the MODIS cloud product.

L75. "Aqua MODIS observations are ideal for this study". Perhaps "ideal" is not the best word here, given the large difference with MLS in terms of observing geometry, spatial resolutions, etc. Suitable?

L95. "The most recent MLS dataset is version 5; however, at the time the ANN was being developed, reprocessing of the entire 16-year MLS record with the v5 software had not yet been completed. Accordingly, L2GP cloudiness flags in this study are provided by the version 4.2x data products (Livesey et al., 2020), and v4.2x is also the source for the Level 1 radiance measurements used herein ". Is there anything significantly different in the V5 radiances that could have an impact on this work? I guess not, but it may be worth commenting that.

L115. "Table 1 lists the 208 days that comprise the global data set used in this study. It consists of eleven random days from each year between 2005 and 2020, as well as a pair of two consecutive days to bring the yearly coverage to thirteen days." In my view Table 1 is not really needed, i.e., knowing that in 2012 d06 was day-of -year 169 is not critical information to pass to the reader.

L118. "Particular attention was paid to ensure that each month is represented (close to) equally in the final data set". I am a bit puzzled here. Why not something as simple as randomly selecting one day per month and year? Then all months are equally represented without any need for further checks.

L131. I was wondering about the choice of two hidden layers and not just one. Perhaps because this is a binary classification problem and the ANN search a decision boundary to separate clear and cloudy? But using the log-sigmoid function of the output node, the ANN will not be just separating into 2 classes, but outputting the probability of having a cloud. This could be interpreted as an ANN approximating a continuous mapping between two finite spaces, the radiances and the probability of having a cloud, i.e., a number between 0 and 1, and therefore one hidden layer may suffice.

L171. Why 60x60 degrees? A finer grid (e.g., 5x5 degrees) would be useful to better understand the geographical distribution of the samples. In fact, what is producing such an uneven distribution? Why the higher sampling over Africa?

L185. "Naturally, these definitions leave some profiles undefined (e.g., those with C in the range 1/3–2/3)". The definition of clear and cloudy classes is perfectly justified, but excludes the undefined profiles from the training dataset, as properly stated in the text. Later on, when applying the ANN to classify the observations, the ANN will have to classify profiles similar to those unseen during the training. So, in principle, the ANN will be extrapolating. Could this be a problem? Could have been an alternative to train with all cases where pCT > 700 hPa, but targeting the continuous variable C with values in the range 0-1, instead of the two defined classes? The ANN would be similar, apart from choosing a loss function more appropriate for a continuous output space.

L245. "Overall, the input matrix for the training and validation of the ANN is of shape 1, 710 × 162, 117". 1710 features are still a relatively large number. Even if a channel selection has been applied, one may wonder about the possibility of further reducing the

dimensionality of the input space with some of the typical feature extraction techniques. Doing so is a common procedure to decrease computational burden and enhance the generalization properties of the ANN. I do not know well the MLS radiances, but I assume that there is some degree of correlation between the 1710 features, which could make possible this further dimensionality reduction.

L253. "The number of neurons per hidden layer is set to 856, which corresponds to the average between the number of nodes in the input and output layers". This sentence may make the reader think that this is a sort of standard procedure to fix the number of neurons, which is not. The number of nodes sets the capacity of the model to learn complex mappings, with small (large) ANNs having the risk to underfit (overfit). But the number of nodes is not typically considered as an hyperparameter to control model complexity. Instead, training techniques similar to the one applied in the paper are applied to control the generalization capacity of the ANN.

L253. The 856 nodes in the hidden layer results in a very large ANN. If you are using a fully connected ANN of the type described, the number of weights to be adjusted during the training is 1710 (inputs) x 856 (hidden layer one) + 856 x 856 (hidden layer 2) + 856 x 1(output), and then you have 856 + 856 + 1 biases. This is around 2 million of model parameters, even larger than the number of samples x input features in the training dataset. I may be missing something, but it is hard to believe that a simpler ANN cannot be setup to classify the radiances into the two clear and cloudy cases.

L255. "(ii) the learning rate, (iii) the mini- batch size, and (iv) the value for the weight decay (i.e., the L2 regularization parameter)". What do you mean by mini-batch? What is L2 regularization? Square of the weights instead of absolute values?

L256. If a weight decay term is used, should not have been included in the loss function of Eq. 8? That will make clear what L2 means.

L257. "of the cost function in Eq. (8)". Perhaps saying loss function, to refer to Eq. 8 with a single name throughout the paper?

L262. "Instead, the models are run with a large number of epochs, and the lowest validation loss is recorded, so an increase in validation loss during the training (i.e., cases where the model is overfitting the training data at some point) has no impact on the overall performance evaluation". It is not clear to me how you apply an "early-stopping technique" to control model complexity here. There can be situations where the validation loss starts to be smaller than the training loss, i.e., an indication of over-fitting, but, nevertheless, it keeps decreasing, although with a smaller rate than the training loss. At some point, the validation loss may reach a minimum. Is that minimum the "lowest validation loss" you describe, where you consider your ANN properly trained, so those are the selected ANN weights? But it could be the case that at that point the training loss was already smaller that the validation loss for a large number of epochs, so in principle the

ANN could be already overfitting. Perhaps it is just not well explained, or I am missing something.

L273. "The ideal setup for the ANN". Perhaps "ideal" is not the best word here, as there will exist a number of ANNs performing very closely (e.g., close but different number of nodes). An "appropriate" setup for the ANN?

L288. "This analysis revealed that the stochastic gradient descent optimizer, using a learning rate of 0.001, and a Nesterov momentum value of 0.9 yielded the overall best validation scores. The best weight decay and mini-batch size values were found to be 5 × 10−4 and 1024 (i.e., 0.8% of the training data), respectively." Stochastic gradient descent? Nesterov momentum?

L292. "By chance, the most obvious cloud cases (e.g., C = 1 and very large QT values) might have ended up in the validation data set, or vice versa, and the trained weights might be inappropriate". I would say that your random selection of training and validation cases, together with the relatively large number of samples, makes this very unlikely to happen. If it is happening, I would revise the sample selection strategy.

L294. "Moreover, a large disparity in validation scores for multiple models might be indicative of an ill-posed problem, where the MLS observations do not provide a reasonable answer to the cloud classification problem". Yes, this is for me the valid reason to undertake these tests.

L297. "In this study, 100 different models are developed ". To be clear, the only thing that changes is the split in the training-validation datasets, the model hyperparameters are set to your final configuration, right?

L300. "The output of each ANN model is a cloudiness probability (P) between 0 (clear) and 1 (cloudy)". Perhaps this should have been already mentioned earlier in the text, e.g., around Eq. 7, as this is the consequence of building the binary classifier with a softmax function in the output node.

L339. Perhaps there is no need to give both absolute number of cases and percentages of the total. I personally found the percentages more informative.

L345. "Only 1.7% of clear profiles are falsely classified as cloudy by the new ANN algorithm, while the current v4.2x status flag mislabels 6.2% of these profiles". I do not have doubts that the new cloud flag performs better than the V4.2 one. But I think it is worth mentioning that the V4.2 flag will always be penalized in these comparisons. The

new flag has been trained on the same dataset used for the evaluation, while the V4.2 is independent of that dataset. For instance, biases in the MODIS cloud properties are likely to be learned by the ANN, and not shown in the evaluation. The V4.2 cloud, on the contrary, knows nothing about those biases.

L358. "It is essential to understand the ANN performance for the undefined, in-between cases". Yes, I fully agree. As mentioned above, the ANN has never seen the undefined cases, so in principle it is extrapolating to classify those profiles.

L379. "Due to the looser definitions, there is a significant drop in performance scores". Could it be not just the looser definitions, but the fact that now the classes include cases never seen by the ANN as they were not part of the training dataset? Perhaps this is a more realistic assessment of how the ANN will perform later when faced with all MSL radiances.

L399. "Each grid box covers an area of 60â□¦ × 60â□¦ (latitude and longitude)". Why not showing in a finer grid? For instance, the 3â□¦ × 5â□¦ used for the remaining plots.

L399. "In contrast to the results for the ANN algorithm, there is a clear latitudinal dependence for the performance of the v4.2x algorithm". As I mentioned above, these comparisons penalize v4.2 as we are evaluating not with true cloud properties, but with the MODIS cloud properties targeted by the new flag. For instance, there may be the case that there is a latitudinal bias in the MODIS cloud parameters, and the ANN may learn it. But as he F1 scores are so poor for v4.2, we can still conclude that the new flag outperforms the v4.2 flag on the basis that MODIS C6 provides a realistic representation of clouds, even if not free from errors.

L429. "Due to the reduced sensitivity towards such clouds (see the discussion in section 3.3), the cloud covers predicted by the ANN are much closer to the MODIS results 430 that do not include low clouds". This was a nice test, with a very reasonable agreement.

L442. "Figure 8 shows two example cloud fields over the North American monsoon region. Nice and very illustrative examples.

L480. "The input layer and the two hidden layers remain unchanged from the cloud classification setup. The labels in the output layer, instead of being set to either "0" or "1" (i.e., clear sky or cloudy), now contain the respective pCT reported by the colocated MLS-MODIS data set". Two hidden layers are probably not required here, as this is definitely a continuous mapping between two finite spaces, so one hidden layer has the capability to approximate this mapping. Because of that, the ANN may look even more over-dimensioned than when acting as a binary classifier.

L483. “Similarly, the model optimizer, learning rate and mini-batch size reported in section 3.4 for the cloud classification ANN provide the best set of hyperparameters”. Does it mean that you are using exactly the same hyperparameters? That looks strange to me, as you have a different mapping to approximate, so the optimal hyperparameters may not be necessarily the same.

L484. “here the only change concerns the weight decay parameter, which is turned off.” Is there a reason for that? Why approximating the new mapping does not require a weight-based regularization of the loss function, while the previous one required one?

L487. “Joint histograms of true (in the sense that they are the prescribed labels to train the ANN)”. This is a good reminder about the fact that true here means MODIS-retrieved. It would have been nice to also introduce a bit more this idea when evaluating the results of the previous ANN with MODIS-retrieved cloudiness.

L488. “are presented in Figure 10”. This seems like a much less detailed analysis of the ANN performance when retrieving pCT, compared with the cloud flag analysis. Even if the goal is only to differentiate between mid-to-low level clouds and high-reaching convection, other metrics than just the correlation, and how those metrics may depend on cloud type, altitude, location, and so on, could have been assessed.

L494. “Three example scenes with the MODIS pCT”. Not suggesting that the presented cases are not interesting, but another option could have been to reuse the cases presented for the cloud flag. Specially the second example there seem to have the right mixture of high and low clouds to permit an analysis of the pCT retrieval. Likewise, it could have been interesting to see the performance of the cloud classifier on the pCT scenes.

L515. “In this study, we present an improved cloud detection scheme based on the popular “Keras” Python library for setting up, testing, and validating feedforward artificial neural networks (ANNs)”. Perhaps more interesting than mentioning the library would have been to briefly describe the setup, i.e., something like a “standard multilayer perceptron configured to act as a binary classifier by using a softmax activation function in the output node and a cross-entropy loss function to derive the weights”.

L533. “A comparison with the current v4.2x status flags reveals that for the complete data set in this study the new ANN results provide a significant improvement in cloud classification”. Perhaps a good place to briefly mention than the “truth” here is the target of the ANN calibration used to derive the new cloud flag.

L550. “in future versions of the MLS v4.2x”. This may seem confusing, as a V5 has already been mentioned in the text. Or perhaps v4.2x does not supersede v5 and both will be coexisting, with both v4.2x and v5 benefiting from the new cloud flag algorithm?