Reply to anonymous referee 2

First of all, we would like to thank the reviewer for reading our paper carefully and providing constructive comments. In the revised manuscript, we have tried to accommodate all the suggested changes. The modifications from the originally submitted version are highlighted in the revised manuscript. Please see our specific responses below.

The paper describes a new technique for satellite measurements of cloud optical thick- ness and cloud droplet effective radius. The key feature of the technique is that it takes into account 3D radiative effects and subpixel variability by considering not one pixel at a time, but by performing simultaneous retrievals over 10 by 10 pixel areas. The most important aspect of the technique is the use of a deep learning algorithm. This is a significant new development, and the study makes an important contribution on the path toward more accurate satellite retrievals of cloud properties. Overall, the method- ology is sound and the presentation is suitable. However, I believe that a few important improvements are needed in the analysis. My recommendation is therefore to make some major revisions. Please find below my detailed comments.

comments

Major issues:
1.
Page 7, Line 8 mentions that "The test dataset used for evaluation should be indepen- dent of the training dataset." My sense is that in this initial study the training and testing datasets are not fully independent, as they come from the very same cloud fields, and that this would be good to mention. (The two datasets include different randomly se- lected locations within the cloud fields, but the statistics of cloud properties are identical in the training and testing datasets.)
As noted in Page 10, Lines 7-8, it will be an important future step to examine the per- formance of the retrieval for a wider range of cloud parameters. It is reasonable to leave this (and the evaluation based on fully independent training and testing datasets) to a future paper, but even the current results could offer further insights into the ro- bustness of the proposed retrieval algorithms. Most importantly, one could examine not only the overall results, but also separately the results for open-cell and closed-cell convection cases. This would demonstrate that the same algorithm and training set improves retrieval accuracy for two very different types of cloud structures. I don't think the currently presented results show this: Overall error statistics may conceivably im- prove due to improvements for open-cell convection only, without any improvements for closed-cell convection. (Because retrieval uncertainties are likely larger for open-cell convection, it may be best to examine by what percentage DNN-2r and DNN-4w reduce the retrieval errors of IPA retrievals for open-cell and for closed-cell convection.) The paper did a good job in examining results as a function of optical thickness, but the new

analysis of already performed retrievals would help because open-cell and closed-cell convection cases differ in horizontal structure even at locations where vertical optical thicknesses are similar.

Response: We have added Table 2 to show RMS errors (in %) for open and closed cell cases, separately. The results show that DNNs generally retrieve more accurate COT and CDER than IPA method. For SZA of 60 deg., DNNs are better than IPA in both open and closed cloud cases, while there is an exception; IPA is better for COT for SZA of 20 deg. in the closed cell case. We have added explanations about this table, as follows:

"Table 2 shows the relative RMSE of estimated COT and CDER by the IPA and DNN retrievals for open-cell and closed-cell cases. In both cases, the retrieval accuracies for DNNs are obviously improved compared to the IPA retrieval. An exception appears for COT in closed-cell case when SZA is 20∘; COT RMSE of 24% for DNN is larger than 16% for IPA. The DNN-4w is better than DNN-2r in both cases."

As the reviewer suggested, we have revised the explanation about test datasets as "the training and test dataset include different randomly selected locations within the cloud fields, but the statistics of cloud properties are identical in the training and test datasets."

2.

Page 5, Lines 13 and 22: I wonder why scene parameters are estimated for 8 X 8 pixel arrays when using the DNN-2r method, but only for the central 6 X 6 pixel arrays when using the DNN-4w method. This could make sense if 3D effects acted over larger distances at 3.75 microns than at the 0.86 and 2.15 microns used by the DNN- 2r method, but neither my own physical reasoning nor the filter weights in Figure 8 suggest this. In fact, Figure 8 shows that DNN-4w retrievals at a pixel are strongly affected by 0.86 micron radiances 2 pixels away. This suggests that (at least for pixels at the edges of 8 X 8 pixel areas) the DNN-2r method cannot capture the portion of 3D effects caused by areas more than a pixel away. This probably contributes to DNN- 2r giving less accurate results than DNN-4w (a tendency mentioned in Page 9, Lines 31-32) and should be mentioned in the discussion of the differences between the two methods at the top of Page 10. (The discussion should also include the impact of additional wavelengths in DNN-4w and algorithmic differences.) Also, it could help to clarify explicitly in the paper whether DNN-2r retrievals inside (not along the edges of) 8 X 8 pixel areas are affected by radiances 2 pixels or more away. If they were, it could even make sense to analyze retrieval accuracy only for pixels in the central 6 X 6 pixels of 10 X 10 pixel areas (similarly to DNN-4w).

Response: We appreciate this comment. Honestly speaking, we had no strong reason to limit the output pixels in DNN-2r and set it as 8x8 pixels. On the other hand, because of convolution filters of 5x5 pixel size, the number of pixels were limited as 6x6 pixels. As the reviewer

pointed out, the difference of the number of pixels in the DNN output layer may have an effect on the retrieval performances. However, we think that it is not a main reason because 1) influences from neighboring pixels generally tend to weaken with increasing distance from the target pixel, 2) there are only 28 edge pixels in the 64 (8x8) pixels area, and 3) most of these edge pixels just miss only one side of neighboring pixels 2 pixels away. The 3D radiative effects actually operate on larger horizontal scales than 2 pixels (560 m) as captured by DNN-4w, while the effects tend to weaken for larger scales. As shown in Fig. 8, the filter coefficients do not vanish at the edges of the window. To improve retrieval accuracy, it is better to increase the number of adjacent pixels used for both DNN-2r and DNN-4w, as this point is mentioned in the conclusion.

In the initial stage of this study, we tested various kinds of DNN architectures with a different number of wavelengths, the use of convolutional layer, different activation function, and so on. For example, we tested one similar to DNN-2r but using four wavelengths or one similar to DNN-4w but using only two wavelengths. However, the DNN-2r and DNN-4w were the best performed ones. Therefore, at least we can say that a main reason for the better retrieval performance for DNN-4w than DNN-2r would be a combined effect of use of additional wavelengths and a different DNN architecture (e.g., use of convolutional layer and release from the reliance on the IPA retrieval). However, unfortunately, we could not make the reason very clear. We have added the description regarding the future improvements in the manuscript, as follows:
"The above two DNN structures were obtained from various trial-and-error experiments. Different DNN structures were also tested. For example, we tested a DNN similar to DNN-2r but with four wavelengths, and one similar to DNN-4w but with only two wavelengths. However, DNN-2r and DNN-4w performed best. There is room for improvement in DNN structures, which should be investigated in the future."

Minor issues:

Page 1, Line 23: What is meant by "cloud state"?

Response: In this context, "cloud state" means cloud vertical and horizontal inhomogeneity. We replaced these words in the new manuscript as "cloud horizontal and vertical inhomogeneity".

Page 2, Line 23: The study by Evans et al. ("The Potential for Improved Boundary Layer Cloud Optical Depth Retrievals from the Multiple Directions of MISR", J. Atmos. Sci., 2008) should also be mentioned, as it also used a neural net for cloud retrievals.

Response: We have added Evans et al. (2008) as a reference.

Page 2, Lines 26-27: What is meant by "feature" and "feature extraction"?

Response: In this context, "feature" means a feature in the training datasets. This word is usually used in the classification problem (e.g., object recognition from input image) using DNN. We have modified the manuscript as follows:
"features in training datasets are learned hierarchically in the DNNs, although it is not very easy know how the features are described in the DNNs"

Figure 2: It would help to indicate the time elapsed during the 60 time steps along the horizontal axis, or to mention the time step in the figure caption.

Response: We have added an explanation in the Fig. 2 caption as
"A time step corresponds to one minute."

Figure 3: It would help to clarify why there is a fully connected layer near the top of the left column that operates only on radiances and not on the IPA-estimated scene parameters.

Response: We thought the layer should be in, at least, an either side (left or right side) of the figure. A layer can be inserted in both sides, but we did not have very strong reason. We believe this is a minor issue, but we have added a sentence as follows:
"In the first layers, radiances and IPA-estimated cloud properties are merged to obtain 8x8x2 elements (two elements per pixel for 8x8 pixels)."

Page 6, Lines 27-29: It would help to clarify whether all pixels within an LES scene are multiplied by the same randomly chosen number, or all individual pixels are multiplied by a different number. (My guess is the first option.)

Response: As the reviewer guessed, all pixels within a single scene are multiplied by the same number. We have added a sentence:
"The cloud extinction coefficients of all pixels within a single cloud scene are multiplied by the same number."

Page 6, Line 25 and Page 7, Lines 8-10: What does the word "samples" refer to? My guess is that each sample is a 10 X 10 pixel area. If my guess is right, can samples overlap? Also, it would help to mention the total number of pixels in the LES dataset, as this could show whether the training set includes almost all LES pixels or just a small fraction of them.

Response: As the reviewer guessed, a "sample" here means a 10 x 10 pixel area, and sample areas can be overlapped by other samples. The horizontal size of a LES scene is 28km x 28km. The total number of pixels in one LES scene is 100 x 100 = 10000 pixels. We have added some descriptions as mentioned above.

Figure 7: It could help to include into one of the panels a PDF of true optical thickness values.

Response: Figure 1 shows means and standard deviations of COT for the two cloud cases, from which rough shapes of COT PDF may be imagined. Although the distribution functions of true COT are not presented in this paper, the following figures show the joint histograms of the DNN-4w retrieval error and COT.
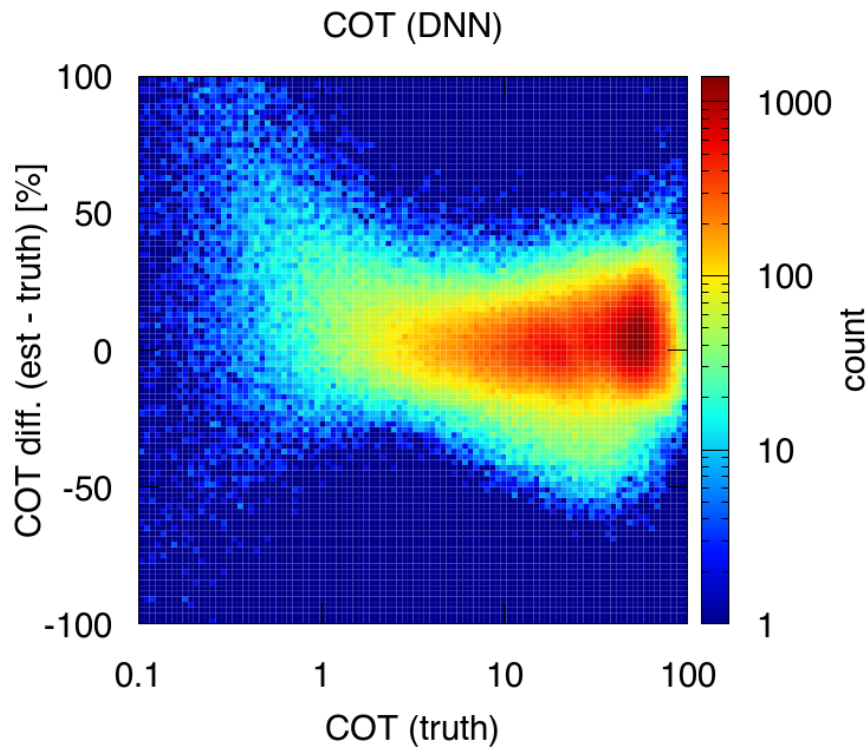

Figure 9: The legend should indicate which color shading corresponds to which line/method.

Response: We have revised Figure 9 as suggested.

Page 10, Lines 5-6: I am not sure the sentence "In the DNN-4w that we tested, we excluded 3D radiative transfer effects that occurred at horizontal scales greater than approximately 1.5 km (5 pixels)" is correct. Based on Figure 8, I thought that DNN-4w retrievals exclude 3D effects that occur at horizontal scales greater than 2 pixels (560 m). This is because I thought the pixel

whose properties we are retrieving is at the center of the filters in Figure 8, which means that only radiances two pixels away are considered. A correction of this sentence or a clarification of the meaning of filters in Figure 8 would help.

Response: The reviewer is right. We have fixed the sentence as
"...scales greater than 560 m (2 pixels)"

Somewhere in the text it would help to comment on whether the speed of calculations would be a concern for using DNN in operational retrievals in the near future. (For example, how does the speed of DNN compare to the speed of IPA and NN retrievals?)

Response: We did not exactly compare the computational costs of DNN, IPA, and NN retrieval. We have added an explanation about rather general facts about the DNN computational cost, in the last part of Section 3.3, as follows:
"As for computational cost, the training requires significant computation time, for which even one GPU helps considerably. Once the DNNs are trained, the retrievals using the present DNNs are generally very quick because they entail only very few simple manipulations of numerical data."