Reply to anonymous Referee 1 comments to *Neural network cloud top pressure and height for MODIS*

Nina Håkansson et al.

1 General comment

1.1 Referee comment:

This paper describes a new approach to retrieving cloud-top height using a neural network. It is an interesting report and gives us hope for improved retrievals. It will be more valuable if additional information is provided. It is much improved from the original submission. I realize that this is a first step, but a bit more analysis would provide the springboard for the next steps. This is an important paper, but too brief.

Reply:

We thank Referee 1 for acknowledging the paper as important and for all interesting comments that will help us extend the analysis of the paper.

2 Specific comments

2.1 Referee comment:

"Nowcasting" should be "nowcasting"

Reply:

We have correct this.

2.2 Referee comment:

Here and elsewhere: please spell out the acronyms the first time they are used (e.g., MODIS, AVHRR)

Reply:

We have correctes this. We had misinterpreted manuscript-preparation guidelines regarding AVHRR and MODIS. We have also updated the manuscript to use the correct acronym *CPR (CloudSat)* (Cloud Profiling Radar for CloudSat (CLOUD SATellite)) everywhere.

2.3 Referee comment:

Sec. 2.2 and 2.3: Please indicate nadir or viewing angles of the CALIOP and CPR.

Reply:

We have added that the viewing angle for CALIOP is 3°, and for CPR 0.16°. In Section 2.1 we have also added information of the satellite zenith angles for the MODIS data. For the matches with CPR the MODIS satellite zenith angle varies between 0.04° and 19.26°; and for matches with CALIOP between 0.04° and 19.08°.

Reply:

2.4 Referee comment:

Sec. 3.2 pg. 4, 25: while the CO2 absorbing band is generally referred to as the $15 - \mu m$ band, the MODIS channels are in the $13.3 - 14.4 \mu m$ range.

Reply:

We have corrected the channel ranges mentioned.

2.5 Referee comment:

Sec 3.3.2: Were the clouds single-layered or both single and multi-layered? It is not clear here. Please indicate if you are training only for single layered clouds or training for the topmost layer. Is there a lower optical depth limit of the clouds detected in the CALIOP 1-km product?

Reply:

To make it clearer we have explicitly stated that both single and multilayer clouds where included. We have also clarified that we used the uppermost layer of the top layer pressure variable as this is missing in the text (also noted by Referee 2).

Clouds optically thick enough to be detected when averaging the lidar data on 1km resolution should be included in the CALIOP 1km data. As we actually have the total optical depth from the 5km included in our match-up data (needed for other studies) we checked the lowest reported optical depth in 5km data for clouds that are detected in the 1km data, it was 1.5e-05.

2.6 Referee comment:

Sec. 4 Are there biases in any of the results for both CALIOP and CloudSat? The mean absolute error does not tell us any tendencies one way or the other. Knowing biases is critical. While MAE is an interesting and informative variable, it gives us less information about variability, which the standard deviation of the differences (SDD) along with the bias would provide us, especially when added to the MAE. Additions of the bias should be included in the tables and discussed. If there is no bias, then the SDD would still provide useful additional information and place the results in the same context as many previously published comparison studies. Addition of biases may help the discussion.

2.6.1 Referee comment (RC3):

Thanks for the explanation for not including the bias and SDD. This is precisely the kind of discussion that belongs in the paper. Without this explanation and discussion, it would appear to many readers that something is being hidden by the authors. The obvious question to most interested parties, particularly those who are potential users of the data, is, "Is the cloud height retrieved with this method, on average, in the right location? If not, how far away from the right altitude is it?" That is essentially the question both reviewers have asked. If I am assimilating or verifying a model output, I will want to put the cloud in the correct layer. An MAE of 500 m can just as easily be produced by all positive or all negative differences and thus I might expect to be within 500 m of the correct height on average, but I will not know if it is plus or minus 500 or if I am always biased high or low. The distributions in the current figures help but are not quantitative. If I look at other cloud height data sources and see that they tell me whether I should expect to be too low or too high on average, I might be more inclined to use one of their datasets. For example, Hamann et al. (AMT, 2014) summarized their differences in bias, stdv and rmsd. Straightforward. It is not the whole story as argued in the response, but an important part. and one most people can relate to. The reader is not well served when obvious statistics are excluded. An explanation for why the bias and SDD are not included has been provided to the reviewers, but not to the readers. There is a lot of good discussion and information in your explanation about the retrievals that are important to understand. For example, the breakdown of biases according to cloud height is very helpful. The differences in bias between CPR and CALIPSO follows from some of my other comments. I find the paper unacceptable without such basic statistics. I think that the paper should include all of it: bias, SDD, Skew, Median, and MAE. The discussion then should be directed at explaining what the best measure should be and why one is better than the other. Part of that is already done in the supplement.

Reply:

In our first reply we argued that bias and SD should be excluded because the error ditributions are non-Gaussian. We where convinced by the arguments of Reviewer 1 that all the measures along with the discussion of them belong in the paper. Specially as the bias and SDD are something most people can relate to, and that the intuitive way is to interpret them as describing a Gaussian distribution.

Therefore the paper is updated with bias, SDD, Skew, median and to help the dicussion and interpretation of the statistics also IQR (interquartile range), RMSE (root mean square error) and mode where included. To give the potential users more quantitative information on the errors to expect and to help the discussion also percentage of absolut error above 0.25, 0.5, 1 and 2 km were included.

The extended discussion and result sections where combined into one section with several subsections: Validation with CALIOP top layer pressure, Discussion of statistics measures for non-Gaussian error distributions, Validation with CALIOP and CPR (CloudSat) height, Validation separated for low, medium and high level clouds, Validation with CALIOP separated for different cloudtypes, Geografical aspects of the NN-CTTH performance and Future work and challenges.

2.7 Referee comment:

Pg. 8, 14: What is the motivation for comparing with CloudSat? Is this a better reference? If so, why use CALIPSO? If not, why is it here? How were the matches made on the larger CPR footprint? Are there sampling differences between CALIOP and CPR? The CPR often misses the top portions of ice clouds and has difficulty detecting clouds with small particles. If the biases discussed earlier are known, the CPR information might be useful if the results are interpreted more in the discussion section. Also, what is the vertical resolution of CloudSat? Would that impact the differences?

Reply:

The CloudSat validation are included to get an independent source of validation, not better just different. We have improved the discussion regarding this, see also reply to comment 2.9. Nearest neighbour matching is used; we have added this information in the article as the description of the matching method was missing.

Clouds not detected at all by CPR (CloudSat) are not a problem as it simply means that we will have less data. That the CPR often misses the top portions could partly explain why results are not improving for NN-MetImage and NN-MetImage-NoCO$_2$ (compared to NN-MERSI-2) when validating with CloudSat. We have added this discussion. The vertical resolution of CPR (CloudSat) is 0.5km this means that we should expect MAE higher than 250m, this information is also added.

2.8 Referee comment:

Pg.8, 26: The plots are distributions of the differences. Bias is the average of those differences. Please correct.

Reply:

We have correct that.

2.9 Referee comment:

Sec. 5 The discussion section is very thin. There is a paucity of what the results shown in the figures and table might mean. For example, what do the differences computed using two different references, CALIOP and CPR, tell us? All samples, except in polar regions are taken in midday or near midnight for Aqua. Could there be any diurnal impacts of training only with this dataset? What happens if the neighbouring pixel is turned off in the training? The conclusions state that that is an important input. Can its impact be quantified to support that conclusion?

Reply:

We thank Referee 1 for the suggestions and valuable comments that will help to improve the discussion.

The usage of two validation truths strengthens our results. The CloudSat results confirm that the improvements are not only due to that the neural networks have learnt to replicate errors of CALIOP. (For the argumentation let us pretend that CALIOP would always place clouds at 5km height if the surface pressure is 1000 hPa, a neural network could learn this but it would not really improve the accuracy of the retrieved cloud top height). Considering the large improvement it was not an alarming risk that the neural network was learning only to mimic CALIOP errors, but with the independent validation truth CloudSat this is confirmed. We have better motivated the inclusion of CloudSat in the paper.

What happens if the neighbouring pixels are not used is better described. We have discussed these results in more detail to support better the statement in the conclusion.

There might be diurnal impact not captured in the current dataset. However results are valid for Aqua which we trained for. Applying similar neural networks to other sensors with different filter functions and ECT will require additional work or validation not in the scope of this paper.

2.10 Referee comment:

Pg. 9, 22: It seems that using matches with Terra will not help much in the non-polar regions. Is this a realistic possibility given the orbital differences?

Reply:

As latitude is not used as a variable, data for higher satellite zenith angles included for Polar regions could help also in non-Polar regions. However it might be that the high latitude matches will not help the network the if varity of weather situations and cloud heights at high latitudes are too small. This must be tested. We have extend the discussion regarding adding Terra matches.

2.11 Referee comment:

Pg. 9, 30: This section is where the futher work on the sources of error (e.g., various cloud types) could be presented. It would help the discussion considerably.

Reply:

The Validation with CALIOP separated for different cloudtypes cloudtypes where included to answer the question on sources of errors from different cloud types. This was not possible to do with version 3 of CALIOP data as several of the classes of the feature classification are empty for CALIOP version 3 data. Therefore the validation was updated to use CALIOP version 4 data. As the validation with CALIOP are now done with the latest version, also the CPR (CloudSat) was updated to use the most recent version. The discussion and result section where merged into one section with several subsections. And the validation for different cloud types where included in section 4.5 *Validation with CALIOP separated for different cloudtypes*.

2.12 Referee comment:

Sec. 6. More analysis in the discussion section would help flesh out this section.

Reply:

We have extended conclusion section, reflecting what was added to the Results and Discussion section.