

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.

Changes in manuscript:

- page 1 line: 2
- page 2 line: 27

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 corrected 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.

Changes in manuscript:

- page 1 line: 3-5, 7-12
- page 2 line: 7, 16-23, 27-28, 31-33
- page 3 line: 11-12, 17, 20
- page 4 line: 4, 15-16
- page 5 line: 31-33
- page 7 line: 5-6
- CloudSat changed to CPR (CloudSat) many occurrences

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° .

Changes in manuscript:

- page 4 line: 4-5, 10, 17
- page 9 line 31-32 made clear what was meant by near NADIR observations.

Reply:**2.4 Referee comment:**

[Sec. 3.2 pg. 4, 25: while the CO₂ absorbing band is generally referred to as the 15 – \$\mu\text{m}\$ band, the MODIS channels are in the 13.3 – 14.4 \$\mu\text{m}\$ range.](#)

Reply:

We have corrected the channel ranges mentioned.

Changes in manuscript:

- page 5 line: 29

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 were 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).

Changes in manuscript:

- page 6 line: 6-7

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 distributions are non-Gaussian. We were 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 discussion and interpretation of the statistics also IQR (interquartile range), RMSE (root mean square error) and mode were included. To give the potential users more quantitative information on the errors to expect and to help the discussion also percentage of absolute error above 0.25, 0.5, 1 and 2 km were included.

The extended discussion and result sections were 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, Geographical aspects of the NN-CTTH performance and Future work and challenges.

Changes in manuscript:

- Discussion of measures and result added to abstract at page 1 line: 17-24.
- Removed lines from abstract to keep it short at page 1 line 27 and page 2 line: 1-2.

- Moved lines to page 8 line 27-33 (from page 13 line 33-35 and page 14 line 1-4) as they make more sense here after the reorganization of the Result section.
- Reformulated (see answer to 2.21 to Reviwer 2) and moved lines to page 9 line 1-5 (from page 15 line 4-6) as it makes more sense here.
- This is most of the part of the old Results section concerning pressure is now found in Section 4.1 at page 9. Notice that lines 28-30 where moved to line 9-11.
- The new section with discussion of measures for non-Gaussian error distributions are added at page 10 and 11 Section 4.2.
- The old part of the Result section that were treating height is now in Section 4.3 and 4.4 page 12-14. Results of the additional measures (mode, median RMSE etc.) are included here.
- The geographical aspects of the NN-CTTH now as a separate section at page 14 Section 4.6.
- Removed page 15 line 8-9 due to section reorganization.
- Moved page 15 line 10-11 moved to conclusion page 17 line 14-16
- Moved page 15 line 13-15 to page 16 line 4-6.
- Moved one sentence from page 16 line 22-23 to page 17 line 16-18. Also reformulated it (See answer 2.9).
- Conclusions regarding the added masures included at page 16 line 29-33 and at page 17 line 1-2.
- Added conclusions related to non-Gaussian error distributions at page 17 line 8-13.
- Table 6 and Table 7 replaced with four new tables pages 27-32.
- The figure referenced in section 4.2 is added at page 41.

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-NoCO₂ (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.

Changes in manuscript:

- Matching described at page 6 line 4-5.
- Motivation for using CPR (CloudSat) included at page 8 line 20-26.
- Discussion of differences in result between CPR (CloudSat) and CALIOP added at page 12 line 11-16, page 13 line 6-11 and page 14 line 5-10.

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.

Changes in manuscript:

- page 35 line: 1

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.

Changes in manuscript:

- page 8 line: 20-26

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.

Changes in manuscript:

- page 9 line: 18-20
- The conclusion was moved and reformulated at page 17 line 16-18 (moved from page 16 line 22-23).

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 variety 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.

Changes in manuscript:

- page 15 line: 20-23

2.11 Referee comment:

Pg. 9, 30: This section is where the further 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 were 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 were merged into one section with several subsections. And the validation for different cloud types were included in section 4.5 *Validation with CALIOP separated for different cloudtypes*.

Changes in manuscript:

- Update of used version page 4 line 12-14, 18.
- A new Section 4.5 with results for different cloud types at page 14.
- Updates to the conclusion section at page 16 line 26-27 and page 17 line 5-6.
- Included the results for different cloud types in a new table at page 33.
- Figures, Tables and results concerned were updated to use the new versions.

2.12 Referee comment:

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

Reply:

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

Changes in manuscript:

- page 16 and 17 Section 5
- One sentence at page 17 line 29-20 were reformulated to be clearer.

Reply to anonymous Referee 2 comments to *Neural network cloud top pressure and height for MODIS*

Nina Håkansson et al.

1 Overall quality of the discussion paper ("general comments"):

1.1 Referee comment:

In the paper a novel retrieval of cloud top pressure and height using neural networks is presented. The presented retrieval technique is state of the art and an accurate technique. To account for different availabilities of channels on different satellites, a few modification of the neural network are investigated revealing the information content of the different channels. The new algorithms are compared to two reference algorithms, the CTTH algorithm of the NWCSAF PPS-v2014 and the MODIS collection 6 L2 height product. Additionally the algorithms is compared to CALIOP and CPR measurements. The quality of the algorithm is evaluated in terms of the mean absolute error (MAE). The improved quality of the results is impressive. From my point of view, I would request for at least another quality measure like standard deviation (similar to Tables 6 and 7). In overall, good work!

Reply:

We thank Referee 2 for this positive comment, and for the other valuable comments that will help us improve the paper further.

Regarding adding standard deviation, which was also requested by Referee 1, we chose the MAE as evaluation metric, over bias and standard deviation of differences (SDD), for many good reasons. However, some of the good reasons became clear to us first when we where faced with the request to include them in the article. Most important is that including bias and SDD of the error distribution intuitively gives the reader the mental picture of a Gaussian error distribution, centred at the bias. However we are dealing with skewed and even bimodal distributions as shown in Figure 2 and the mean is not at the centre of the distribution; i.e. the bias is not located at the peak of the error distribution.

The overall standard deviation is much affected by the largest errors. Some large errors are expected due to the differences between the passive and active sensors and the different FOV (field of view). Therefore we argue that the MAE is a better measure of variation of the error compared to SDD. The largest errors are of course also interesting, but when investigating these some care should be made to separate true errors from expected differences due to for example cloud edges in the FOV. We have included also the interquartile range (IQR) as it is more robust measures of variability less sensitive to outliers compared to SDD.

As the bias and SDD are traditionally used when evaluating cloud top height retrieval algorithms these are included along with the discussion of why they are not so useful or even misleading. To help the discussion also IQR, RMSE (root mean square error), mode, median and percentage of absolute errors above 0.25, 0.5, 1 and 2 km were included.

Changes in manuscript:

- See reply to Referee 1 point 2.6.

2 Individual scientific questions/issues ("specific comments")

2.1 Referee comment:

p1 line 23: CTH might also be used in data assimilation of atmospheric motion vectors.

Reply:

We have added this to the text.

Changes in manuscript:

- page 2 line 8

2.2 Referee comment:

Introduction: A short description of the traditional techniques to retrieve cloud top pressure and height could be added to the introduction. (or cite an overview paper like Hamann et al. "Remote sensing of cloud top pressure/height from SEVIRI: analysis of ten current retrieval algorithms." Atmospheric Measurement Techniques 7.9 (2014): 2839-2867.)

Reply:

We have added the suggested reference.

Changes in manuscript:

- page 2 line 12-13

2.3 Referee comment:

The introduction should motivate, why it is expected that using machine learning, in particular neural networks, could improve the expected results.

Reply:

Many CTH retrieval algorithms including MODIS-C6 and PPS-v2014 include some fitting of temperatures to NWP temperature profiles. This is most difficult in the case of inversions both as one temperature occurs at several pressure heights in the profile but also as the inversions are often not captured accurately enough in the NWP temperature profile. Many different techniques are used to deal with this, for example PPS-v2014 will place the cloud at the inversion height if the temperature is not more than 0.5 to 2K lower than the temperature at the inversion. MODIS-C6 has another approach using climatological lapse rates over sea for clouds likely to be low. These kinds of fitting techniques a statistical machine learning technique could probably do better. A motivation is included in the introduction.

Changes in manuscript:

- page 3 line 1-3

2.4 Referee comment:

Merge chapter 2.1.1 into chapter 3.2. (and skip the sentence (p3 line 10) “The MODIS Collection 6 cloud product were used as an independent. . .”, you said that before).

Reply:

We have removed section 2.1.1 and moved the information from it to section 2.1 and removed the repeated information.

Changes in manuscript:

- page 3 line 28, 30, 31

2.5 Referee comment:

Chapter 2.2 Add a short sentence, why you chose the CALIOP 1km product and not also 5km or 10km product which are more sensitive to optically thin clouds.

Reply:

The 1km CALIOP product was selected because it has the resolution closest to the MODIS resolution. It is expected that the thinnest cloud seen by CALIOP lidar is invisible to the passive imagers, so it should not be a problem that the thinnest clouds are missing in the 1km data. However we have also done some tests using AVHRR-GAC data and CALIOP 5km (Version 4) resolution for training (this is outside the scope of this article). The first tests show that results improve if the thinnest (0.05 or 0.1 in optical depth) clouds are excluded from the training. If these networks (trained on AVHRR-GAC) are applied on the validation data (MODIS) of this article the MAEs of the retrievals are between 76hPa to 79hPa. We have added a sentence about why 1km data was chosen.

Changes in manuscript:

- page 4 line 12

2.6 Referee comment:

Chapter 2.4 add the version number of the ECMWF model and add product name and version number of the OSISAF data used in this study.

Reply:

We have added the versions and product names in section 2.4.

Changes in manuscript:

- page 4 line 24-26

2.7 Referee comment:

You might consider to add the PPS-v2014 and MODIS C6 algorithm to table 3 and 4.

Reply:

We have added them to Table 4, this will give the clear view of what channels are used for which method also for them. We suggest that they are not included in Table 3 as it describes *Network specific variables*. There was also an error in table 4, the NN-OPAQUE uses channel 12 μ m as described in Table 3 not channel 11 μ m. We have correct this as well and sorted the columns from lowest to highest wave length.

Changes in manuscript:

- page 25 Table 4

2.8 Referee comment:

Please make the order of algorithms in table 3, 4 and 5 consistent.

Reply:

We did this in the first revision, and can not find any remaning inconsistencies.

2.9 Referee comment:

p5 line 5: how often is a pressure lower than 70 hPa retrieved?

Reply:

It varies with each network from 0 up to 0.05%, we have added this to the text.

Changes in manuscript:

– page 6 line 9-10

2.10 Referee comment:

p5 line 10: Why did you choose this number of levels? Is it sufficient to use 6 levels to represent the boundary layer inversions or other small scale features?

Reply:

Five of the levels (surface, 950, 850, 700, 500) were already used in the PPS-v2014, so this was our starting point. We tested to use the troposphere pressure, but then the networks became very sensitive to the type of NWP-data used. Instead we added the 250hPa level to have one more high level. We did test increasing the number of levels near the ground by adding levels at 800, 900 and 1000hPa, but the improvement was not large enough to motivate the extra computational time. One common problem for cloud height retrieval algorithms is that inversions are not represented accurately enough in the NWP data. As mentioned previously, MODIS-C6 instead uses climatological lapse-rates over sea to avoid this problem, other algorithms use sharpening techniques at the inversion. So it is not clear that more levels which would better represent the inversions in the NWP data would improve the neural network results, but this could be further investigated.

2.11 Referee comment:

p5 line 21: do you skip non cloudy pixels in the 5x5 pixel standard deviations?

Reply:

No, all pixels are included.

2.12 Referee comment:

p5 line 20: the $B_{3.7}$ has a solar component. Did you correct for this during day/night?

Reply:

No correction was made, that channel is used just like the others. The information that the solar component is not treated was added to the text.

Changes in manuscript:

– page 9 line 26-27

2.13 Referee comment:

p6 chapter 3.3.2: You chose to use specific days for training and others for validation. Given that you only use a limited number of days, wouldn't it be more to randomly select independent pixels from all available dates for training, validation, and testing to represent a larger variety of weather situations?

Reply:

Unfortunately all pixels in the dataset we have are not independent. A typical cloud is much larger than one pixel; there could be hundreds of pixels with almost identical data in the datasets. If we would randomly select independent pixels from all days for each dataset we would in practice use the same data all the time. This would cause the neural network to overtrain as the *during training validation data* would be the same as the *training data*. And in the last validation step results would be overly positive as also the *validation data* would be in practice the same as the *training data*.

Random sampling from all data has been suggested to us previously. We have therefore added the discussion to the article why it is not possible.

Changes in manuscript:

– page 3 line 25-27

2.14 Referee comment:

p 6 line 19: Did you test other configurations that 30/15 neurons in the first/second layer? If yes, how was the performance?

Reply:

We did one test with 20/15 this network (NN-AVHRR) was 1 hPa worse and one with 30/45/45 (NN-AVHRR) this was 2.5hPa better but also took 5 times as long time to retrieve pressure. This information have been included in the text.

Changes in manuscript:

– page 15 line 28-31

2.15 Referee comment:

Chapter 3.3.3: May batch size and momentum be changed during the training process?

Reply:

No they can't.

2.16 Referee comment:

p 7 line 27: consider to discuss the solar component of the 3.7 μm channel. To my opinion this NN could perform better when corrected for that (e.g. adding the solar time as input variable).

Reply:

We considered adding the sun zenith angle as variable; however we can not decide how the neural network would use it. In the data we do not have all sun zenith angles present globally. It could be that the neural network would use the sun zenith angle to decide that during this time of day clouds of a particular height are most common. It does not have to be bad though and can be tested in future studies. The performance of NN-AVHRR1 could probably be improved if the solar component of 3.7 is treated explicitly. We have added discussion about the solar component of 3.7 μm .

Changes in manuscript:

- page 9 line 26-27

2.17 Referee comment:

p8 line 5: Maybe express it positively: All NN can reproduce a clear bi-modal pdf very similar to CALIPSO, the pdf of PPS-v2014 deviates from this shape . . .

Reply:

We have changed the formulation, thank you for the suggestion.

Changes in manuscript:

- page 10 line 3-7

2.18 Referee comment:

p8 line 7: It is written “for the best performing network”. Did you train several networks for one channel configuration? If so, could you describe the number of trained networks in chapter 3.2.2, please?

Reply:

With the *the best performing networks* we meant that the NN-NWP, NN-OPAQUE, NN-BASIC and NN-BASIC-CIWV was excluded. We have clarified this in the text.

Changes in manuscript:

- page 10 line 22-23

2.19 Referee comment:

p 8, line 11: according to my table 6, the NN-MetImage is better than the NN-MetImage- NoCO₂.

Reply:

Yes it does! However the NN-MetImage does not perform well for higher satellite zenith angles. Only networks that perform well for all satellite zenith angles are discussed in this sentence. We have added a sentence to make it clearer.

Changes in manuscript:

- page 12 line 21-22

2.20 Referee comment:

p 8, line 15: do you have an idea, why the MAE against CPR is larger than the MAE against CALIOP for NN-MetImage and NN-MetImage-NoCO₂?

Reply:

This we think it partly because the NN-MetImage and NN-MetImage-NoCO₂ have some skill in predicting very thin high clouds that are not detected by the CPR-Radar. We have added discussion about this.

Changes in manuscript:

- page 12 line 11-16

2.21 Referee comment:

p 9, line 9: could you please describe a bit more in detail the differences seen in Figure 7?

Reply:

We have described the differences in more detail.

Changes in manuscript:

- page 9 line 1-5 (moved from page 15 line 4-6 due to the reorganization of the Result section)

2.22 Referee comment:

Chapter 5 Discussion: Could you also comment on applying your NN technique on geostationary satellites? What would be the main differences/challenges?

Reply:

This technique should not be limited to polar orbiting satellites. As the instrument SEVIRI has the two most important channels at 11 μ m and 12 μ m it should be possible to apply the technique to SEVIRI data. More data (in terms of number of days) compared to MODIS may be needed to produce enough matches. As the SEVIRI resolution is coarser results might be degraded compared to MODIS. Matches of SEVIRI with CALIOP will occur at many different satellite zenith angles. This might make it possible to use the CO₂ channel on SEVIRI to improve results without losing performance skill at high satellite zenith angles. We have commented on using the NN-CTTH technique for geostationary satellites in the paper.

Changes in manuscript:

- page 16 line 16-19

A compact listing of purely technical corrections ("technical corrections": typing errors, etc.)

2.23 Referee comment:

- p5 line 29 and thereafter: don't write CO₂ with cursive letters
- consider to write NoCO₂ (in MetImageNoCO₂) not in cursive letters.

Reply:

We have kept the notation with subscript but without cursive letters. All CO₂ are updated to be written without cursive letters.

2.24 Referee comment:

Figure 1 (p21): consider to have the figures in the same order as the algorithms are mentioned in table 3, 4, and 5.

Reply:

This is a reasonable request. However it is also nice to have the two AVHRR based algorithms next to each other so they can be compared. Also this order makes the two networks performing bad at high satellite zenith angles appear on the last row. And changing the order increases the risk to mix them up in later references. If someone refers to the bad satellite zenith angle behaviour in Figure 2 (h) in the discussion paper, and an interested reader by accident finds the final revised paper (assuming

there will be one) and finds the result for NN-AVHRR1 in that sub figure that is not good. As there are also good reasons to keep the current order of sub figures we argue that their order should not be changed.

2.25 Referee comment:

- p 8, line 34 and following: avoid the abbreviation NN-CTTH, e.g. change: that the NN-CTTH all have -> that all NN retrievals have . . .
- avoid NN-CTTH abbreviation (which one do you mean? all NN retrievals or NN-MetImage or another one?)
- p 10, line 12: avoid NN-CTTH -> specify which retrieval you referring to

Reply:

We have kept the NN-CTTH as the name for the neural network method in the paper. We have present it as the name and avoided using it where it might be confusing.

Changes in manuscript:

- page 8 line 29 Reformulated from *NN-CTTH all have*. Note that the lines were moved from page 13 line 35.
- page 16 line 24 Clarified that NN-CTTH means all networks in this sentence.
- page 16 line 33 Reformulated to not use NN-CTTH abbreviation.

2.26 Referee comment and changes in manuscript:

C: Please check consistent spelling of NWC SAF (e.g. p2 line 11) and NWCSAF (e.g. p1 line 19)

R: Changed to use NWC SAF page 1 line 7 and page 17 line 23.

C: Please check consistent spelling of PPS-2018 (e.g. p1 line 19) and PPS-v2014 (e.g. p2 line 23)

R: Sentence removed page 2 line 2.

C: Please check the space between numbers and units and the typeset of the units.

R: Done

C: Try to reduce number of paragraphs in the abstract, e.g. p1 line 8 is a one sentence paragraph.

R: The one sentence paragraph was merged with the following paragraph. However the new paragraph about the statistical measures means that there are still 5 paragraphs in the abstract.

C: please check capital letters, e.g. Neural network (p3 line 26), neural network (p2 line 14) or Neural Network.

R: Updated to use neural network without capital letters at page 4 line 23.

C: p3 line 22: (change . to ,) . . . of the networks, see Table 1 for selected Dates

R: Changed at page 4 line 19.

C: Move p4 line 9-14 to line 6.

R: Moved page 5 line 15-18 to line 7-10.

C: p4 line 27: introduce abbreviation GDAS (as written in line 30)

R: Abbreviation introduced page 5 line 31-32.

C: p4 line 28: add “the”: . . . and the PFAAST radiative transfer model. . .

R: Added “the“ at page 5 line 32.

C: p5 line 3 add: The “uppermost cloud” top layer. . .

R: Added at page 6 line 6.

C: p5 line 11: reformulate “much of what”

R: Reformulated page 6 line 15-16.

C: p5 line 19: introduce physical unit “B” (in the lines before)

R: Unit introduced at page 6 line 15.

C: p5 line 23: B₁₁ “for” neighboring pixels -> B₁₁ “of the” neighboring Pixels

R: Changed at page 6 line 28.

C: p5 line 25: avoid brackets

R: Reformulated without brackets at page 6 line 29-30.

C: p 15 line 7, add “and”: BT for water vapour channels at 6.7 “and” 7.3 μm

R: Added or at page 23 line 7.

C: p15 line 9/10/11, remove “.” at the end of first entry, e.g. BT differences “”

R: Removed the “.” at page 23 line 9-11.

C: p5 line 29 and thereafter: don't write CO2 with cursive letters

R: Changed to use non-cursive letters.

C: p 6 line 19: hidden layer "for" the neural network -> hidden layer "of" the neural network

R: Changed at page 7 line 25.

C: p 6 line 17-32: reduce the number of paragraphs. Don't create one sentence paragraphs.

R: Paragraph 2 and 3 of Section 3.3.3 were merged.

C: p 7 line 18: write Ciwv in cursive letters

R: Changed at page 9 line 14.

C: p 7 last line: N-VIIRS -> NN-VIIRS

R: Corrected at page 9 line 33.

C: Figure 7 (p27): Could you please add a color scale instead of describing it with words.

R: Color scale added and the description with words removed at page 40.

C: p 9, line 11 and thereafter: cloud heights -> cloud "top" heights,

R: Done

C: p 10, line 15: The NN CTTH retrievals all have better results for low, medium and high clouds . . . (Clouds don't show results. . .)

R: Reformulated at page 16 line 26.

2.27 Additional changes:

Three additional sentences were slightly reformulated to be clearer:

- page 7 line 18
- page 9 line 12
- page 15 line 1

Neural network cloud top pressure and height for MODIS

Nina Håkansson¹, Claudia Adok², Anke Thoss¹, Ronald Scheirer¹, and Sara Hörnquist¹

¹Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden

²Regional Cancer Center Western Sweden, Gothenburg, Sweden

Correspondence to: Nina Håkansson (nina.hakansson@smhi.se)

Abstract.

Cloud top height retrieval from imager instruments is important for ~~Nowcasting~~ nowcasting and for satellite climate data records. A neural network approach for cloud top height retrieval from the imager instrument MODIS (Moderate-resolution Imaging Spectro-radiometer) is presented. The neural networks are trained using cloud top layer pressure data from the CALIOP (Cloud-Aerosol Lidar with Orthogonal Polarisation) dataset.

Results are compared with two operational reference algorithms for cloud top height: the MODIS Collection 6 level 2 height product and the cloud top temperature and height algorithm (~~CTTH~~) in the 2014 version of the ~~NWCSAF~~ NWC SAF (EUMETSAT (European Organisation for the Exploitation of Meteorological Satellites) Satellite Application Facility for nowcasting and very shortrange forecasting) PPS (Polar Platform System (PPS-v2014)). All three techniques are evaluated using both CALIOP and CPR (CloudSat) (Cloud Profiling Radar for CloudSat (CLOUD SATellite)) height.

Instruments like AVHRR ~~and VHRS~~ (Advanced Very High Resolution Radiometer) and VIIRS (Visible Infrared Imaging Radiometer Suite) contain fewer channels useful for cloud top height retrievals than MODIS, therefore several different neural networks are investigated to test how infrared channel selection influences retrieval performance.

Also a network with only channels available for the AVHRR1 instrument is trained and evaluated. To examine the contribution of different variables, networks with fewer variables are trained. It is shown that variables containing imager information for neighbouring pixels are very important.

~~Overall results for the neural network height retrievals are very promising.~~ The error distributions of the involved cloud top height algorithms are found to be non-Gaussian. Different descriptive statistic measures are presented and it is exemplified that bias and SD (standard deviation) can be misleading for non-Gaussian distributions. The median and mode are found to better describe the tendency of the error distributions and IQR (interquartile range) and MAE are found to give the most useful information of the spread of the errors.

For all descriptive statistics presented MAE, IQR, RMSE (root mean square error), SD, mode, median, bias and percentage of absolute errors above 0.25, 0.5, 1 and 2 km the neural network perform better than the reference algorithms both validated with CALIOP and CPR (CloudSat). The neural networks using the brightness temperatures at 11 μm and 12 μm show at least ~~33% (or 62732 % (or 623 m))~~ lower mean absolute error (MAE) compared to the two operational reference algorithms when validating with CALIOP height. Validation with CPR (CloudSat) height gives at least ~~25% (or 433 % (or 430 m))~~ reduction of MAE. ~~For the network trained with a channel combination available for AVHRR1, the MAE is at least 542~~ better when

~~validated with CALIOP and 414 when validated with CPR (CloudSat) compared to the two operational reference algorithms. The NWCSAF PPS-2018 release will contain a neural network based cloud height algorithm.~~

1 Introduction

The retrieval of cloud top temperature, pressure and height from imager data from polar orbiting satellites is used both as a vital product in global cloud climatologies (Stubenrauch et al., 2013) and for nowcasting at high latitudes where data from geostationary satellites are either not available or not available in sufficient quality and spatial resolution. Cloud top height products from VIS/IR ([visible/infrared](#)) imagers are used in the analysis and early warning of thunderstorm development ~~and~~, for height assignment in aviation forecasts [and in data assimilation of atmospheric motion vectors](#). The cloud [top](#) height can serve as input to mesoscale analysis and models for use in nowcasting in general, or as input to other satellite retrievals used in nowcasting (e.g. cloud micro physical properties retrieval, or cloud type retrieval). It is important that climatologists and forecasters have reliable and accurate cloud [top](#) height products from recent and past satellite measurements.

~~There are different traditional techniques to retrieve cloud top height see Hamann et al. (2014) for a presentation of ten cloud top height retrieval algorithms applied to the SEVIRI (Spinning Enhanced Visible Infra-Red Imager).~~ Several algorithms to retrieve cloud top height from polar orbiting satellites are available and used operationally for nowcasting purposes or in cloud climatologies. These include the CTTH (cloud top temperature and height) from the PPS (Polar Platform System) package (Dybbroe et al., 2005), which is also used in the CLARA-A2 ~~climate data record of CMSAF (EUMETSAT (CM SAF (EUMETSAT (European Organisation for the Exploitation of Meteorological Satellites) Satellite Application Facility for Climate Monitoring) cloud, albedo and surface radiation dataset) climate data record~~ (Karlsson et al., 2017), ACHA (~~cloud height algorithm~~) ~~Algorithm Working Group (AWG) Cloud Height retrieval Algorithm~~ used in PATMOS-x ([Pathfinder Atmospheres - Extended](#)) (Heidinger et al., 2014), CC4CL ([Community Cloud Retrieval for Climate](#)) used in ESA ([European Space Agency](#)) Cloud_CCI ([Cloud Climate Change Initiative](#)) (Stengel et al., 2017), MODIS ([Moderate-resolution Imaging Spectro-radiometer](#)) Collection-6 algorithm (Ackerman et al., 2015) and the ISCCP ([International Satellite Cloud Climatology Project](#)) algorithm (Rossow and Schiffer, 1999).

We will use both the MODIS Collection-6 (MODIS-C6) and the version 2014 CTTH from PPS (PPS-v2014) as references to evaluate the performance of neural network based cloud [top](#) height retrieval. The MODIS-C6 algorithm is developed for the MODIS instrument. The PPS, delivered by the NWC SAF (EUMETSAT Satellite Application Facility for ~~Nowcasting nowcasting~~ and very shortrange forecasting), is adapted to handle data from instruments AVHRR, ~~VHRS (Advanced Very High Resolution Radiometer)~~, VIIRS ([Visible Infrared Imaging Radiometer Suite](#)) and MODIS.

Artificial neural networks are widely used for non-linear regression problems, see for example Gardner and Dorling (1998), Meng et al. (2007) or Milstein and Blackwell (2016) for neural network applications in atmospheric science. In CC4CL a neural network is used for the cloud detection (Stengel et al., 2017). Artificial neural networks have also been used on MODIS data to retrieve cloud optical depth (Minnis et al., 2016). The COCS ~~algorithm (cirrus optical properties derived from CALIOP and SEVIRI algorithm during day and night) algorithm~~ uses artificial neural networks to retrieve cirrus cloud optical thickness

and cloud top height for the SEVIRI instrument (Kox et al., 2014). Considering that neural networks in the mentioned examples have successfully derived cloud properties, and that cloud top height retrievals often include fitting of brightness temperatures to temperature profiles, neural network can be expected to retrieve cloud top pressure for MODIS with some skill.

One type of neural network is the multilayer perceptron described in (Gardner and Dorling, 1998) which is a supervised learning technique. If the output for a certain input, when training the multilayer perceptron, is not equal to the target output an error signal is propagated back in the network and the weights of the network are adjusted resulting in a reduced overall error. This algorithm is called the back-propagation algorithm.

In this study we will compare the performance of back-propagation neural network algorithms for retrieving cloud top height (NN-CTTH) with the CTTH algorithm from PPS version 2014 (PPS-v2014) and MODIS Collection 6 (MODIS-C6) algorithm. Several networks will be trained to estimate the contribution of different training variables to the overall result. The networks will be validated using both CALIOP ~~and CloudSat~~ (Cloud-Aerosol Lidar with Orthogonal Polarisation) and CPR (CloudSat) (Cloud Profiling Radar for CloudSat (CLOUD SATellite)) height data.

In section 2 the different datasets used are briefly described and in section 3 the three algorithms are described. Results are presented ~~in section ??, and~~ discussed in section ~~4.1-4~~ and final conclusions are found in section 5.

15 2 Instruments and data

For this study we used data from the MODIS instrument on the polar orbiting satellite Aqua in the A-Train, as it is co-located with both CALIPSO ~~and (Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations)~~ and CloudSat at most latitudes and has multiple channels useful for cloud top height retrieval.

2.1 Aqua - MODIS

20 The MODIS (~~Moderate-resolution Imaging Spectro-radiometer~~) is a spectro-radiometer with 36 channels covering the solar and thermal spectra. We are using level 1 data from the MODIS instrument on the polar orbiter Aqua. For this study the MYD021km (MODIS Science Data Support Team, 2015a) and MYD03 (MODIS Science Data Support Team, 2015b) for all orbits from 24 dates were used (1st and 14th of every month of 2010). The data were divided into four parts which were used for training, validation during training (used to decide when to quit training), testing under development (used to test different combinations of variables during prototyping) and final validation. The data contains many pixels that are almost identical, because a typical cloud is larger than one pixel. Therefore randomly dividing the data into four datasets is not possible as this would in practice give four identical datasets, which would cause the network to over-train. See Table 1 for distribution of data.

2.1.1 ~~Aqua -- MODIS Collection 6 cloud products~~

The MODIS Collection-6 climate data records produced by the National Aeronautics and Space Administration (NASA) Earth Observation System ~~are using data from the MODIS sensor. The MODIS Collection 6 cloud products were used as an independent algorithm with which to compare the performance of the NN-CTTH. The 1-km was used for comparison. The~~

1km cloud top height and cloud top pressure from the MYD06_L2-product (Ackerman et al., 2015) for the dates in Table 1 were used.

2.2 ~~CALIPSO~~-CALIOP

~~The CALIOP (Cloud-Aerosol Lidar with Orthogonal Polarisation)~~ The satellite zenith angles for MODIS when matched with CALIOP varies between 0.04 and 19.08° and then matched with CPR (CloudSat) it varies between 0.04 and 19.26°.

2.2 CALIOP

The CALIOP lidar on the polar orbiting satellite CALIPSO is an active sensor and therefore more sensitive to particle conglomerates with low density than typical imagers. The horizontal pixel resolution is ~~0.07km x km x 0.333km~~ km, this means that when co-locating with MODIS one should remember that CALIOP samples only a small part of each MODIS pixel. The ~~CALIOP~~ vertical resolution for CALIOP is 30 m. The viewing angle for CALIOP is 3°. The CALIOP 1km Cloud Layer product (version 3) data were used (for the dates, see Table 1) as the truth to train the networks against, and for validation of the networks. The 1km product was selected because the resolution is closest to the MODIS resolution. For training version 3 was used and for validation version 4, be able to access the improved cloud type information in the feature classification flag in version 4.

15 2.3 ~~CloudSat~~-CPR (CloudSat)

The CPR (~~Cloud Profiling Radar for~~ CloudSat) is a radar ~~on CloudSat~~ which derives a vertical profile of cloud water. Its horizontal resolution is ~~1.4km x km x 3.5km~~ km, and its vertical resolution is 0.5 km and the viewing angle is 0.16°. The CPR ~~product 2B-GEOPROF-R04 (CloudSat) product 2B-GEOPROF-R05~~ (Marchand et al., 2008) was used as an additional source for independent validation of the networks. ~~See, see~~ Table 1 for selected dates. The validation with ~~CloudSat~~ CPR (CloudSat) will have a lower percentage of low clouds compared to CALIOP because ground clutter is a problem for space borne radar instruments.

2.4 Other data

Numerical weather prediction (NWP) data are needed as input for the PPS-v2014 and the ~~Neural-neural~~ network algorithm. In this study the operational 91-level short-range archived forecasted NWP data from ECMWF (European Centre for Medium-range Weather Forecasting) were used. The analysis times at 00:00 and 12:00 was used and the forecast times (6, 9, 12 and 15 h). Under the period IFS cycles Cycle 35r3, 36r1, 36r3 and 36r4 were operational. Also ice maps (OSI-409 version 1.1) from OSISAF (Satellite Application Facility on Ocean and Sea Ice) were used as input for the PPS cloud mask algorithm.

3 Algorithms

3.1 PPS-v2014 cloud top temperature and height

The cloud top height algorithm in PPS-v2014, uses two different algorithms for cloud top height retrieval, one for pixels classified as opaque and another for semi-transparent clouds. The reason for having two different algorithms is that the straight forward opaque algorithm can not be used for pixels with optically thin clouds like cirrus or broken cloud fields like cumulus. The signals for these pixels are a mixture of contributions from the cloud itself and underlying clouds and/or the surface.

The ~~retrieval~~ algorithm uses a split-window technique to decide whether to apply the opaque or semi-transparent retrieval. All pixels with a difference between the 11 μm and 12 μm brightness temperatures of more than 1.0 K are treated as semi-transparent. This is a slight modification of the PPS version 2014 algorithm where also the clouds classified as non-opaque by cloud type product are considered semi-transparent.

The retrieval for opaque clouds matches the observed brightness temperatures at 11 μm against a temperature profile derived from a short term forecast or (re)analysis of a NWP model, adjusted for atmospheric absorption. The first match, going along the profile from the ground and upwards, gives the cloud top height and pressure. Temperatures colder or warmer than the profile are fitted to, respectively, the coldest or warmest temperature of the profile below tropopause.

The algorithm ~~uses a split-window technique to decide whether to apply the opaque or semi-transparent retrieval. All pixels with a difference between the 11 and 12 brightness temperatures of more than 1.0 K are treated as semi-transparent. This is a slight modification of the PPS version 2014 algorithm where also the clouds classified as non-opaque by cloud type product are considered semi-transparent.~~

The ~~algorithm~~ for semi-transparent pixels uses a histogram method, based on the work of Inoue (1985) and Derrien et al. (1988), which fits a curve to the brightness temperature difference between the 11 μm and 12 μm bands as a function of 11 μm brightness temperatures for all pixels in a segment (32x32 pixels). One parameter of this fitting is the cloud top temperature. The solution is checked for quality (low root mean square error) and sanity (inside physically meaningful interval and not predicted too far from data). The solution is accepted if both tests are passed. The height and pressure are then retrieved from the temperature, in the same way as for opaque clouds. For more detail about the algorithms see SMHI (2015).

PPS height uses the unit altitude above ground. For all comparisons this is transformed to height above mean sea level, using elevations given in the CloudSat CPR (CloudSat) or CALIOP datasets.

3.2 MODIS Collection 6 Aqua Cloud Top Properties product

In MODIS Collection 6 the ~~CO_2~~ -slicing method (described in Menzel et al., 2008) is used to retrieve cloud top pressure using the 13 and 14 μm channels for ice clouds (as determined from MODIS phase algorithm). For low level clouds the 11 μm channel and the IR-window approach (IRW) with a latitude dependent lapse rate is used over ocean (Baum et al., 2012). Over land the 11 μm temperature is fitted against a 11 μm temperature profile calculated from GDAS (Global Data Assimilation System) temperature, water vapour and ozone profiles and ~~PFAAST~~ the PFAAST (Pressure-Layer Fast Algorithm for Atmospheric Transmittance) radiative transfer model are used for low clouds (Menzel et al., 2008). For more details about

the updates in Collection 6 see Baum et al. (2012). Cloud pressure is converted to temperature and height using the National Centers for Environmental Prediction Global Data Assimilation System (Baum et al., 2012).

3.3 Neural network cloud top temperature and height NN-CTTH

Neural networks are trained using MODIS data co-located with CALIOP data. Nearest neighbour matching was used with the pyresample package in the pyTroll project (Raspaud et al., in press). The Aqua and CALIPSO satellites are both part of the A-Train and the matched FOV (field of view) are close in time (only 75s apart). The uppermost top layer pressure variable, for both multi- and single-layer clouds, from CALIOP data was used as training truth. Temperature and height for the retrieved cloud top pressure are extracted using NWP-data. Pressure predicted higher than surface pressure are set to surface pressure. For pressures lower than 70 hPa neither height nor temperature values are extracted. The amount of pixels with pressure lower than 70 hPa varies between 0 and 0.05 % for the networks.

3.3.1 Neural network variables

To reduce sun-zenith angle dependence and to have the same algorithm for all illumination conditions it was decided to use only infra-red channels to train the neural networks. Several different types of variables were used to train the network. The most basic ones were the NWP temperatures at pressure levels (surface, 950, 850, 700, 500 and 250 hPa). This together with the 11 μm or 12 μm brightness temperature (B_{11} or B_{12}) gives the network ~~much of~~ what is needed to ~~predict~~ make a radiance fitting to retrieve cloud top pressure for opaque clouds, although with very coarse vertical resolution in the NWP data. For opaque clouds that are geometrically thin, with little or no water vapour above the cloud, the 11 μm and 12 μm brightness temperatures will be the same as the cloud top temperature. If the predicted NWP temperatures are correct the neural network could fit the 11 μm brightness temperature to the NWP temperatures and receive the cloud pressure (similar to what is done in PPS-v2014 and MODIS-C6). For cases without inversions in the temperature profile, the retrieved cloud top pressure should be accurate. The cases with inversions are more difficult to fit correctly, since multiple solutions exist and the temperature inversion might not be accurately captured regarding its strength and height in the NWP data. For semi-transparent clouds the network needs more variables to make a correct retrieval.

To give the network information on opacity of the pixel, brightness temperature difference variables were included (~~$B_{11} - B_{12}$, $B_{11} - B_{3.7}$, $B_{11} - B_{3.75}$~~ , $B_{11} - B_{3.7}$, $B_{8.5} - B_{11}$). Texture variables with the standard deviation of brightness temperature, or brightness temperature difference, for ~~5x5~~ 5 x 5 pixels were included. These contain information about whether pixels with large $B_{11} - B_{12}$ are more likely to be semi-transparent or more likely to be fractional or cloud edges.

As described in section 3.1, PPS-v2014 uses $B_{11} - B_{12}$ and B_{11} ~~for~~ of the neighbouring pixels to retrieve temperatures for semitransparent clouds. In order to feed the network with some of this information the neighbouring warmest and coldest pixels (~~in brightness temperature 11 in a 5x5 pixel neighbourhood~~) in B_{11} in a 5 x 5 pixel neighbourhood were identified. Variables using the brightness temperature at these warmest and coldest pixels were calculated, for example the 12 μm brightness temperature for the coldest pixel minus the same for the current pixel: $B_{12}^C - B_{12}$, see Table 2 for more information about what variables were calculated.

The surface pressure was also included, which provides the network with a value for the maximum reasonable pressure. Also the brightness temperature for the $\text{CO}_2\text{-CO}_2$ channel at $13.3\ \mu\text{m}$ and the water vapour channels at $6.7\ \mu\text{m}$ and $7.3\ \mu\text{m}$ were included as variables. The $\text{CO}_2\text{-CO}_2$ channel at $13.3\ \mu\text{m}$ is used in the $\text{CO}_2\text{-CO}_2$ -slicing method of MODIS-C6 and should improve the cloud [top](#) height retrieval for high clouds.

- 5 The instruments AVHRR, VIIRS, MERSI-2, ~~MetImage~~ ([Medium Resolution Spectral Imager -2](#)), [MetImage \(Meteorological Imager\)](#) and MODIS all have different selections of IR channels. Most of them have the $11\ \mu\text{m}$ and $12\ \mu\text{m}$ channels. The first AVHRR instrument AVHRR1 had only two IR channels at $11\ \mu\text{m}$ and $3.7\ \mu\text{m}$ and no channel at $12\ \mu\text{m}$. Networks were trained using combinations of MODIS IR-channels corresponding to the channels available for the other instruments. See Table 3 for specifications of the networks trained. Table 4 gives an overview of what imager channels were used for which network.
- 10 To see how much the different variable types contribute to the result, some basic networks were trained using less or no imager data. These are also described in Table 3. Also one network using only NWP data was included as a sanity check. For this network we expect bad results. However good results for this network would indicate that height information retrieved was already available in the NWP-data.

3.3.2 Training

- 15 For the training 1.5 million pixels were used, with the distribution ~~50%~~ % low clouds, ~~25%~~ % medium level clouds and ~~25%~~ % high clouds. A higher percentage of low clouds was included because the mean square error (MSE) is often much higher for high clouds. Previous tests showed that less low clouds caused the network to focus too much on predicting the high clouds correctly and showed degraded results for low clouds. For [training-validation-the-validation-dataset-used-during-training](#) 375000 pixels were randomly selected with the same low/medium/high distribution as for the training data.
- 20 The machine learning module Scikit-learn (Pedregosa et al., 2011), the Keras package (Chollet et al., 2015), the Theano (Theano Development Team, 2016) backend and the language Python were used for training the network.

3.3.3 Parameters and configurations

During training of the network the MSE was used as the loss function that is minimized during training. The data were standardized by subtracting the mean and dividing with the standard deviation before training.

- 25 Choosing the number of hidden neurons and hidden layers ~~for~~ [of](#) the neural network is also important for the training to be effective. Too few hidden neurons will result in under-fitting. We used two hidden layers with 30 neurons in the first layer and 15 neurons in the second.

The initialization of weights before training the network is important for the neural network to learn faster. There are many different weight initialization methods, for training the networks the glorot uniform weight initialization was used.

- 30 The activation function used for the hidden layers was the tangent hyperbolic (see Karlik and Olgac, 2011) and for the output layer a linear activation function was used.

To determine the changes in the weights an optimization method is used during the back-propagation algorithm. The optimization method used for the multilayer perceptron is mini-batch stochastic gradient descent which performs mini-batch

training. A mini-batch is a sample of observations in the data. Several observations are used to update weights and biases, which is different from the traditional stochastic gradient descent where one observation at a time is used for the updates (Cotter et al., 2011). Having an optimal mini-batch size is important for the training of a neural network because overly large batches can cause the network to take a long time to converge. We used a mini-batch size of 250.

5 When training the neural network there are different learning parameters that need to be tuned to ensure an effective training procedure. During prototyping several different combinations were tested. The learning rate is a parameter that determines the size of change in the weights. A too large learning rate will result in large weight changes and can result in an unstable model (Hu and Weng, 2009). If a learning rate on the other hand is too small the training time of the network will be long. We used a learning rate of 0.01.

10 The momentum is a parameter which adds a part of the weight change to the current weight change, using momentum can help avoid the network getting trapped in local minima (Gardner and Dorling, 1998). A high value of momentum speeds up the training of the network. We had a momentum of 0.9. The parameter *learning rate decay*, set to 10^{-6} , in Keras, is used to decrease the learning rate after each update as the training progresses.

To avoid the neural network from over-fitting (which makes the network extra sensitive to unseen data), a method called
15 *early stopping* was used. In early stopping the validation error is monitored during training to prevent the network from over-fitting. If the validation error is not improved for some (we used 10) epochs training is stopped; this helps to reduce risk of over-fitting. The network for which the validation error was at its lowest is then used. The neural networks were trained for a maximum of 2650 epochs, but the early stopping method caused the training to stop much earlier.

4 Results and Discussion

20 ~~First~~The validation data was matched with CALIOP layer top pressure and layer top altitude or CPR (CloudSat) height using nearest neighbour matching in the same way as the training data was matched. The CPR (CloudSat) data includes less clouds as both some very low clouds and some very thin clouds are not detected by the radar. CPR (CloudSat) is included to strengthen the results. There is always a risk that a neural network approach learns also or only the errors of the training truth; however if results are improved also when validated with an independent truth it is made sure that it is not only the errors that are learnt.
25 A cloudy threshold of 30% is used for CPR (CloudSat) to include only strong detections. The coarser vertical resolution for CPR (CloudSat) of 500m means that MAE is expected to be higher than 250m compared to 15m for CALIOP.

The scatter plots in Figure 4 show how the cloud top pressure retrievals of the neural networks and the reference methods are distributed compared to CALIOP. Figure 3 show the same type of scatter plots for cloud top height with CloudSat as truth. These scatter plots show that all neural networks have similar appearance with most of the data retrieved close to the truth. All
30 methods (NN-CTTH, PPS-v2014 and MODIS-C6) retrieve some heights and pressures that are very far from the true values of CPR (CloudSat) or CALIOP. It is important to remember that some of these seemingly bad results are due to the different FOV for the MODIS and the CALIOP or CPR (CloudSat) sensors.

Figure 7 compares the NN-AVHRR and PPS-v2014 for one scene. The blue squares for PPS-v2014 (c) are due to the temperature retrieval for 32x32 pixels in one go. We can see that a lot of high clouds are by NN-AVHRR placed higher (pixels that are blue in (c), are white in (a)). For NN-AVHRR in (a) we can see that the large area with low clouds in the lower left corner gets a consistent cloud top height (the same orange colour everywhere). Note that the NN-AVHRR has a less noisy appearance and has less nodata.

4.1 Validation with CALIOP top layer pressure

First we consider the performance of all the trained networks ~~were validated with CALIOP~~ validated with the uppermost CALIOP top layer pressure in terms of mean absolute error (MAE). Results in Table 5 show that both PPS-v2014 and MODIS-C6 have a MAE close to ~~120hPa. Even the 120~~ hPa. Notice that the network using only the NWP information and no imager channels (NN-NWP) shows high MAE. This was included as a sanity check to see that the neural networks are using mainly the satellite data, and the high MAE for NN-NWP is supporting this. The NN-OPAQUE network using only B_{12} and the basic NWP-data has a ~~109~~ hPa improvement in MAE compared to the reference algorithms. By including the variable $B_{11} - B_{12}$, the MAE improves by an additional 19 hPa because $B_{11} - B_{12}$ contains information about the semi-transparency of the pixel. Adding the NWP variable ~~$C_{iww}C_{iww}$~~ , which allows the network to attempt to predict the expected values of $B_{11} - B_{12}$, has a smaller effect of 2 hPa on MAE. However adding all variables containing information on neighboring pixels improves the result by additional 20 hPa. The NN-AVHRR network using 11 μm and 12 μm from MODIS provides an MAE which is reduced by about ~~50hPa~~ 50 hPa compared to both from MODIS-C6 and PPS-v2014. Notice also that the scores improve for all categories (low, medium and high) when compared with both PPS-v2014 and MODIS-C6. The inclusion of the neighbouring pixels gives almost 40 % of the improvement. Note that for medium level clouds NN-BASIC-CIWW, without information from neighbouring pixels, has higher MAE compared to PPS-v2014.

Adding more IR channels improves the results further. Adding channel 8.5 μm ($B_{8.5} - B_{11}$, NN-VIIRS) improves MAE by 7 hPa and adding 7.3 μm ($B_{7.3}$, NN-MERSI-2) improves MAE by 5 hPa. Including the other water vapor channel at 6.7 μm ($B_{6.7}$, ~~NN-MetImage-NoCO₂~~ NN-MetImage-NoCO₂) improves MAE only by 1 hPa. The ~~CO₂~~ CO₂ channel at 13.3 μm ($B_{13.3}$, NN-MetImage) improves the MAE by an additional 6 hPa.

The NN-AVHRR1 network trained using 3.7 μm and 11 μm (MAE 76.1 hPa) is a little worse compared to NN-AVHRR (MAE ~~72.2~~ 72.4 hPa). Note that $B_{3.7}$ has a solar component which currently is not treated in any way. If $B_{3.7}$ was corrected for the solar component, by the network or in a preparation step, the results for AVHRR1 might improve. Also NN-AVHRR1 shows better scores for all categories (low, medium, and high) compared to PPS-v2014 and MODIS-C6. ~~Notice that the network using only the NWP information and no imager channels (NN-NWP) shows high MAE. This was included as a sanity check to see that the predicted height is using mainly the satellite data, and the high MAE for NN-NWP is supporting this.~~

The training with CALIOP using only MODIS from Aqua includes only near NADIR observations with all satellite zenith angles for MODIS below 20°. Figure 1 shows that NN-AVHRR and NN-AVHRR1 networks perform robustly also for higher satellite zenith angles. The ~~N-VIIRS and NN-MetImage-NoCO₂~~ NN-VIIRS and NN-MetImage-NoCO₂ results deviate for satellite zenith angles larger than 60 degrees. The NN-MERSI-2 results deviate for satellite zenith angles larger than 40 degrees.

The NN-MetImage retrieval shows deviations already above 20 degree satellite zenith angles and for satellite zenith angles larger than 40 the retrieval has no predictive skill. Notice that the distribution for MODIS-C6 also depend on the satellite zenith angle (with less high clouds at higher angles). For PPS-v2014 instead ~~there are~~ less low clouds at higher satellite zenith angles ~~are found~~. The neural networks (NN-AVHRR, NN-AVHRR1, NN-VIIRS and NN-MetImage-NoCO₂) can reproduce the bi-modal cloud top pressure distribution similar to CALIOP, PPS-v2014 ~~also has the least reasonable pressure distribution deviates from this shape~~ with one peak for mid-level clouds ~~instead of capturing the two peaks for low and high clouds observed in~~.

4.2 Discussion of statistics measures for non-Gaussian error distributions

For pressure we choose a single measure, MAE, to describe the error; however which (and how many) measures are needed to adequately describe the error distribution can be discussed. For a Gaussian error distribution the obvious choices are bias and SD (standard deviation) as the Gaussian error distribution is completely determined from bias and SD and all other interesting measures could be derived from bias and SD. Unfortunately the error distributions considered here are non-Gaussian. This is expected, as we know that apart from the errors of the ~~CALIOP data~~ algorithm and the errors due to different FOV we expect the lidar to detect some thin cloud layers not visible to the imager. These thin layers, not detected by the imager, should result in underestimated cloud top heights. In Figure 8 the error distributions for MODIS-C6, PPS-v2014 and NN-AVHRR are shown. The Gaussian error distribution with the same bias and SD are plotted in grey. It is clear that the bias is not at the center (the peak) of the distribution. The median is not at the center either, but closer to it. For validation with CALIOP we can see the expected negative bias for all algorithms and for all cases we can see that assuming a Gaussian distribution underestimates the amount of small errors.

$$PE_x = \frac{\text{number of absolut errors} > x \text{ km}}{\text{number of errors}} \quad (1)$$

Results ~~for MAE in meters~~ compared to CALIOP top layer height and CPR (CloudSat) height are provided for the best performing networks in Table 6. ~~Notice that the MAE for high clouds is 8~~ (i.e. NN-OPAQUE, NN-BASIC and NN-BASIC-CI WV was excluded). The skewness show that the distributions are skewed and non-Gaussian. The mode is calculated using the half-range method to robustly estimate the mode from the sample (for more info see Bickel, 2002). The bias should be interpreted with caution. Consider PPS-v2014 compared to CALIOP Table 8 if we add 1465 to all retrievals creating a “corrected” retrieval we would have an error distribution with the same SD and zero bias but the center (peak) of the distribution would not be closer zero. The PE₁ (percentage of absolute errors above 1 km, see Equation 1) for this “corrected” retrieval would increase from 54 % to 73 %! For the user this is clearly not an improvement. The general over estimation of cloud top heights of this “corrected” retrieval would however be detected by the median and the mode which would be further away from zero but now on the positive side. This example illustrates the risk of misinterpretation of the bias for non-Gaussian error distributions.

Several different measures of variation are presented in Table 8 MAE, IQR (Interquartile range), SD and RMSE. The measures have different benefits; IQR are robust against outliers and RMSE and SD focuses on the worst retrievals as errors are squared. Considering that it is likely not interesting if useless retrievals with large errors are 10km off or 15km off, in

combination with that some large errors are expected due to different FOV and different sensitivities of the instruments, the MAE and IQR provide more interesting measures of variation compared to SD and RMSE. In the example discussed in the previous section the MAE for the “corrected” retrieval would change only 10 m but the RMSE (Root mean square error) would improve with 356 m indicating a much better algorithm; when in fact it is a degraded algorithm. If the largest errors are considered very important RMSE is preferred over SD for skewed distributions, especially if bias is also presented; as RMSE and bias have a smaller risk to be misinterpreted by the reader as a Gaussian error distribution.

For low level clouds we have even stronger reasons to expect skewed distributions as there is always a limit (ground) to how low clouds top heights can be underestimated and Table 9 shows that the skewness is large for low level clouds. The bias for low level clouds is difficult to interpret as it is the combination of the main part of the error distribution located close to zero and the large positive errors (which are to some extent expected due to different FOV). In Figure 2 (f) and (e) the error distributions for MODIS-C6, PPS-v2014 and NN-AVHRR for low level clouds are shown. We can see, in Figure 2, that the NN-AVHRR less often underestimates the cloud top height for low level clouds which partly explains the higher bias for NN-AVHRR.

To exemplify the problem with bias and SD for skewed distributions consider PPS-v2014 and NN-AVHRR validated with CPR (CloudSat) in Table 9 and for the argument let us falsely assume a Gaussian error distribution. With this assumption the PPS-v2014 with a 232 m better bias and only 24 m worse SD clearly is the better algorithm. The PE_2 and RMSE are very similar between the two algorithms. However all other measures MAE, IQR, $PE_{0.25}$, $PE_{0.5}$, PE_1 , median and mode all indicate that NN-AVHRR is the better algorithm and it is clear in Figure 2 (e) that the NN-AVHRR has the highest and best centered distribution; contrary to what was indicated by the bias and SD given a false assumption of Gaussian error distribution.

One explanation of the low bias for PPS-v2014 validated with CPR (CloudSat) in Table 9 is seen in Figure 2 (e) where the error distribution of PPS-v2014 is shown to be bi-modal; the general small underestimation of cloud top heights compensates for the mode located close to 1.8km. The low bias can also be explained by less low level clouds predicted much too high. The lowest values for PE_2 , SD and RMSE supports this. If we look at the result for the high clouds (Table 11) we see a large negative tendency for PPS-v2014 (mode and median) and this is also part of the explanation of the small RMSE for PPS-v2014 for low level clouds. If high clouds are generally placed 1.5 km better-for-the too low; this should improve results for low level clouds mistaken for high. This includes cases where the different FOVs causes the imager to see mostly a high cloud but the lidar and radar see only the part of the FOV with a low cloud. This has a large impact on SD and RMSE as the errors are squared.

Comparing the RMSE, SD for NN-AVHRR ~~than~~ and PPS-v2014 for low level clouds in the validation with CPR (CloudSat) also highlights why the RMSE and SD are less useful as measures of variation of the error distribution. The RMSE and SD are very similar between the two algorithms and do not reflect the narrower and better centered error distribution seen for NN-AVHRR for low level clouds in Figure 2 (e). The NN-AVHRR has a larger amount of small errors (see $PE_{0.25}$, $PE_{0.5}$) and only 16 % of the errors are larger than 1 km compared to 29 % for PPS-v2014. ~~Compared to MODIS-C6 the NN-AVHRR is 0.8 better for both medium level clouds and high clouds~~ But NN-AVHRR has 1 % more absolute errors larger than 2 km and the absolute error for this percent is larger. As the MAE does not square the errors, it indicates instead that the NN-AVHRR has

smaller variation of the error distribution. The IQR that does not regard the largest errors at all is more than 500 m better for NN-AVHRR. The

The bias of -117 m for NN-AVHRR compared to -1203 m for MODIS-C6 in Table 11 in the validation with CPR (CloudSat) for a Gaussian error distribution would be a large improvement of tendency; however when also considering the mode and the median we can see that the improvement of the tendency is more realistically between 150 to 500 m compared to CPR (CloudSat) and not as large as indicated by the bias.

4.3 Validation results with CALIOP and CPR (CloudSat) height

All measures in Table 8 have better values for all neural networks compared to both the reference algorithms and both validation truths. Considering the improvement in all the other measures in Table 8 it is safe to conclude that also the lower bias for the neural networks actually is an improvement. However the mode and median better describe the improvement of tendency and for the mode the worst performing network is just a few meters better than the best mode of the reference algorithms. For the comparison to CALIOP in Table 8 we see that the most measures improve as we add more channels to the neural network. Validated with CPR (CloudSat) the results are not improving for the NN-MetImage-NoCO₂ and NN-MetImage. A possible explanation for this can be that some high thin clouds layers are not detected by the radar but the neural network places them higher than the detected CPR (CloudSat) layer below. Thin single layer clouds not detected by the radar are of course not included in the analysis.

In the validation with CALIOP the NN-AVHRR MAE is 627623 m lower (corresponding to 33% 32 % reduction of MAE) than MODIS-C6 and 797795 m (corresponding to 38% % reduction of MAE) lower than PPS-v2014. The NN-MetImage-NoCO₂ NN-MetImage-NoCO₂ has the best result while performing well at all satellite zenith angles, with a 44% 43 % reduction in MAE when compared to MODIS-C6 and a 48% % reduction when compared to PPS-v2014.

Co-located comparisons with CloudSat were also performed. In Table 7 the MAE in meters compared to height from The NN-MetImage have even better scores but are not useful for satellite zenith angles exceeding 20°. In the validation with CPR (CloudSat) are presented. All neural networks show at least 400 better MAE compared to PPS-v2014 and MODIS-C6. Note that when using CPR (CloudSat) as reference, the neural networks show better results for all categories (low, medium and high) compared to both PPS-v2014 and MODIS-C6. The the NN-AVHRR shows 433430 m lower MAE (corresponding corresponding to 25% % reduction of MAE) compared to MODIS-C6 and 483482 m (corresponding to 27% 28 % reduction of MAE) compared to PPS-v2014. The NN-MetImage-NoCO₂ shows 550 lower MAE (NN-MetImage-NoCO₂ shows corresponding to 32% % reduction of MAE) compared to MODIS-C6 and 600 lower MAE (corresponding to 34% % reduction of MAE) compared to PPS-v2014.

Notice that

4.4 Validation results separated for low, medium and high level clouds

Results for low level clouds (Table 9) show that all distributions are well centered around zero and the median and mode are within 250 m from zero for all algorithms except the mode for PPS-v2014 and NN-MetImageNoCO₂ validated with CPR

(CloudSat). The $PE_{0.25}$, $PE_{0.5}$ and PE_1 and most useful measures of variation, IQR and MAE, show better values for the neural networks than both reference algorithms as compared to both validation truths. This indicates that the neural networks have a larger amount of good retrievals with small errors. When validation with CALIOP, only 31 % of the absolute errors for NN-AVHRR exceed 0.5 km, compared to 58 % for MODIS-C6 the MAE for low clouds is high, 1206, in Table 7, and when compared to CALIOP (Table 6) MODIS-C6 and 47 % for PPS-v2014.

For low level clouds validation with CPR (CloudSat) one needs to keep in mind that some thin cloud layers are not detected by the radar. This means that the CPR (CloudSat) height does not reflect the true upper most layer for these clouds. Correct cloud top height retrievals for these clouds will give large positive errors in the CPR (CloudSat) validation for low level clouds. This can explain why the PE_2 and RMSE for all the neural networks are better than both reference algorithms when validated with CALIOP but when validated with CPR (CloudSat) PPS-v2014 have the best PE_2 and RMSE. In Section 4.2 it is discussed why the bias and SD are not very informative for these highly skewed distributions.

Notice that MODIS-C6 has a high MAE (1192 m) for low level clouds when validated with CPR (CloudSat). Also in the CALIOP validation MODIS-C6 has the highest MAE for low-, IQR, RMSE, $PE_{0.25}$, $PE_{0.5}$, PE_1 and PE_2 for low level clouds. When checking the MAE per month we found that scores for MODIS-C6 for low clouds were worst for December (at the same time the scores for high clouds were best in December). There turned out to be a bug in the algorithm for low marine cloud top height (Richard Frey, MODIS Team, 2017 pers. comm.) which likely affected the results and the bug has been corrected in Version 6.1. However overall validation scores for MODIS-C6 were not affected by the bug (Steve Ackerman, MODIS Team, 2017 pers. comm.).

In Figure 2 the height bias distributions comparing MODIS-C6, PPS-v2014 and NN-AVHRR to For medium level clouds (see Table 10) the neural networks have better measures for MAE, IQR, RMSE, SD, PE_1 , PE_2 compared to both reference algorithms when validated both with CALIOP and CPR (CloudSat) are shown. The NN-AVHRR has the highest and best centered distributions especially for low clouds. For medium level clouds both the NN-AVHRR and the For the validation with CPR (Cloudsat) the neural network also have the best $PE_{0.25}$, $PE_{0.5}$, median and bias. In the validation with CALIOP we can see that also PPS-v2014 show good results, has good values for $PE_{0.25}$, $PE_{0.5}$, median and the bias even better than some of the neural networks. This is also seen in Figure 2 (d) where we can see that PPS-v2014 has a well centered and high peak for the error distribution, but a larger amount of underestimated cloud top heights compared to NN-AVHRR. All algorithms report good values for the mode within 300 m from zero for medium level clouds.

For high clouds, in Figure 2, we can see that the NN-AVHRR have less clouds predicted too low, especially compared to PPS-v2014. In the validation with CALIOP (Table 11) the neural networks perform better than the two reference algorithms. For the high clouds comparison to CloudSat, validation with CPR (CloudSat) MODIS-C6 has the highest peak, however MODIS-C6 also has a large amount of clouds with bias (Figure 2), but also a bi-modal error distribution with another peak close to -6 km, which. This explains why the overall MAE (Table 7) for high clouds is better for the NN-AVHRR.

The scatter plots in Figure 3 show how the height data of the neural networks and the reference methods are distributed compared to CloudSat. Figure 4 show the same type of scatter plots for pressure with CALIOP as truth. These scatter plots show that the NN-CTTH all have similar appearance with most of the data retrieved close to the truth. All methods (NN-CTTH,

PPS-v2014 and The higher peak for MODIS-C6) retrieve some heights and pressures that are very far from the true values of CloudSat or CALIOP for validation with CPR (CloudSat) is also reflected in a good IQR, $PE_{0.5}$, PE_1 and mode inline with the neural networks. It is important to remember that some of these seemingly bad results are due to the different FOV for the MODIS and the CALIOP or CloudSat sensors

5 The median and mode for high level clouds for most neural networks are positive when compared to CPR (CloudSat) but negative when validated with CALIOP. This supports the idea that some high thin clouds, or upper part of clouds, are not detected by the radar but by the lidar and the imager. The median for the neural networks for high level clouds are increasing for neural networks with more variables. This suggests that the extra channels help the neural networks to detect the very thin clouds detected by CALIOP. The medians for the validation with CPR (CloudSat) are also increasing, becoming more positive,
10 and this can be explained by some very thin cloud layers not detected by CPR (CloudSat).

In Table 11 we can also note that SD for the PPS-v2014 validated with CPR (CloudSat) is in line with SD for the neural networks. This in combination with the large negative values on mode and median, and the high MAE and quite good IQR indicates that PPS-v2014 systematically underestimates the cloud top height for high-level clouds.

4.5 Validation with CALIOP separated for different cloudtypes

15 In Table 12, the MAE, median and $PE_{0.5}$ are shown for the different cloud types from the CALIOP feature classification flag. We can see that the MAE and $PE_{0.5}$ for all the neural networks is better than both reference algorithms, except that PPS-v2014 also has a low MAE and $PE_{0.5}$ for *opaque altostratus*. Large improvements in MAE are seen for the *altocumulus transparent transparent cirrus* and *deep convective (opaque)* classes. For $PE_{0.5}$ the largest improvements is seen for the four low cloud classes and the *deep convective (opaque)* class for which the neural networks have at least 12 % less errors above 0.5 km
20 compared to both reference algorithms.

All algorithms have medians closer to zero than 250 m for the classes *low overcast (transparent)* and *transition stratocumulus*. For the *low overcast (opaque)* and *low broken cumulus* the neural networks and PPS-v2014 show good values for the median. For the classes *altocumulus transparent*, *transparent cirrus* and *deep convective (opaque)* clouds the neural network show medians at least 450 m closer to zero than both reference algorithms. For the *opaque altostratus* class the median of the
25 reference algorithms is better than the neural networks. PPS-v2014 also have a MAE and $PE_{0.5}$ that is better than NN-AVHRR and NN-AVHRR1 for the *opaque altostratus* class. The good performance of PPS-v2014 for *opaque altostratus* are also reflected in Figure 2 (d) where PPS-v2014 have the highest peak.

It is most difficult for all algorithms to correctly retrieve cloud top height for the largest class *cirrus (transparent)*. If we compare NN-MetImage with PPS-v2014 for the *cirrus (transparent)* class we see that MAE is improved with 2.4 km, the
30 median with 3 km and 21 % less absolute errors are larger than 500 m.

4.6 Geographical aspects of the NN-CTTH performance

To show how performance varies between surfaces and different parts of the globe, the MAE in meters compared to CALIOP are calculated on a Fibonacci grid (constructed using the method described in González, 2009) with a grid evenly spread out

on the globe approximately 250 km apart and All observations are matched to the closest grid point and results are plotted in Figure 5. We can see that all algorithms have problems with clouds around the equator in areas where very thin high cirrus is common. The MAE-difference (Figure 6) shows that the NN-AVHRR is better than MODIS-C6 in most parts of the globe, with the greatest benefit observed closer to the poles. At a few isolated locations MODIS-C6 is better than NN-AVHRR. ~~Figure 7 compares the NN-AVHRR and PPS-v2014 for one scene. Note that the NN-AVHRR has a less noisy appearance and has less nodata.~~

5 Discussion

~~The neural networks show great potential for retrieving cloud height. The NN-CTTH is better in terms of MAE than both PPS-v2014 and the MODIS Collection 6. The neural network algorithms are also useful for instruments with fewer channels than MODIS, including the channels available for AVHRR1. This is important for climate data records which include AVHRR1 data to produce a long, continuous time series.~~

4.1 Future work and challenges

~~Neural networks can behave unexpectedly for unseen data. By using a large training dataset and early stopping the risk for unexpected behaviour is decreased. Also the risk for unexpected results in a neural network algorithm can be a fair price to pay given the significant improvements when compared to the current algorithms.~~

Only near nadir satellite zenith angles were used for training. This might limit the performance for the neural networks at other satellite zenith angles. The NN-MetImage network using the CO_2 channel at $13.3\ \mu\text{m}$ shows strong satellite zenith angle dependence and is not useful for higher satellite zenith angles. A solution to train networks to perform better at higher satellite zenith angles could be to include MODIS data from satellite Terra co-located with CALIPSO in the training data, as they will get matches at any satellite zenith angle ~~although only at high latitudes. As latitude is not used as a variable, data for higher satellite zenith angles included for high latitude regions could help also in other regions. However it might be that the high latitude matches will not help the network if the variety of weather situations and cloud top heights at high latitudes is too small.~~ Radiative transfer calculations for the CO_2 -channels for different satellite-zenith angles could be another way to improve the performance for higher satellite-zenith angles.

Several technical parameters influence the performance of the neural network, for example: learning rate, learning rate decay, momentum, number of layers, number of neurons, weight initialization function and early stopping criteria. For several combinations tested, the differences were in the order of a few hPa. Networks tested using two hidden layers were found to perform better than those using only one hidden layer. We did train one network with less neurons and one with more layers and neurons with the same variables as NN-AVHRR. The network with fewer neurons in the two hidden layers (20/15) was 1 hPa worse. The network with more neurons in three layers (30/45/45) was 2.5hPa better than NN-AVHRR but also took 5 times as long time to retrieve pressure. The best technical parameters and network setup to use could be further investigated.

The NN-CTTH algorithm currently has no pixel specific error estimate. The MAE provides a constant error estimate (the same for all pixels). However for some clouds the height retrieval is more difficult, e.g. thin clouds and sub-pixel clouds. Further work to include pixel specific error estimates could be valuable.

5 Neural networks can behave unexpectedly for unseen data. By using a large training dataset and early stopping the risk for unexpected behaviour is decreased. Also the risk for unexpected results in a neural network algorithm can be a fair price to pay given the significant improvements when compared to the current algorithms. The training of neural networks requires reference data (truth). For optimal performance a neural network approach for upcoming new sensors (e.g. MERSI-2, MetImage) being launched, when data from CALIPSO or CloudSat are no longer available, either another truth is needed or a method to robustly transform network trained for one sensor to other sensors is needed. A way forward could be to include variables
10 with radiative transfer calculations of cloud free brightness temperatures and brightness temperature differences. Further work is needed to test how the networks trained for the MODIS sensor perform for AVHRR, AVHRR1, VIIRS and other sensors. Our results show that networks can be trained using only the channels available on AVHRR, but they might need to be re-trained with actual AVHRR data as the spectral response functions of the channels differ. The spectral response functions also differ between different AVHRR instruments, and more investigations are needed to see how networks trained for one AVHRR
15 instrument will perform for other AVHRR instruments.

The results here are valid for the MODIS imager on the polar orbiting satellite Aqua. However nothing in the method restricts it to polar orbiting satellites. The method should be applicable for imagers like SEVIRI, which has the two most important channels at 11 μm and 12 μm , on geostationary satellites. However the network trained on MODIS data might need to be retrained with SEVIRI data to get the best performance as the spectral response functions between SEVIRI and MODIS differ.
20

5 Conclusions

The neural network approach shows high potential to improve cloud height retrievals. ~~Including variables with information on neighbouring pixel values was very important to get good results. Compared to two existing algorithms (MODIS Collection 6 and PPS-v2014)~~
25 The NN-CTTH (for all trained neural networks) is better in terms of MAE in meters than both PPS-v2014 and the MODIS Collection 6. This is seen for validation with CALIOP and CPR (CloudSat) and for low, medium, high level clouds. The neural networks also show best MAE for all cloud types except *altostratus* (*opaque*) for which PPS-v2014 is better than some of the neural networks~~show an~~. The neural networks show an overall improvement of mean absolute error (MAE) of at least from 400 m. This is valid both for validation with CloudSat and CALIOP height products. Low and up to 1km. Considering overall performance in terms of IQR, RMSE, SD, PE_{0.25}, PE_{0.5}, PE₁, PE₂, median, mode and bias the neural
30 network performs better than both the reference algorithms both when validated with CALIOP and CPR (CloudSat). In the validation with CALIOP the neural networks have between 7 and 20 percentages more retrievals with absolute errors smaller than 250 m compared to the reference algorithms. Considering low, medium and high level clouds all show better results for the NN-CTTH compared with both PPS-v2014 and MODIS-C6. levels separately the neural networks perform better or for

some cases in line with the best of the two reference algorithms in terms of MAE, IQR, $PE_{0.25}$, $PE_{0.5}$, PE_1 , median and mode. This indicates that the neural networks have well centered, narrow error distributions with large amount of retrievals with small errors.

5 The two reference algorithms have been shown to have different strengths MODIS-C6 validated with CPR (CloudSat) for high clouds shows a well centered and narrow error distribution in line with (and better than some of) the neural networks, although the MAE is higher for MODIS-C6. PPS-v2014 validated with CALIOP for the cloud type *altostratus (opaque)* show scores in line with (and better than some of) the neural networks.

10 The error distributions for the cloud top height retrievals were found to be skewed for all algorithms considered in the paper, especially for low level clouds. It was exemplified why the bias and SD should be interpreted with caution and how they can easily be misinterpreted. The median and mode were found to be better measures of tendency than the bias. The IQR and MAE were found to better describe the spread of the errors, compared to SD and RMSE, as the absolute values of the largest errors are not the most interesting. Measuring the amount of absolute error above for example 1km (PE_1) was found to provide valuable information on the amount of large/small errors and useful retrievals.

15 The neural network algorithms are also useful for instruments with fewer channels than MODIS, including the channels available for AVHRR1. This is important for climate data records which include AVHRR1 data to produce a long, continuous time series. Including variables with information on neighbouring pixel values was very important to get good results about and 40 % of the improvement of MAE for the cloud top pressure retrieval for NN-AVHRR was due to the variables with neighbouring pixels. The networks trained using only two IR-channels at 11 μm and 12 μm or 3.7 μm showed the most robust performance at higher satellite zenith angles. Including more IR channels does improve results for nadir observations, but
20 ~~introduces some differences between~~ degrades performance at higher satellite zenith angles. ~~The NN-CTFH could run for the AVHRR1 instrument, in contrast to the other two algorithms neither of which could be applied for the AVHRR1 instrument.~~

A neural network cloud top pressure, temperature and height algorithm will be part of the PPS-v2018 release. The PPS software package is accessible via the [NWCSAF-NWC SAF](http://nwcsaf-nwc-saf.eumetsat.int) site nwc-saf.eumetsat.int.

25 *Author contributions.* All authors contributed to designing the study. Claudia Adok and Nina Håkansson wrote the code and carried out the experiments. Nina Håkansson drafted the manuscript and prepared the figures and tables. All authors discussed results and revised the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors acknowledge that the work was mainly funded by EUMETSAT. The CALIOP data were obtained from the NASA Langley Research Center Atmospheric Science Data Center. The CloudSat Data Processing Center (DPC) and Science Teams

are further acknowledged for providing CloudSat datasets. NWP data were downloaded from ECMWF. The MODIS/Aqua dataset was acquired from the Level-1 & Atmosphere Archive and Distribution System (LAADS) Distributed Active Archive Center (DAAC), located in the Goddard Space Flight Center in Greenbelt, Maryland (<https://ladsweb.nascom.nasa.gov/>). The authors thank Thomas Heinemann (EUMETSAT) for suggesting adding [CloudSat-CPR \(CloudSat\)](#) as an independent validation truth.

References

- Ackerman, S., Menzel, P., and Frey, R.: MODIS Atmosphere L2 Cloud Product (06_L2), https://doi.org/http://dx.doi.org/10.5067/MODIS/MYD06_L2.006, 2015.
- Baum, B. A., Menzel, W. P., Frey, R. A., Tobin, D. C., Holz, R. E., Ackerman, S. A., Heidinger, A. K., and Yang, P.: MODIS Cloud-Top
5 Property Refinements for Collection 6, *Journal of Applied Meteorology and Climatology*, 51, 1145–1163, <https://doi.org/10.1175/JAMC-D-11-0203.1>, <http://dx.doi.org/10.1175/JAMC-D-11-0203.1>, 2012.
- Bickel, D. R.: Robust Estimators of the Mode and Skewness of Continuous Data, *Comput. Stat. Data Anal.*, 39, 153–163, [https://doi.org/10.1016/S0167-9473\(01\)00057-3](https://doi.org/10.1016/S0167-9473(01)00057-3), [http://dx.doi.org/10.1016/S0167-9473\(01\)00057-3](http://dx.doi.org/10.1016/S0167-9473(01)00057-3), 2002.
- Chollet, F. et al.: Keras, <https://github.com/fchollet/keras>, 2015.
- 10 Cotter, A., Shamir, O., Srebro, N., and Sridharan, K.: Better Mini-Batch Algorithms via Accelerated Gradient Methods, in: *Advances in Neural Information Processing Systems 24*, edited by Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., pp. 1647–1655, Curran Associates, Inc., <http://papers.nips.cc/paper/4432-better-mini-batch-algorithms-via-accelerated-gradient-methods.pdf>, 2011.
- Derrien, M., Lavanant, L., and Le Gleau, H.: Retrieval of the cloud top temperature of semi-transparent clouds with AVHRR, in: *Proceedings of the IRS'88*, pp. 199–202, Deepak Publ., Hampton, Lille, France, 1988.
- 15 Dybbroe, A., Karlsson, K.-G., and Thoss, A.: AVHRR cloud detection and analysis using dynamic thresholds and radiative transfer modelling - part one: Algorithm description, *Journal of Applied Meteorology*, 41, 39–54, <https://doi.org/http://dx.doi.org/10.1175/JAM-2188.1>, <http://journals.ametsoc.org/doi/pdf/10.1175/JAM-2188.1>, 2005.
- Gardner, M. and Dorling, S.: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences, *Atmospheric Environment*, 32, 2627 – 2636, [https://doi.org/https://doi.org/10.1016/S1352-2310\(97\)00447-0](https://doi.org/https://doi.org/10.1016/S1352-2310(97)00447-0), <http://www.sciencedirect.com/science/article/pii/S1352231097004470>, 1998.
- 20 González, Á.: Measurement of Areas on a Sphere Using Fibonacci and Latitude–Longitude Lattices, *Mathematical Geosciences*, 42, 49, <https://doi.org/10.1007/s11004-009-9257-x>, <http://dx.doi.org/10.1007/s11004-009-9257-x>, 2009.
- Hamann, U., Walther, A., Baum, B., Bennartz, R., Bugliaro, L., Derrien, M., Francis, P. N., Heidinger, A., Joro, S., Kniffka, A., Le Gléau, H.,
25 Lockhoff, M., Lutz, H.-J., Meirink, J. F., Minnis, P., Palikonda, R., Roebeling, R., Thoss, A., Platnick, S., Watts, P., and Wind, G.: Remote sensing of cloud top pressure/height from SEVIRI: analysis of ten current retrieval algorithms, *Atmospheric Measurement Techniques*, 7, 2839–2867, <https://doi.org/10.5194/amt-7-2839-2014>, <https://www.atmos-meas-tech.net/7/2839/2014/>, 2014.
- Heidinger, A. K., Foster, M. J., Walther, A., and Zhao, X. T.: The Pathfinder Atmospheres–Extended AVHRR Climate Dataset, *Bulletin of the American Meteorological Society*, 95, 909–922, <https://doi.org/10.1175/BAMS-D-12-00246.1>, [https://doi.org/10.1175/](https://doi.org/10.1175/BAMS-D-12-00246.1)
30 [BAMS-D-12-00246.1](https://doi.org/10.1175/BAMS-D-12-00246.1), 2014.
- Hu, X. and Weng, Q.: Estimating impervious surfaces from medium spatial resolution imagery using the self-organizing map and multi-layer perceptron neural networks, *Remote Sensing of Environment*, 113, 2089 – 2102, <https://doi.org/https://doi.org/10.1016/j.rse.2009.05.014>, <http://www.sciencedirect.com/science/article/pii/S0034425709001655>, 2009.
- Inoue, T.: On the Temperature and Effective Emissivity Determination of Semi-Transparent Cirrus Clouds by Bi-Spectral Measurements in the $10\mu\text{m}$ Window Region, *Journal of the Meteorological Society of Japan. Ser. II*, 63, 88–99, 1985.
- 35 Karlik, B. and Olgac, A. V.: Performance analysis of various activation functions in generalized mlp architectures of neural networks, *International Journal of Artificial Intelligence and Expert Systems*, 1, 111–122, 2011.

- Karlsson, K.-G., Anttila, K., Trentmann, J., Stengel, M., Fokke Meirink, J., Devasthale, A., Hanschmann, T., Kothe, S., Jääskeläinen, E., Sedlar, J., Benas, N., van Zadelhoff, G.-J., Schlundt, C., Stein, D., Finkensieper, S., Håkansson, N., and Hollmann, R.: CLARA-A2: the second edition of the CM SAF cloud and radiation data record from 34 years of global AVHRR data, *Atmospheric Chemistry and Physics*, 17, 5809–5828, <https://doi.org/10.5194/acp-17-5809-2017>, <https://www.atmos-chem-phys.net/17/5809/2017/>, 2017.
- 5 Kox, S., Bugliaro, L., and Ostler, A.: Retrieval of cirrus cloud optical thickness and top altitude from geostationary remote sensing, *Atmospheric Measurement Techniques*, 7, 3233–3246, <https://doi.org/10.5194/amt-7-3233-2014>, <https://www.atmos-meas-tech.net/7/3233/2014/>, 2014.
- Marchand, R., Mace, G. G., Ackerman, T., and Stephens, G.: Hydrometeor Detection Using Cloudsat—An Earth-Orbiting 94-GHz Cloud Radar, *Journal of Atmospheric and Oceanic Technology*, 25, 519–533, <https://doi.org/10.1175/2007JTECHA1006.1>, <https://doi.org/10.1175/2007JTECHA1006.1>, 2008.
- 10 Meng, L., He, Y., Chen, J., and Wu, Y.: Neural Network Retrieval of Ocean Surface Parameters from SSM/I Data, *Monthly Weather Review*, 135, 586–597, <https://doi.org/10.1175/MWR3292.1>, <https://doi.org/10.1175/MWR3292.1>, 2007.
- Menzel, W. P., Frey, R. A., Zhang, H., Wylie, D. P., Moeller, C. C., Holz, R. E., Maddux, B., Baum, B. A., Strabala, K. I., and Gumley, L. E.: MODIS Global Cloud-Top Pressure and Amount Estimation: Algorithm Description and Results, *Journal of Applied Meteorology and*
- 15 *Climatology*, 47, 1175–1198, <https://doi.org/10.1175/2007JAMC1705.1>, <http://dx.doi.org/10.1175/2007JAMC1705.1>, 2008.
- Milstein, A. B. and Blackwell, W. J.: Neural network temperature and moisture retrieval algorithm validation for AIRS/AMSU and CrIS/ATMS, *Journal of Geophysical Research: Atmospheres*, 121, 1414–1430, <https://doi.org/10.1002/2015JD024008>, <http://dx.doi.org/10.1002/2015JD024008>, 2015JD024008, 2016.
- Minnis, P., Hong, G., Sun-Mack, S., Smith, W. L., Chen, Y., and Miller, S. D.: Estimating nocturnal opaque ice cloud optical depth from
- 20 MODIS multispectral infrared radiances using a neural network method, *Journal of Geophysical Research: Atmospheres*, 121, 4907–4932, <https://doi.org/10.1002/2015JD024456>, <http://dx.doi.org/10.1002/2015JD024456>, 2015JD024456, 2016.
- MODIS Science Data Support Team: MYD021KM, <https://doi.org/http://dx.doi.org/10.5067/MODIS/MYD021KM.006>, 2015a.
- MODIS Science Data Support Team: MYD03, <https://doi.org/http://dx.doi.org/10.5067/MODIS/MYD03.006>, 2015b.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.,
- 25 Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 12, 2825–2830, 2011.
- Raspaud, M., Dybbroe, A., Devasthale, A., Lahtinen, P., Rasmussen, L. Ø., Hoese, D., Nielsen, E., Leppelt, T., Maul, A., Hamann, U., Kliche, C., Itkin, M., and Thorsteinsson, H.: PyTroll: An open source, community driven Python framework to process Earth Observation satellite data, *Bulletin of the American Meteorological Society*, in press.
- 30 Rossow, W. B. and Schiffer, R. A.: Advances in Understanding Clouds from ISCCP, *Bulletin of the American Meteorological Society*, 80, 2261–2287, [https://doi.org/10.1175/1520-0477\(1999\)080<2261:AIUCFI>2.0.CO;2](https://doi.org/10.1175/1520-0477(1999)080<2261:AIUCFI>2.0.CO;2), [https://doi.org/10.1175/1520-0477\(1999\)080<2261:AIUCFI>2.0.CO;2](https://doi.org/10.1175/1520-0477(1999)080<2261:AIUCFI>2.0.CO;2), 1999.
- SMHI: Algorithm Theoretical Basis Document for Cloud Top Temperature, Pressure and Height of the NWC/PPS, NWCSAF, 4.0 edn., http://www.nwcsaf.org/AemetWebContents/ScientificDocumentation/Documentation/PPS/v2014/NWC-CDOP2-PPS-SMHI-SCI-ATBD-3_v1_0.pdf, 2015.
- 35 Stengel, M., Stapelberg, S., Sus, O., Schlundt, C., Poulsen, C., Thomas, G., Christensen, M., Carbajal Henken, C., Preusker, R., Fischer, J., Devasthale, A., Willén, U., Karlsson, K.-G., McGarragh, G. R., Proud, S., Povey, A. C., Grainger, D. G., Meirink, J. F., Feofilov, A., Bennartz, R., Bojanowski, J., and Hollmann, R.: Cloud property datasets retrieved from AVHRR, MODIS, AATSR and MERIS

in the framework of the Cloud_cci project, Earth System Science Data Discussions, 2017, 1–34, <https://doi.org/10.5194/essd-2017-48>, <https://www.earth-syst-sci-data-discuss.net/essd-2017-48/>, 2017.

5 Stubenrauch, C. J., Rossow, W. B., Kinne, S., Ackerman, S., Cesana, G., Chepfer, H., Girolamo, L. D., Getzewich, B., Guignard, A., Heidinger, A., Maddux, B. C., Menzel, W. P., Minnis, P., Pearl, C., Platnick, S., Poulsen, C., Riedi, J., Sun-Mack, S., Walther, A., Winker, D., Zeng, S., and Zhao, G.: Assessment of Global Cloud Datasets from Satellites: Project and Database Initiated by the GEWEX Radiation Panel, *Bulletin of the American Meteorological Society*, 94, 1031–1049, <https://doi.org/10.1175/BAMS-D-12-00117.1>, <https://doi.org/10.1175/BAMS-D-12-00117.1>, 2013.

Theano Development Team: Theano: A Python framework for fast computation of mathematical expressions, arXiv e-prints, abs/1605.02688, <http://arxiv.org/abs/1605.02688>, 2016.

Table 1. MODIS data from 2010 used for training and validation of the neural networks.

Dataset	Days used
Training	1 st January March July September
	14 th February April May
	14 th August October December
Validation during training	1 st May
	14 th March July November
Testing under development	1 st November
	14 th January June September
Final validation	1 st February April June
	1 st August October December

Table 2. Description of variable types used to train [the](#) neural networks.

Variable type	Variable names	Note
Surface pressure	P_S	Max pressure for pixel
NWP temperatures at surface, 950, 850, 700, 500, 250 hPa	$T_S, T_{950}, T_{850}, T_{700}, T_{500}, T_{250}$	BT to pressure conversion
NWP column integrated water vapour	C_{iww}	Expected BT differences
Brightness Temperature (BT) for 11 μm or 12 μm	B_{11} or B_{12}	Opaque temperature
BT for water vapour channels at 6.7 μm or 7.3 μm	$B_{6.7}, B_{7.3}$	High or low
BT for CO_2 channel at 13.3 μm	$B_{13.3}$	High or low
BT differences $-$	$B_{11} - B_{12}, B_{11} - B_{3.7}, B_{8.5} - B_{11}$	Opacity or phase
BT differences <i>to</i> warmest/coldest neighbour $-$	$B_{12}^W - B_{12}, B_{12}^C - B_{12}$ or $B_{11}^W - B_{11}, B_{11}^C - B_{11}$	Edge or thin
BT differences <i>for</i> warmest/coldest neighbour $-$	$B_{11}^W - B_{12}^W, B_{11}^C - B_{12}^C$ or $B_{11}^W - B_{3.7}^W, B_{11}^C - B_{3.7}^C$	Opacity
Texture: standard deviation of variable for 5x5 pixels	$S_{B_{11}-B_{12}}, S_{B_{11}}, S_{B_{3.7}}$	Edge or thin

Table 3. Description of the different networks. See Table 2 for explanation of the variables. The NWP variables: $P_S, T_S, T_{950}, T_{850}, T_{700}, T_{500}, T_{250}$ are used in all networks.

Network name	Network specific variables
NN-NWP	$Ciww$
NN-OPAQUE	B_{12}
NN-BASIC	$B_{12}, B_{11} - B_{12},$
NN-BASIC-CIWW	$B_{12}, B_{11} - B_{12}, Ciww,$
NN-AVHRR	$B_{12}, B_{11} - B_{12}, Ciww,$ $B_{11}^W - B_{12}^W, B_{11}^C - B_{12}^C,$ $B_{12}^W - B_{12}, B_{12}^C - B_{12},$ $S_{B_{11}-B_{12}}, S_{B_{11}},$
NN-VIIRS	$B_{12}, B_{11} - B_{12}, Ciww,$ $B_{11}^W - B_{12}^W, B_{11}^C - B_{12}^C,$ $B_{12}^W - B_{12}, B_{12}^C - B_{12},$ $S_{B_{11}-B_{12}}, S_{B_{11}},$ $B_{8.5} - B_{11}$
NN-MERSI-2	$B_{12}, B_{11} - B_{12}, Ciww,$ $B_{11}^W - B_{12}^W, B_{11}^C - B_{12}^C,$ $B_{12}^W - B_{12}, B_{12}^C - B_{12},$ $S_{B_{11}-B_{12}}, S_{B_{11}},$ $B_{8.5} - B_{11}, B_{7.3}$
NN-MetImage-NoCO₂ - <u>NN-MetImage-NoCO₂</u>	$B_{12}, B_{11} - B_{12}, Ciww,$ $B_{11}^W - B_{12}^W, B_{11}^C - B_{12}^C,$ $B_{12}^W - B_{12}, B_{12}^C - B_{12},$ $S_{B_{11}-B_{12}}, S_{B_{11}},$ $B_{8.5} - B_{11}, B_{7.3}, B_{6.7}$
NN-MetImage	$B_{12}, B_{11} - B_{12}, Ciww,$ $B_{11}^W - B_{12}^W, B_{11}^C - B_{12}^C,$ $B_{12}^W - B_{12}, B_{12}^C - B_{12},$ $S_{B_{11}-B_{12}}, S_{B_{11}},$ $B_{8.5} - B_{11}, B_{7.3}, B_{6.7}, B_{13.3}$
NN-AVHRR1	$B_{11}, B_{11} - B_{3.7}, Ciww,$ $B_{11}^W - B_{3.7}^W, B_{11}^C - B_{3.7}^C,$ $B_{11}^W - B_{11}, B_{11}^C - B_{11}$ $S_{B_{3.7}}, S_{B_{11}}$

Table 4. Description of the imager channels used for the different networks algorithms. For MODIS-C6 channels used indirectly, to determine if CO₂-slicing should be applied, are noted with brackets.

Network name	Imager channel: B₁₁ - <u>B_{3,7}</u>	B₁₂ - <u>B_{6,7}</u>	<u>B_{7,3}</u>	B _{8,5}	B_{7,3} - <u>B₁₁</u>	B_{6,7} - <u>B₁₂</u>	B _{13,3}	B_{3,7} - <u>B_{13,6}</u>	<u>B_{13,9}</u>
<u>PPS-v2014</u>					<u>x</u>	<u>x</u>			
<u>MODIS-C6</u>			(x)	(x)	<u>x</u>	(x)	<u>x</u>	<u>x</u>	<u>x</u>
NN-NWP									
NN-OPAQUE							x		
NN-BASIC					x	x			
NN-BASIC-CIWV					x	x			
NN-AVHRR					x	x			
NN-VIIRS				x	x	x			
NN-MERSI-2			x	x	x	x			
NN-MetImage-NoCO₂ - <u>NN-MetImage-NoCO₂</u>		x	x	x	x	x			
NN-MetImage		x	x	x	x	x	x		
NN-AVHRR1	x				x				

Table 5. Mean absolute error (MAE) for different algorithms compared to CALIOP top layer pressure. The final validation dataset (see Table 1), containing ~~1796428~~1832432 pixels (45% ~~high~~, 39% ~~low~~ and 16% ~~medium level clouds~~) is used. Pixels with valid pressure for PPS-v2014, MODIS-C6, and CALIOP are considered. The low, medium and high classes are from CALIOP feature classification flag.

	MAE [hPa]			
	all	low	medium	high
PPS-v2014	122.9 <u>122.6</u>	79.4 <u>80.2</u>	88.6 <u>88.0</u>	173.5 <u>172.9</u>
MODIS-C6	124.3 <u>123.9</u>	90.4 <u>90.7</u>	140.0 <u>139.8</u>	148.4 <u>147.3</u>
NN-NWP	191.6 <u>191.7</u>	140.8 <u>141.7</u>	110.5 <u>110.3</u>	266.0 <u>265.8</u>
NN-OPAQUE	113.3 <u>113.2</u>	81.3 <u>82.1</u>	105.1 <u>105.0</u>	144.5 <u>143.8</u>
NN-BASIC	93.9	66.7 <u>67.7</u>	92.8	118.3 <u>117.6</u>
NN-BASIC-CIWV	92.1	66.4 <u>67.5</u>	91.4 <u>91.3</u>	115.0 <u>114.2</u>
NN-AVHRR	72.2 <u>72.4</u>	54.1 <u>55.4</u>	67.4 <u>67.6</u>	89.9 <u>89.2</u>
NN-VIIRS	65.7 <u>65.9</u>	49.1 <u>50.5</u>	59.2 <u>59.3</u>	82.7 <u>81.9</u>
NN-MERSI-2 <u>NN-MERSI2</u>	61.2 <u>61.4</u>	46.7 <u>48.2</u>	52.0 <u>52.1</u>	77.3 <u>76.6</u>
NN-MetImage-NoCO₂ <u>NN-MetImage-NoCO₂</u>	60.0 <u>60.3</u>	45.5 <u>47.1</u>	54.3 <u>54.5</u>	74.8 <u>74.1</u>
NN-MetImage	53.6 <u>54.2</u>	42.7 <u>44.5</u>	51.3 <u>51.6</u>	64.1 <u>63.8</u>
NN-AVHRR1	76.1	53.6 <u>54.7</u>	70.0 <u>69.9</u>	98.1 <u>97.3</u>

Table 6. Mean absolute error (MAE) in meters for different algorithms compared to CALIOP top layer altitude. The final validation dataset (see Table 1), containing 1793142 pixels (45% high, 39% low and 16% medium level clouds), where all algorithms had a cloud height is used. The low, medium and high classes are from CALIOP feature classification flag. A small amount 0.2% of the pixels were excluded because of missing height or pressure below 70hPa for any of the algorithms.

	MAE [m]			
	all	low	medium	high
PPS-v2014	2087	837	1124	3542
MODIS-C6	1917	944	1759	2833
NN-AVHRR	1290	567	962	2049
NN-VIIRS	1177	514	828	1891
NN-MERSI-2	1110	488	727	1797
NN-MetImage- <i>NoCO₂</i>	1081	478	757	1732
NN-Metimage	964	451	707	1510
NN-AVHRR1	1375	560	978	2239

Table 7. Mean absolute error (MAE) in meters for different algorithms compared to CPR (CloudSat) Height. The final validation dataset (see Table 1), containing 1121199 pixels (53% high, 27% low and 21% medium level clouds) is used. The low, medium and high classes are derived comparing the CloudSat height to the NWP height at 440hPa and 680hPa. A cloudy threshold of 30% is used for CloudSat.

	MAE [m]			
	all	low	medium	high
PPS-v2014	1761	977	1365	2315
MODIS-C6	1711	1206	1912	1888
NN-AVHRR	1278	771	1218	1559
NN-VIIRS	1223	766	1144	1486
NN-MERSI-2	1135	748	1061	1362
NN-MetImage- <i>NoCO₂</i>	1161	768	1095	1386
NN-Metimage	1186	802	1119	1407
NN-AVHRR1	1297	858	1225	1548

Table 8. Statistic measures for the error distributions for all clouds. For all measures except skewness it is the case that values closer to zero are better. The statistics are calculated for 1198599 matches for CPR (CloudSat) and 1803335 matches for CALIOP. A small amount 0.2% of the matches were excluded because of missing height or pressure below 70 hPa for any of the algorithms. PE_X describes percentage of absolute errors above X km, see Equation 1.

	<u>MAE</u>	<u>IQR</u>	<u>RMSE</u>	<u>SD</u> ¹	<u>PE</u> _{0.25}	<u>PE</u> _{0.5}	<u>PE</u> ₁	<u>PE</u> ₂	<u>median</u>	<u>mode</u>	<u>bias</u> ¹	<u>skew</u>
	[m]	[m]	[m]	[m]	[%]	[%]	[%]	[%]	[m]	[m]	[m]	
CALIOP all clouds												
<u>PPS-v2014</u>	<u>2095</u>	<u>2832</u>	<u>3188</u>	<u>2832</u>	<u>82</u>	<u>69</u>	<u>54</u>	<u>29</u>	<u>-639</u>	<u>-118</u>	<u>-1465</u>	<u>-1.0</u>
<u>MODIS-C6</u>	<u>1923</u>	<u>2177</u>	<u>3105</u>	<u>2883</u>	<u>85</u>	<u>72</u>	<u>51</u>	<u>23</u>	<u>-612</u>	<u>-262</u>	<u>-1153</u>	<u>-1.5</u>
<u>NN-AVHRR</u>	<u>1300</u>	<u>1326</u>	<u>2234</u>	<u>2197</u>	<u>73</u>	<u>56</u>	<u>36</u>	<u>14</u>	<u>50</u>	<u>106</u>	<u>-405</u>	<u>-1.8</u>
<u>NN-VIIRS</u>	<u>1187</u>	<u>1189</u>	<u>2114</u>	<u>2074</u>	<u>71</u>	<u>52</u>	<u>33</u>	<u>12</u>	<u>28</u>	<u>100</u>	<u>-410</u>	<u>-1.9</u>
<u>NN-MERSI-2</u>	<u>1120</u>	<u>1107</u>	<u>2039</u>	<u>1996</u>	<u>69</u>	<u>50</u>	<u>30</u>	<u>11</u>	<u>-2</u>	<u>73</u>	<u>-420</u>	<u>-2.0</u>
<u>NN-MetImage-NoCO₂</u>	<u>1091</u>	<u>1040</u>	<u>2009</u>	<u>1966</u>	<u>68</u>	<u>48</u>	<u>29</u>	<u>11</u>	<u>-49</u>	<u>44</u>	<u>-416</u>	<u>-2.0</u>
<u>NN-MetImage</u>	<u>979</u>	<u>909</u>	<u>1840</u>	<u>1817</u>	<u>65</u>	<u>46</u>	<u>26</u>	<u>9</u>	<u>-17</u>	<u>15</u>	<u>-294</u>	<u>-1.9</u>
<u>NN-AVHRR1</u>	<u>1383</u>	<u>1547</u>	<u>2354</u>	<u>2281</u>	<u>75</u>	<u>58</u>	<u>38</u>	<u>16</u>	<u>-42</u>	<u>50</u>	<u>-584</u>	<u>-1.8</u>
CPR (CloudSat) all clouds												
<u>PPS-v2014</u>	<u>1744</u>	<u>2255</u>	<u>2432</u>	<u>2160</u>	<u>87</u>	<u>74</u>	<u>56</u>	<u>24</u>	<u>-833</u>	<u>-426</u>	<u>-1118</u>	<u>-0.1</u>
<u>MODIS-C6</u>	<u>1692</u>	<u>1928</u>	<u>2607</u>	<u>2533</u>	<u>84</u>	<u>70</u>	<u>48</u>	<u>20</u>	<u>-375</u>	<u>-259</u>	<u>-614</u>	<u>-0.1</u>
<u>NN-AVHRR</u>	<u>1262</u>	<u>1473</u>	<u>1928</u>	<u>1923</u>	<u>77</u>	<u>61</u>	<u>41</u>	<u>14</u>	<u>88</u>	<u>-141</u>	<u>143</u>	<u>0.2</u>
<u>NN-VIIRS</u>	<u>1207</u>	<u>1368</u>	<u>1901</u>	<u>1896</u>	<u>76</u>	<u>58</u>	<u>38</u>	<u>13</u>	<u>69</u>	<u>-146</u>	<u>137</u>	<u>0.5</u>
<u>NN-MERSI-2</u>	<u>1120</u>	<u>1275</u>	<u>1793</u>	<u>1788</u>	<u>75</u>	<u>56</u>	<u>35</u>	<u>11</u>	<u>40</u>	<u>-201</u>	<u>136</u>	<u>0.5</u>
<u>NN-MetImage-NoCO₂</u>	<u>1146</u>	<u>1315</u>	<u>1834</u>	<u>1828</u>	<u>76</u>	<u>57</u>	<u>35</u>	<u>12</u>	<u>9</u>	<u>-218</u>	<u>147</u>	<u>0.7</u>
<u>NN-MetImage</u>	<u>1170</u>	<u>1421</u>	<u>1865</u>	<u>1843</u>	<u>76</u>	<u>58</u>	<u>37</u>	<u>11</u>	<u>84</u>	<u>-243</u>	<u>285</u>	<u>0.9</u>
<u>NN-AVHRR1</u>	<u>1281</u>	<u>1523</u>	<u>1953</u>	<u>1953</u>	<u>79</u>	<u>63</u>	<u>41</u>	<u>14</u>	<u>30</u>	<u>-128</u>	<u>-14</u>	<u>0.0</u>

¹ Interpret bias and SD with caution as distributions are non-Gaussian. Bias is not located at the center of the distribution.

Table 9. Statistic measures for the error distributions for low level clouds. For all measures except skewness it is the case that values closer to zero are better. The statistics are calculated for 328015 matches for CPR (CloudSat) and 709434 matches for CALIOP. The low class comes from CALIOP feature classification flag (class 0, 1, 2 and 3) and for CPR (CloudSat) it is the pixels with heights lower or exactly at the NWP height at 680 hPa. PE_X describes percentage of absolute errors above X km, see Equation 1.

	<u>MAE</u>	<u>IQR</u>	<u>RMSE</u>	<u>SD</u> ¹	<u>PE_{0.25}</u>	<u>PE_{0.5}</u>	<u>PE₁</u>	<u>PE₂</u>	<u>median</u>	<u>mode</u>	<u>bias</u> ¹	<u>skew</u>
	[m]	[m]	[m]	[m]	[%]	[%]	[%]	[%]	[m]	[m]	[m]	
Low level clouds CALIOP												
<u>PPS-v2014</u>	<u>847</u>	<u>1035</u>	<u>1469</u>	<u>1436</u>	<u>68</u>	<u>47</u>	<u>27</u>	<u>5</u>	<u>-46</u>	<u>-117</u>	<u>312</u>	<u>3.0</u>
<u>MODIS-C6</u>	<u>952</u>	<u>1230</u>	<u>1576</u>	<u>1561</u>	<u>78</u>	<u>58</u>	<u>29</u>	<u>6</u>	<u>-17</u>	<u>-150</u>	<u>219</u>	<u>2.9</u>
<u>NN-AVHRR</u>	<u>586</u>	<u>584</u>	<u>1121</u>	<u>1027</u>	<u>56</u>	<u>31</u>	<u>14</u>	<u>3</u>	<u>215</u>	<u>101</u>	<u>449</u>	<u>4.0</u>
<u>NN-VIIRS</u>	<u>533</u>	<u>515</u>	<u>1080</u>	<u>1006</u>	<u>52</u>	<u>27</u>	<u>11</u>	<u>3</u>	<u>182</u>	<u>126</u>	<u>391</u>	<u>4.8</u>
<u>NN-MERSI-2</u>	<u>509</u>	<u>490</u>	<u>1063</u>	<u>998</u>	<u>49</u>	<u>25</u>	<u>10</u>	<u>3</u>	<u>159</u>	<u>86</u>	<u>365</u>	<u>4.8</u>
<u>NN-MetImage-NoCO₂</u>	<u>499</u>	<u>504</u>	<u>1068</u>	<u>1024</u>	<u>48</u>	<u>24</u>	<u>10</u>	<u>3</u>	<u>98</u>	<u>40</u>	<u>303</u>	<u>4.9</u>
<u>NN-MetImage</u>	<u>476</u>	<u>450</u>	<u>1103</u>	<u>1069</u>	<u>45</u>	<u>21</u>	<u>8</u>	<u>3</u>	<u>74</u>	<u>14</u>	<u>271</u>	<u>5.4</u>
<u>NN-AVHRR1</u>	<u>574</u>	<u>646</u>	<u>1045</u>	<u>969</u>	<u>58</u>	<u>33</u>	<u>13</u>	<u>3</u>	<u>197</u>	<u>49</u>	<u>391</u>	<u>3.8</u>
Low level clouds CPR (CloudSat)												
<u>PPS-v2014</u>	<u>949</u>	<u>1197</u>	<u>1571</u>	<u>1556</u>	<u>78</u>	<u>56</u>	<u>29</u>	<u>5</u>	<u>-173</u>	<u>-413</u>	<u>211</u>	<u>2.8</u>
<u>MODIS-C6</u>	<u>1192</u>	<u>1335</u>	<u>2145</u>	<u>2097</u>	<u>79</u>	<u>60</u>	<u>33</u>	<u>9</u>	<u>46</u>	<u>-110</u>	<u>450</u>	<u>2.9</u>
<u>NN-AVHRR</u>	<u>743</u>	<u>685</u>	<u>1595</u>	<u>1532</u>	<u>56</u>	<u>31</u>	<u>16</u>	<u>6</u>	<u>16</u>	<u>-132</u>	<u>443</u>	<u>3.8</u>
<u>NN-VIIRS</u>	<u>739</u>	<u>637</u>	<u>1690</u>	<u>1633</u>	<u>55</u>	<u>30</u>	<u>15</u>	<u>6</u>	<u>-6</u>	<u>-139</u>	<u>432</u>	<u>4.2</u>
<u>NN-MERSI-2</u>	<u>721</u>	<u>605</u>	<u>1652</u>	<u>1602</u>	<u>55</u>	<u>28</u>	<u>14</u>	<u>6</u>	<u>-31</u>	<u>-181</u>	<u>403</u>	<u>4.1</u>
<u>NN-MetImage-NoCO₂</u>	<u>742</u>	<u>608</u>	<u>1670</u>	<u>1637</u>	<u>60</u>	<u>31</u>	<u>13</u>	<u>6</u>	<u>-105</u>	<u>-255</u>	<u>328</u>	<u>4.2</u>
<u>NN-MetImage</u>	<u>773</u>	<u>578</u>	<u>1813</u>	<u>1775</u>	<u>58</u>	<u>30</u>	<u>13</u>	<u>6</u>	<u>-102</u>	<u>-217</u>	<u>369</u>	<u>4.1</u>
<u>NN-AVHRR1</u>	<u>827</u>	<u>852</u>	<u>1676</u>	<u>1602</u>	<u>64</u>	<u>38</u>	<u>18</u>	<u>7</u>	<u>48</u>	<u>-198</u>	<u>491</u>	<u>3.6</u>

¹ Interpret bias and SD with caution as distributions are non-Gaussian. Bias is not located at the center of the distribution.

Table 10. Statistic measures for the error distributions for medium level clouds. For all measures except skewness it is the case that values closer to zero are better. The statistics are calculated for 244885 matches for CPR (CloudSat) and 295186 matches for CALIOP. The high class comes from CALIOP feature classification flag (class 4 and 5) and for CPR (CloudSat) it is the pixels with heights between the NWP height at 440 hPa and 680 hPa. PE_X describes percentage of absolute errors above X km, see Equation 1.

	<u>MAE</u>	<u>IQR</u>	<u>RMSE</u>	<u>SD¹</u>	<u>PE_{0.25}</u>	<u>PE_{0.5}</u>	<u>PE₁</u>	<u>PE₂</u>	<u>median</u>	<u>mode</u>	<u>bias¹</u>	<u>skew</u>
	[m]	[m]	[m]	[m]	[%]	[%]	[%]	[%]	[m]	[m]	[m]	
Medium level clouds CALIOP												
<u>PPS-v2014</u>	<u>1121</u>	<u>1600</u>	<u>1651</u>	<u>1614</u>	<u>78</u>	<u>59</u>	<u>37</u>	<u>12</u>	<u>-68</u>	<u>124</u>	<u>-348</u>	<u>0.2</u>
<u>MODIS-C6</u>	<u>1759</u>	<u>2590</u>	<u>2304</u>	<u>2192</u>	<u>87</u>	<u>76</u>	<u>60</u>	<u>27</u>	<u>-654</u>	<u>205</u>	<u>-708</u>	<u>0.6</u>
<u>NN-AVHRR</u>	<u>969</u>	<u>1243</u>	<u>1394</u>	<u>1339</u>	<u>78</u>	<u>59</u>	<u>34</u>	<u>7</u>	<u>304</u>	<u>273</u>	<u>387</u>	<u>0.8</u>
<u>NN-VIIRS</u>	<u>832</u>	<u>1048</u>	<u>1227</u>	<u>1206</u>	<u>74</u>	<u>53</u>	<u>28</u>	<u>5</u>	<u>186</u>	<u>23</u>	<u>223</u>	<u>0.7</u>
<u>NN-MERSI-2</u>	<u>731</u>	<u>935</u>	<u>1102</u>	<u>1093</u>	<u>70</u>	<u>47</u>	<u>23</u>	<u>4</u>	<u>83</u>	<u>16</u>	<u>144</u>	<u>0.9</u>
<u>NN-MetImage-NoCO₂</u>	<u>762</u>	<u>984</u>	<u>1148</u>	<u>1145</u>	<u>71</u>	<u>49</u>	<u>24</u>	<u>4</u>	<u>28</u>	<u>-1</u>	<u>86</u>	<u>1.1</u>
<u>NN-MetImage</u>	<u>714</u>	<u>905</u>	<u>1091</u>	<u>1090</u>	<u>69</u>	<u>46</u>	<u>22</u>	<u>3</u>	<u>4</u>	<u>-63</u>	<u>36</u>	<u>1.1</u>
<u>NN-AVHRR1</u>	<u>980</u>	<u>1330</u>	<u>1381</u>	<u>1364</u>	<u>79</u>	<u>61</u>	<u>35</u>	<u>7</u>	<u>187</u>	<u>176</u>	<u>213</u>	<u>0.5</u>
Medium level clouds CPR (CloudSat)												
<u>PPS-v2014</u>	<u>1364</u>	<u>1978</u>	<u>1927</u>	<u>1858</u>	<u>82</u>	<u>66</u>	<u>46</u>	<u>18</u>	<u>-300</u>	<u>53</u>	<u>-512</u>	<u>0.5</u>
<u>MODIS-C6</u>	<u>1909</u>	<u>2698</u>	<u>2532</u>	<u>2475</u>	<u>88</u>	<u>78</u>	<u>62</u>	<u>30</u>	<u>-597</u>	<u>69</u>	<u>-534</u>	<u>0.9</u>
<u>NN-AVHRR</u>	<u>1215</u>	<u>1541</u>	<u>1817</u>	<u>1770</u>	<u>81</u>	<u>64</u>	<u>40</u>	<u>12</u>	<u>209</u>	<u>-113</u>	<u>409</u>	<u>1.2</u>
<u>NN-VIIRS</u>	<u>1139</u>	<u>1325</u>	<u>1788</u>	<u>1760</u>	<u>77</u>	<u>59</u>	<u>36</u>	<u>11</u>	<u>114</u>	<u>-81</u>	<u>310</u>	<u>1.5</u>
<u>NN-MERSI-2</u>	<u>1059</u>	<u>1203</u>	<u>1706</u>	<u>1686</u>	<u>75</u>	<u>55</u>	<u>32</u>	<u>10</u>	<u>15</u>	<u>-150</u>	<u>264</u>	<u>1.7</u>
<u>NN-MetImage-NoCO₂</u>	<u>1091</u>	<u>1259</u>	<u>1752</u>	<u>1740</u>	<u>76</u>	<u>57</u>	<u>33</u>	<u>10</u>	<u>-44</u>	<u>-154</u>	<u>205</u>	<u>1.8</u>
<u>NN-MetImage</u>	<u>1113</u>	<u>1217</u>	<u>1832</u>	<u>1818</u>	<u>75</u>	<u>56</u>	<u>33</u>	<u>11</u>	<u>-45</u>	<u>-174</u>	<u>225</u>	<u>1.9</u>
<u>NN-AVHRR1</u>	<u>1221</u>	<u>1591</u>	<u>1776</u>	<u>1751</u>	<u>81</u>	<u>65</u>	<u>41</u>	<u>13</u>	<u>146</u>	<u>-25</u>	<u>301</u>	<u>1.0</u>

¹ Interpret bias and SD with caution as distributions are non-Gaussian. Bias is not located at the center of the distribution.

Table 11. Statistic measures for the error distributions for high level clouds. For all measures except skewness it is the case that values closer to zero are better. The statistics are calculated for 625699 matches for CPR (CloudSat) and 798715 matches for CALIOP. The high class comes from CALIOP feature classification flag (class 6 and 7) and for CPR (CloudSat) it is the pixels with heights higher or exactly at the NWP height at 440 hPa. PE_X describes percentage of absolute errors above X km, see Equation 1.

	<u>MAE</u> [m]	<u>IQR</u> [m]	<u>RMSE</u> [m]	<u>SD</u> ¹ [m]	<u>PE_{0.25}</u> [%]	<u>PE_{0.5}</u> [%]	<u>PE₁</u> [%]	<u>PE₂</u> [%]	<u>median</u> [m]	<u>mode</u> [m]	<u>bias</u> ¹ [m]	<u>skew</u>
High level clouds CALIOP												
<u>PPS-v2014</u>	<u>3564</u>	<u>3367</u>	<u>4475</u>	<u>2842</u>	<u>96</u>	<u>92</u>	<u>84</u>	<u>57</u>	<u>-2918</u>	<u>-1897</u>	<u>-3456</u>	<u>-0.9</u>
<u>MODIS-C6</u>	<u>2846</u>	<u>3095</u>	<u>4196</u>	<u>3342</u>	<u>92</u>	<u>84</u>	<u>68</u>	<u>36</u>	<u>-1586</u>	<u>-917</u>	<u>-2537</u>	<u>-1.5</u>
<u>NN-AVHRR</u>	<u>2057</u>	<u>2775</u>	<u>3072</u>	<u>2704</u>	<u>87</u>	<u>76</u>	<u>57</u>	<u>27</u>	<u>-799</u>	<u>-130</u>	<u>-1457</u>	<u>-1.4</u>
<u>NN-VIIRS</u>	<u>1899</u>	<u>2459</u>	<u>2916</u>	<u>2581</u>	<u>86</u>	<u>74</u>	<u>53</u>	<u>23</u>	<u>-716</u>	<u>-18</u>	<u>-1356</u>	<u>-1.6</u>
<u>NN-MERSI-2</u>	<u>1807</u>	<u>2258</u>	<u>2818</u>	<u>2486</u>	<u>85</u>	<u>72</u>	<u>51</u>	<u>21</u>	<u>-705</u>	<u>-192</u>	<u>-1326</u>	<u>-1.7</u>
<u>NN-MetImage-NoCO₂</u>	<u>1739</u>	<u>2134</u>	<u>2760</u>	<u>2464</u>	<u>84</u>	<u>70</u>	<u>48</u>	<u>20</u>	<u>-606</u>	<u>-248</u>	<u>-1242</u>	<u>-1.8</u>
<u>NN-MetImage</u>	<u>1524</u>	<u>1906</u>	<u>2476</u>	<u>2298</u>	<u>83</u>	<u>67</u>	<u>44</u>	<u>16</u>	<u>-360</u>	<u>-83</u>	<u>-920</u>	<u>-2.0</u>
<u>NN-AVHRR1</u>	<u>2250</u>	<u>2913</u>	<u>3292</u>	<u>2791</u>	<u>89</u>	<u>79</u>	<u>61</u>	<u>30</u>	<u>-1099</u>	<u>-475</u>	<u>-1746</u>	<u>-1.3</u>
High level clouds CPR (CloudSat)												
<u>PPS-v2014</u>	<u>2309</u>	<u>2384</u>	<u>2930</u>	<u>2092</u>	<u>93</u>	<u>87</u>	<u>74</u>	<u>36</u>	<u>-1789</u>	<u>-1428</u>	<u>-2052</u>	<u>-0.5</u>
<u>MODIS-C6</u>	<u>1869</u>	<u>2142</u>	<u>2845</u>	<u>2578</u>	<u>86</u>	<u>73</u>	<u>51</u>	<u>22</u>	<u>-614</u>	<u>-506</u>	<u>-1203</u>	<u>-1.2</u>
<u>NN-AVHRR</u>	<u>1553</u>	<u>2244</u>	<u>2121</u>	<u>2118</u>	<u>87</u>	<u>75</u>	<u>54</u>	<u>19</u>	<u>143</u>	<u>348</u>	<u>-117</u>	<u>-0.6</u>
<u>NN-VIIRS</u>	<u>1479</u>	<u>2095</u>	<u>2043</u>	<u>2041</u>	<u>86</u>	<u>73</u>	<u>52</u>	<u>17</u>	<u>168</u>	<u>332</u>	<u>-85</u>	<u>-0.7</u>
<u>NN-MERSI-2</u>	<u>1353</u>	<u>1876</u>	<u>1894</u>	<u>1893</u>	<u>85</u>	<u>71</u>	<u>48</u>	<u>14</u>	<u>177</u>	<u>326</u>	<u>-54</u>	<u>-0.9</u>
<u>NN-MetImage-NoCO₂</u>	<u>1379</u>	<u>1843</u>	<u>1944</u>	<u>1944</u>	<u>85</u>	<u>71</u>	<u>48</u>	<u>15</u>	<u>219</u>	<u>292</u>	<u>29</u>	<u>-0.7</u>
<u>NN-MetImage</u>	<u>1399</u>	<u>1871</u>	<u>1904</u>	<u>1885</u>	<u>87</u>	<u>74</u>	<u>52</u>	<u>14</u>	<u>463</u>	<u>511</u>	<u>265</u>	<u>-0.8</u>
<u>NN-AVHRR1</u>	<u>1542</u>	<u>2275</u>	<u>2145</u>	<u>2107</u>	<u>87</u>	<u>74</u>	<u>53</u>	<u>19</u>	<u>-67</u>	<u>281</u>	<u>-403</u>	<u>-0.8</u>

¹ Interpret bias and SD with caution as distributions are non-Gaussian. Bias is not located at the center of the distribution.

Table 12. Mean absolute error (MAE) and median in meters for different algorithms compared to CALIOP top layer altitude. The final validation dataset (see Table 1), containing 1803335 pixels (5 % low overcast (transparent), 12 % low overcast opaque, 19 % transition stratocumulus, 2 % low, broken cumulus, 7 % altocumulus (transparent), 8 % altostratus (opaque), 30 % cirrus (transparent) and 14 % deep convective (opaque)), where all algorithms had a cloud top height is used. The cloud types are from CALIOP feature classification. PE_{0.5} describes percentage of absolute errors above 0.5 km.

	low overcast, transparent	low overcast, opaque	transition stratocumulus	low, broken cumulus	altocumulus (transparent)	altostratus (opaque)	cirrus (transparent)	deep convective (opaque)
MAE [m]								
PPS-v2014	709	637	886	1695	1609	699	4343	1901
MODIS-C6	903	1028	901	1058	2343	1254	3567	1308
NN-AVHRR	519	442	627	1027	1134	825	2608	883
NN-VIIRS	454	407	571	938	1011	678	2398	833
NN-MERSI-2	408	381	550	946	900	584	2283	791
NN-MetImage-NoCO ₂	395	372	541	929	929	617	2210	734
NN-MetImage	365	364	509	912	885	565	1905	711
NN-AVHRR1	516	448	617	911	1156	827	2847	977
median [m]								
PPS-v2014	-183	50	-90	220	-633	63	-3835	-1716
MODIS-C6	-91	331	-138	-477	-1953	85	-2243	-912
NN-AVHRR	223	160	241	380	109	410	-1605	71
NN-VIIRS	185	143	201	315	7	279	-1360	46
NN-MERSI-2	160	116	177	313	-34	145	-1268	-35
NN-MetImage-NoCO ₂	110	70	102	226	-138	119	-1133	-19
NN-MetImage	53	46	86	214	-163	87	-787	125
NN-AVHRR1	188	140	232	313	-180	380	-1895	-82
PE _{0.5} [%]								
PPS-v2014	46	38	49	67	76	44	95	87
MODIS-C6	58	59	56	63	89	64	87	76
NN-AVHRR	33	23	34	47	67	53	83	60
NN-VIIRS	27	19	30	43	63	44	81	58
NN-MERSI-2	24	17	28	42	58	37	79	57
NN-MetImage-NoCO ₂	22	16	27	40	59	40	78	53
NN-MetImage	20	15	23	37	58	36	74	52
NN-AVHRR1	34	24	37	45	68	54	86	64

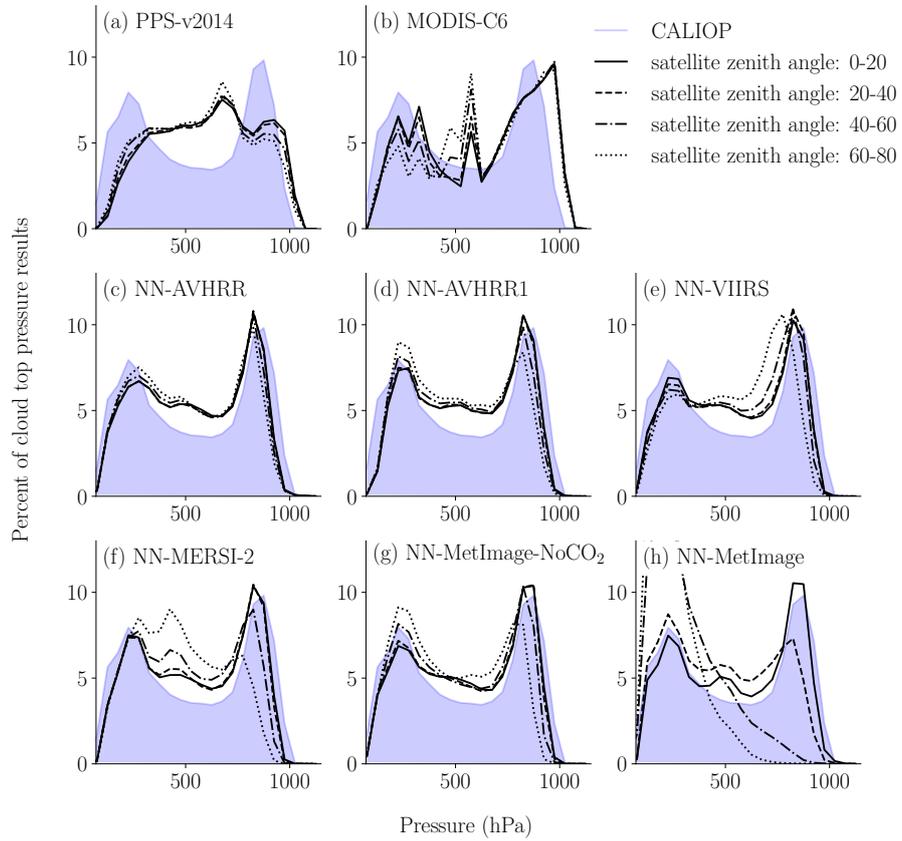


Figure 1. Retrieved pressure dependence on satellite zenith angle. CALIOP pressure distribution is shown in light blue. The percent of results are calculated in 50 hPa bins. The final validation dataset is used (see Table 1).

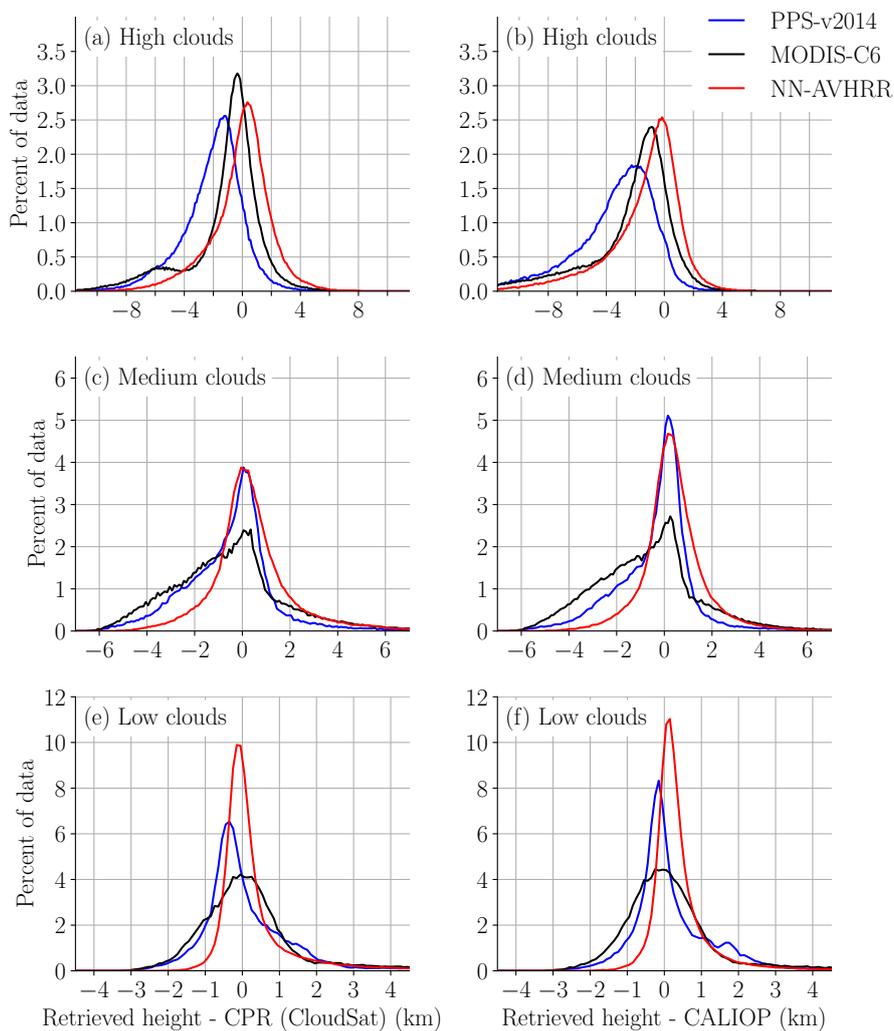


Figure 2. Bias-Error distribution compared to CPR (CloudSat) (left) and CALIOP (right). The percent of data is calculated in 0.1 km bins. For CALIOP the low, medium and high clouds are determined from CALIOP feature classification flag. For CPR (CloudSat) the low, medium, high clouds are determined from CPR (CloudSat) height compared to NWP geopotential height at 440 hPa and 680 hPa. The final validation dataset (see Table 1) where all algorithms had a height reported is used. Note that the values on the y-axis are dependent of the bin size. The peak at 11% for NN-AVHRR in subplot (f), means that 11% of the retrieved heights are between the CALIOP height and the CALIOP height + 0.1 km.

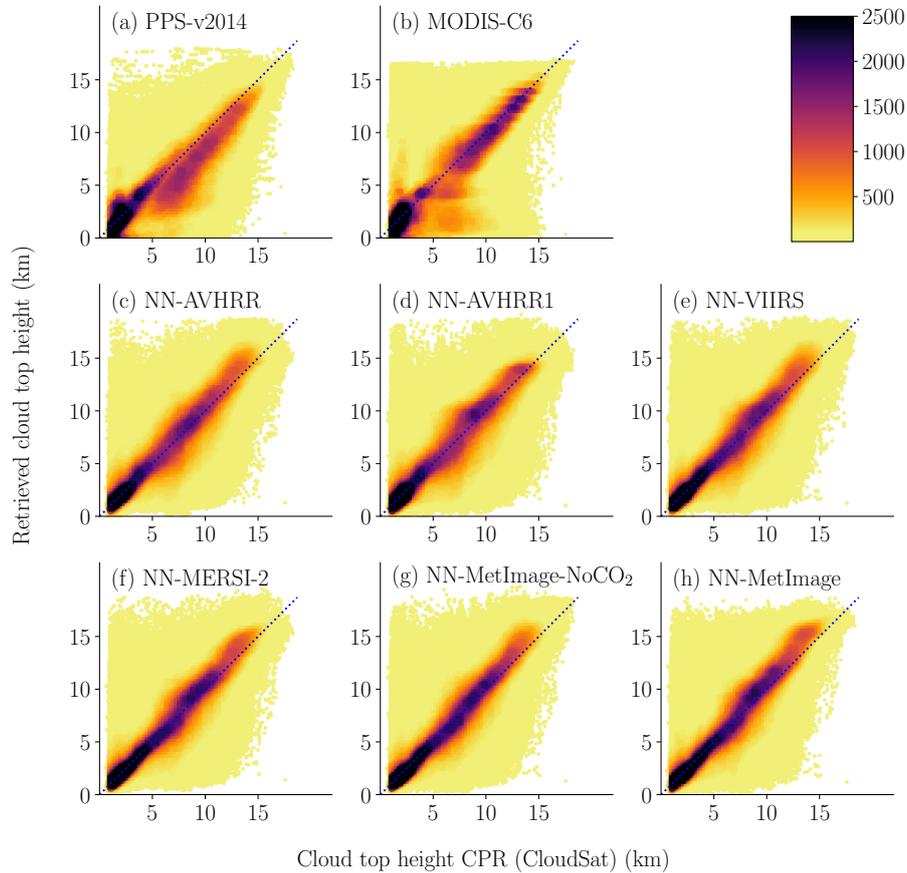


Figure 3. Scatters plot of the height for the neural networks and for the reference methods against CPR (CloudSat) height. The data were divided in bins of size 0.25×0.25 (kmkm) for colour coding. The number of points in each bin determines the colour of the point. The final validation dataset (see Table 1) where all algorithms had a height reported is used. ~~Five-Two~~ points where CPR (CloudSat) had a height above ~~22km-22~~ km were excluded. A cloudy threshold of ~~30%-~~ % is used for CPR (CloudSat).

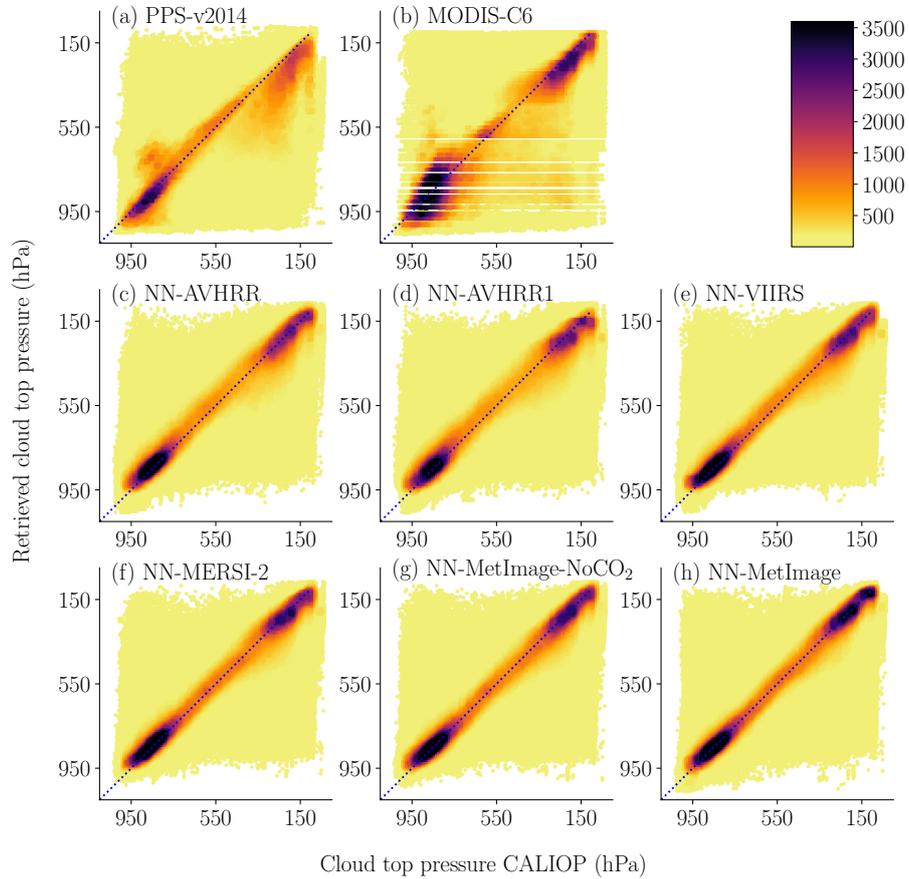


Figure 4. Scatter plots of the pressure for the neural networks and for the reference methods against CALIOP cloud top pressure. The data were divided in bins of size 10 x 10 (hPa hPa) for colour coding. The number of points in each bin determines the colour of the point. The final validation dataset (see Table 1) where all algorithms had a height reported is used.

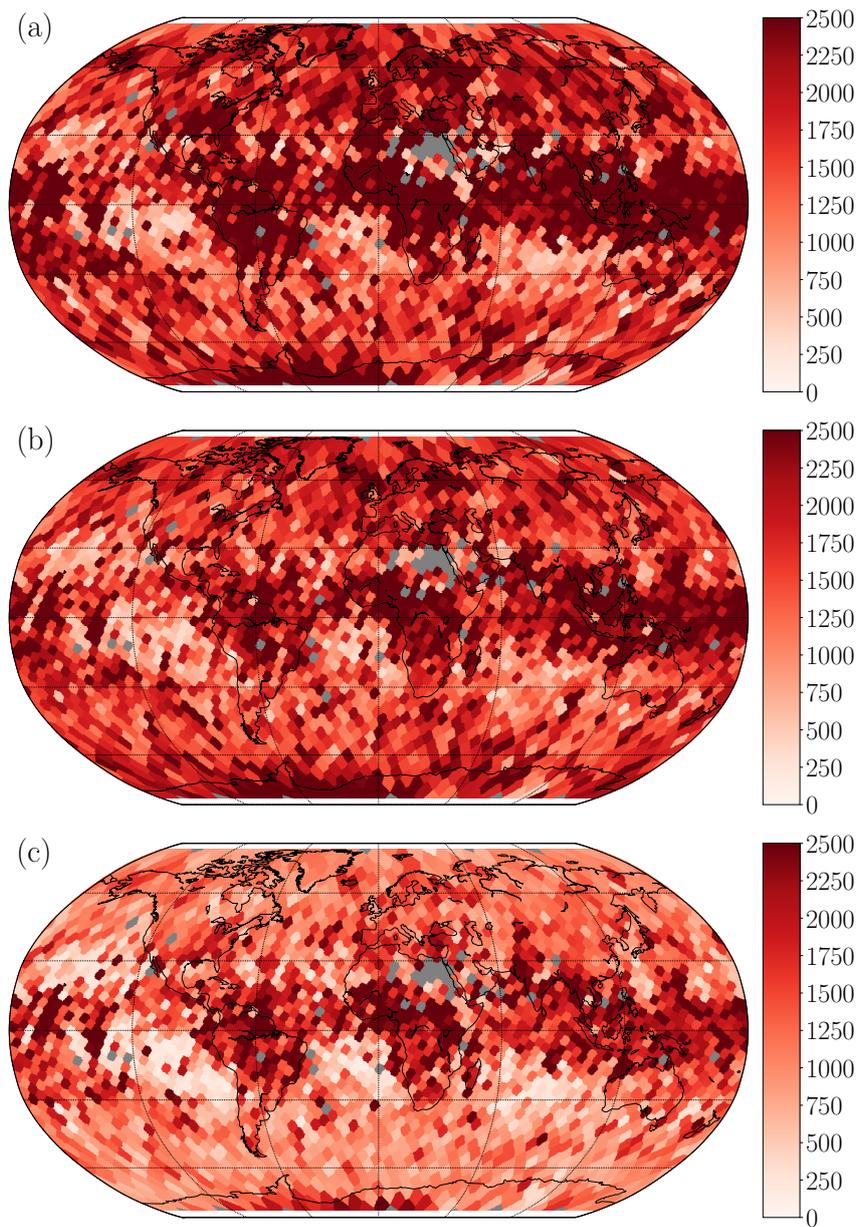


Figure 5. Mean absolute error in meters compared to CALIOP height. From the top a) PPS-v2014, b) MODIS-C6, and c) NN-AVHRR. Results are calculated for bins evenly spread out 250 km apart. Bins with less than 10 cloudy pixels are excluded (plotted in dark grey). The final validation and testing under development data (see Table 1) are included to get enough pixels.

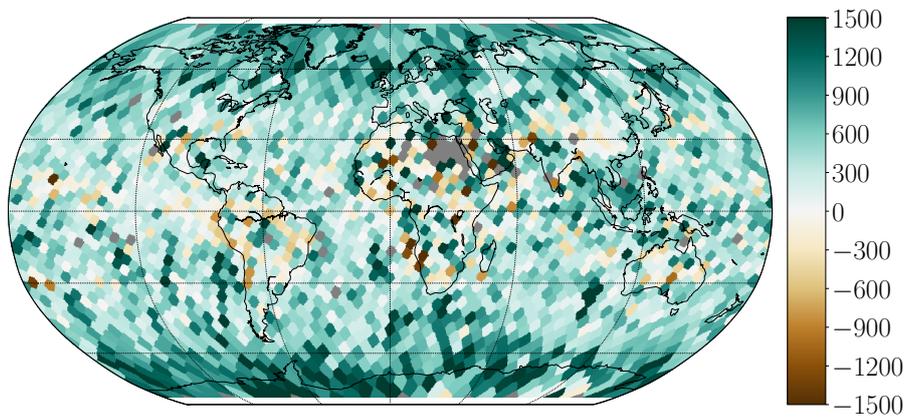


Figure 6. Mean absolute error difference in meters between MODIS-C6 and NN-AVHRR compared to CALIOP. Results are calculated for bins evenly spread out 250 km apart. Bins with less than 10 cloudy pixels are excluded (plotted in dark grey). Dark green means NN-AVHRR is 1.5 km better than MODIS-C6, dark brown means MODIS-C6 is 1.5 km better than NN-AVHRR. The final validation and testing under development data (see Table 1) are included to get enough pixels.

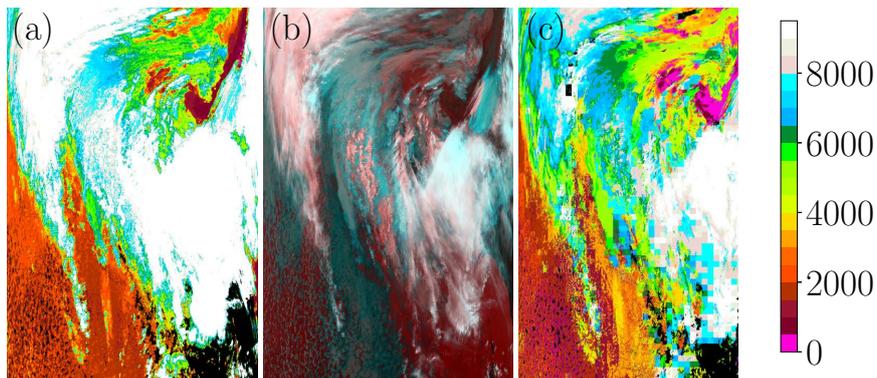


Figure 7. Comparing the cloud top height from the NN-AVHRR (left) to PPS-v2014 (right) with a RGB in the middle using channels at 3.7 μm , 11 μm , 12 μm . *Cloud retrievals below 2 are red, brown or purple. Cloud retrievals between 2 and 5 are orange and yellow. Cloud retrievals between 5 and 8 are green and blue. White pixels are cloud retrievals above 8.* Notice that the NN-AVHRR is smoother, contain less nodata and that the small high ice clouds in the lower part of the figure are better captured. This is from MODIS on Aqua 14th of January 2010, 00:05UTC.

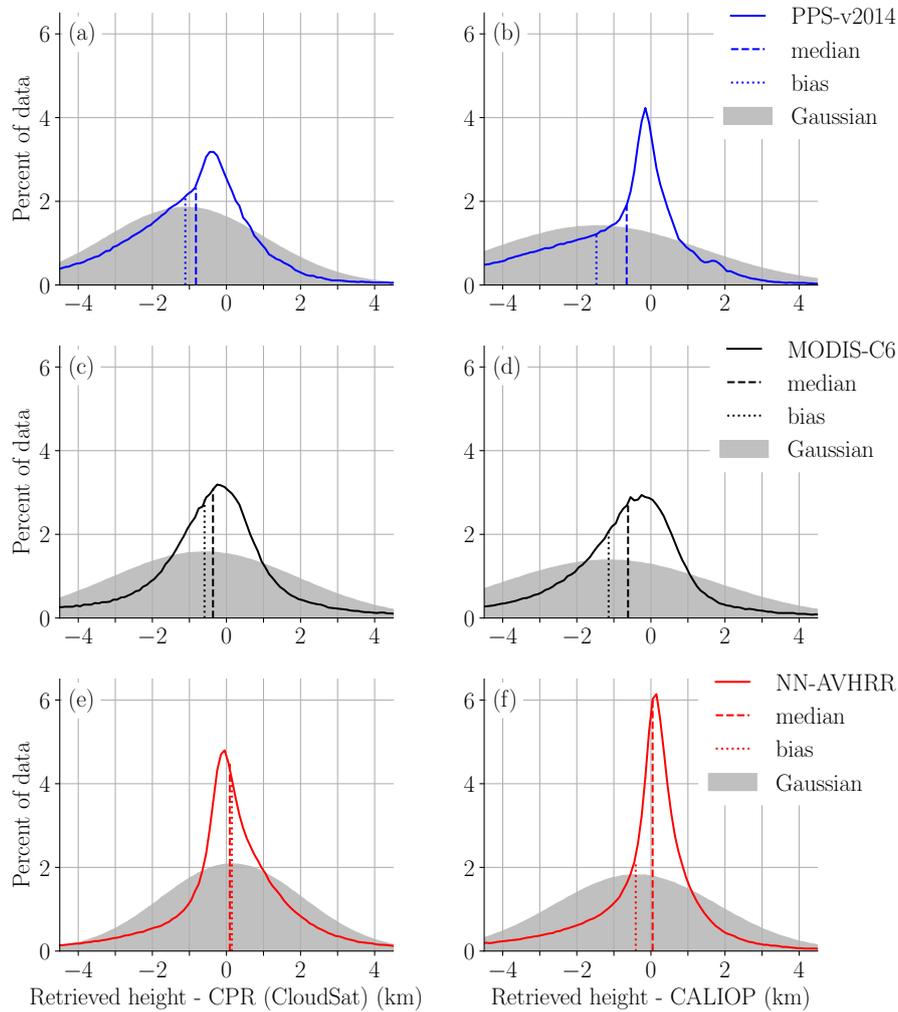


Figure 8. Error distribution compared to CPR (CloudSat) (left) and CALIOP (right) with biases and medians marked. The percent of data is calculated in 0.1 km bins. The final validation dataset (see Table 1) where all algorithms had a height reported is used. Note that the values on the y-axis are dependent of the bin size. The peak at 6 % for NN-AVHRR in subplot (f), means that 6 % of the retrieved heights are between the CALIOP height and the CALIOP height + 0.1 km. In grey the Gaussian distribution with the same bias and standard derivation is shown.