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 will correct this.

2.2 Referee comment:

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

Reply:

As both acronyms MODIS and AVHRR are better known than their written-out form we argue to keep just the acronym for these two, following the manuscript-preparation guidelines. However we will check all acronyms again as there are others (some noted also by Referee 2) that we should define.

2.3 Referee comment:

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

Reply:

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

Reply:

2.4 Referee comment:

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

Reply:

We will correct the channel ranges mentioned.

2.5 Referee comment:

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

Reply:

Because it is currently not clear enough we will explicitly state that both single and multilayer clouds where included. We will also clarify that we used the uppermost layer of the top layer pressure variable as this is missing in the text (also noted by Referee 2).

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

2.6 Referee comment:

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

Reply:

We do not agree that biases are at all critical, but rather we claim that there is a large risk of misinterpretation of the biases as we are handling non-Gaussian skewed distributions. Our preference for the MAE over bias and SDD has grown over the years, but first faced with the direct request to include also bias in this article did we fully investigate why they are less useful or even misleading. We thank Referee 1 for raising this question. And as bias and SDD are often included in previously published comparison studies the exclusion of them need to be motivated in the paper.

To make our point clear, tables with biases and SDD are included in this answer for CALIPSO (Table A1), and CloudSat (Table A2) but we argue that they should best be left out of the article.

Consider bias and SDD for low level clouds for PPS-v2014 and NN-AVHRR in Table A2, seeing these results the PPS-v2014 must be the better algorithm for low level clouds. The small improvement in SDD with $16\mathrm{m}$ can surely not be worth the $233\mathrm{m}$ higher bias! However considering the distribution on the differences (Figure 2 (e)) in combination with the better MAE in Table 7 we would claim that the NN-AVHRR is the better algorithm also for low level clouds. How could the bias and SDD indicate the opposite, what is going on here? The low bias for PPS-v2014 is because it underestimates the height for most low clouds which nicely compensates for amount of clouds being places 1.5-2km too high. The $472\mathrm{m}$ bias for NN-AVHRR indicates a difference distribution for NN-AVHRR centred at $472\mathrm{m}$ but that again is not reflected in Figure 2 (e). The peak of the difference distribution seams to be located much closer to zero. The NN-AVHRR distribution in fact places 89.3% of the low clouds within +/-$1.5\mathrm{km}$, to compare with 83% PPS-v2014. And considering this part (most) of data the bias for NN-AVHRR is $19\mathrm{m}$, to compare with -$214\mathrm{m}$ for PPS-v2014. So the large positive bias of $472\mathrm{m}$ does not mean that the difference distribution peaks at $472\mathrm{m}$ it is instead a consequence of the fact that the error distribution is skewed; low level clouds are more often placed too high than too low. This naturally is the case as there is always fixed limit (ground) for how low clouds could realistically be placed. If we consider bias and SDD also the skewness of the distributions should be considered at the same time, but this makes interpretation of the results even harder.

Regarding SDD we argue that MAE already provides a better estimate of the spread of the data. The MAE is not dependent of the bias and it is less influenced by outliers and the largest errors compared to the SDD. Focusing on SDD, during algorithm development, could make it seem much more important to improve pixels that are $15\mathrm{km}$ off to just $10\mathrm{km}$ off compared to improve pixels that are $2\mathrm{km}$ off. Considering again Figure 2 (e), how could there be such a small difference in SDD in Table A2

between NN-AVHRR and PPS-v2014, is that not unrealistic? Surely the NN-AVHRR distribution seems more centred around zero? If we look at the cases with difference larger than 4km which are not visible in Figure 2 (e) for NN-AVHRR they have a bias of 6.7km and (and consist of 4.03$\%$ of the low cloud data) and for PPS-v2014 they will have a bias of 6.3km and (consist of 3.7$\%$ of the low clouds data). This is what makes the SDD improvement so small. For the data with bias below 4km (more than 95$\%$ of the data) the SDD for NN-AVHRR it is 815m, compared to 1029m for PPS-v2014. As the SDD puts more focus on the largest errors, the increase of 0.3$\%$ low level clouds being predicted much (>4km) too high and the 0.4km increase in absolute error for these clouds evens out the improvement seen in the error distribution for the majority (96$\%$) of the data (Figure 2 (e)).

Of course the worst cases are important. The increased amount (0.3$\%$) of absoulte errors above 4km, discussed in previous section, could be a true degradation of the performance for NN-AVHRR compared to PPS-v2014. However we should specially remember that we are handling different FOVs and different instruments and we do expect differences for some pixels for example at cloud edges. When focusing on the worst performing part of data, neighbouring pixels both for the imager and the radar (or lidar) should be considered to decide if we are at a cloud edge. When investigating the worst cases it is important to make an effort to separate the true failures of the algorithms from errors expected from instrument and field of view differences.

To make clear that there is no general increase of outliers for the neural network method we made the same analysis as above and calculated amount of data with large (>4km) absolute errors for all cases in Table 6 and Table 7. In the example discussed above, PPS-v2014 low level clouds validated with CloudSat, is the only case where the neural networks have a larger percentage of large errors compared to PPS-v2014. MODIS-C6 data have a greater amount of large errors compared to all of the neural networks in all cases.

In addition to the other problems with interpreting bias for cloud top height retrieval, the overall bias are very dependent on the frequency distribution of high, medium and low clouds which is dependent on the validation dataset used and the performance of the cloud mask algorithm. This means that even if bias and SDD are included, the total bias and SDD should really not be compared between different studies that use different validation datasets and different cloud masks. The possibility to compare statistical measures between studies is further limited by the usage of different validation strategies including: filtering, averaging, only single layers, only low clouds etc.

As another illustration of the problem with bias consider a fictional cloud height retrieval algorithm with an error distribution that are Gaussian distributed ($\sim N(0, 2km)$). The zero bias here actually marks the centre of the distribution and together with the SDD describes the expected errors well. Now an improvement to the algorithm is made and clouds that where before placed more than 2km too low are now placed exactly at the truth. This improvement effects one of the tails and the largest part of the error distribution stays the same, although we now have quite a high percentage of data with zero error. Our mental picture with the error centred at zero is still valid (and the mode and the median are still 0). However if we calculate the bias for this truncated (and therefore skewed distribution) it is now 480m. A bias of 480 meters indicates an algorithm that generally places clouds too high. But as we know the underlying distribution we know that is not true; the algorithm just never places any clouds much too low.

It is also possible for a distribution to have a bias close to zero and not be exceptionally good. See for example the NN-AVHRR1 bias of only -8m in Table A2, this stands out as clearly better than all the others. However considering the MAE in Table 7 and the median in Table A4 it is clear that also this network performs in line with the others.

We agree that information about tendencies are lacking in the tables in the result section, the MAE does not provide information if clouds are generally over or under estimated. Figure 2 provides a view of the error distribution for the three algorithms (although only one of the neural networks is included). Here also the sometimes bimodal behaviour is evident (for low level clouds PPS-v2014, and high level clouds MODIS-C6) and the skewness of the distributions are evident. We will extend the discussion about the results in Figure 2. We will add a tendency measure, the median, which is more suitable for skewed distributions to Table 6 and Table 7 and we thank Referee 1 for suggesting tendency measures as an improvement to the article. We will also add a discussion why MAE and median are used and not the traditional bias and SDD. We will mention that the median should be interpreted with caution and that it will not be the same as the mode (where the peak of the error distribution is) again because the distributions are skewed and in some cases even bimodal. We suggest that we add Figure A1 to the article to help the discussion by visually showing why numercial valuse for the biases where excluded. In Figure A1 the error distributions are plotted together with Gaussian distributions with the same bias and SDD and the differences are evident. For the completeness we here included the same figure for low, medium and high (Figure A2-A4) cloud classes however we suggest that these will be left out of the article. These could be included as supplementary figures. We will also consider including the Median absolute deviation (MAD) or the interquartile range (IQR) as these are more robust measures of variablity less sensitve to outliers compared to SDD.

To summarize the above discussion: we are dealing with skewed, non-normal and even bimodal distributions and this makes the median a better measure of tendency than the bias. As we are not very interested in how large the largest errors are, especially as these are partly expected due to the sensor differences, the MAE is a better measure of spread compared to SDD.

See also further comments in the reply to Referee 2 who also wanted SDD included.

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 will improve the discussion regarding this, see reply to comment 2.9. Nearest neighbour matching is used; we will add this information in the article as the description of the matching method is now missing.

Clouds not detected at all by CloudSat are not a problem as it simply means that we will have less data. That the CPR often misses the top portions could partly explain why results are not improving for NN-MetImage and NN-MetImage-NoCO$_2$ (compared to NN-MERSI-2) when validating with CloudSat. We will add this in the discussion.

The vertical resolution of CPR is 0.5km this means that we should expect MAE higher than 250m. We will add discussion about this and about the medians that will be added to the result section instead of biases.

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 will correct that.

2.9 Referee comment:

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

Reply:

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

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 will better motivate the inclusion of CloudSat in the paper.

What happens if the neighbouring pixels are not used is to some extent seen in Table 5, but not well enough described in the results (page 7, line 19) and overlooked in the discussion section. We will discuss these results in more detail in the discussion section to support better the statement in the conclusion.

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

2.10 Referee comment:

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

Reply:

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

2.11 Referee comment:

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

Reply:

We will extend the discussion section with help of the questions raised by the Referees.

2.12 Referee comment:

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

Reply:

We will extend Section 6, reflecting what is added to Section 5.

Table A1. Bias, standard deviation (SDD) and skewness of the error distribution in meters for different algorithms compared to CALIOP top layer altitude. Skewness where calcualted with the scipy.stats skew function. Running the skewtest function show that all distribution have a skewness that differ from the normal distribution. Interpret with caution; as the distributions are skewed the bias are not at the center of the distribution.

	Bias [m]				Std [m]				Skewness			
	all	low	medium	high	all	low	medium	high	all	low	medium	high
PPS-v2014	-1473	295	-373	-3436	2797	1409	1611	2799	-1.0	3.0	0.2	-0.9
MODIS-C6	-1162	204	-730	-2524	2854	1518	2179	3313	-1.5	2.7	0.6	-1.5
NN-AVHRR	-416	427	362	-1447	2175	973	1330	2691	-1.8	3.9	0.7	-1.4
NN-VIIRS	-422	369	199	-1350	2049	941	1194	2565	-2.0	4.6	0.5	-1.6
NN-MERSI-2	-431	342	123	-1318	1967	930	1080	2466	-2.1	4.7	0.7	-1.7
NN-MetImage-NoCO$_2$	-428	279	63	-1234	1938	955	1128	2448	-2.1	4.8	0.8	-1.8
NN-MetImage	-306	242	12	-907	1774	979	1063	2271	-2.1	5.4	0.7	-2.0
NN-AVHRR1	-593	374	193	-1736	2260	931	1360	2771	-1.8	3.7	0.4	-1.3

Table A2. Bias, standard deviation (SDD) and skewness of the error distribution in meters for different algorithms compared to CPR (CloudSat) Height. Skewness where calcualted with the scipy.stats skew function. Running the skewtest function show that all distribution have a skewness that differ from the normal distribution. Interpret with caution; as the distributions are skewed the bias are not at the center of the distribution.

	Bias [m]				Std [m]				Skewness			
	all	low	medium	high	all	low	medium	high	all	low	medium	high
PPS-v2014	-1122	239	-509	-2056	2179	1605	1859	2097	-0.1	2.7	0.5	-0.5
MODIS-C6	-630	445	-528	-1217	2555	2123	2481	2602	-0.1	2.9	0.9	-1.2
NN-AVHRR	151	472	411	-114	1945	1589	1774	2128	0.2	3.8	1.2	-0.6
NN-VIIRS	142	459	314	-85	1918	1687	1765	2052	0.5	4.2	1.5	-0.7
NN-MERSI-2	139	429	263	-55	1807	1651	1684	1903	0.5	4.0	1.7	-0.9
NN-MetImage-NoCO$_2$	149	352	204	25	1849	1689	1742	1955	0.7	4.1	1.8	-0.7
NN-MetImage	290	397	228	260	1867	1831	1824	1899	0.9	4.0	1.9	-0.8
NN-AVHRR1	-8	520	303	-400	1974	1651	1752	2117	0.0	3.5	1.0	-0.8

Table A3. Median error in meters for different algorithms compared to CALIOP top layer altitude. The final validation dataset, 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. Interpret with caution; it is not the same as the mode of the error distribution. These columns is suggested to be included in Table 6.

	Median [m]			
	all	low	medium	high
PPS-v2014	-653	-50	-77	-2923
MODIS-C6	-620	-19	-666	-1588
NN-AVHRR	47	211	291	-797
NN-VIIRS	25	179	177	-719
NN-MERSI-2	-6	156	75	-706
NN-MetImage-NoCo$_2$	-53	94	20	-607
NN-MetImage	-20	71	-3	-359
NN-AVHRR1	-46	193	177	-1101

Table A4. Median error in meters for different algorithms compared to CPR (CloudSat) Height. The final validation dataset, 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. Interpret with caution; it is not the same as the mode of the error distribution. These columns is suggested to be included in Table 7.

	Median [m]			
	all	low	medium	high
PPS-v2014	-849	-156	-298	-1797
MODIS-C6	-384	49	-598	-619
NN-AVHRR	94	25	213	141
NN-VIIRS	75	3	120	165
NN-MERSI-2	46	-24	17	177
NN-MetImage-NoCo$_2$	13	-101	-43	216
NN-MetImage	91	-97	-41	459
NN-AVHRR1	36	54	151	-57

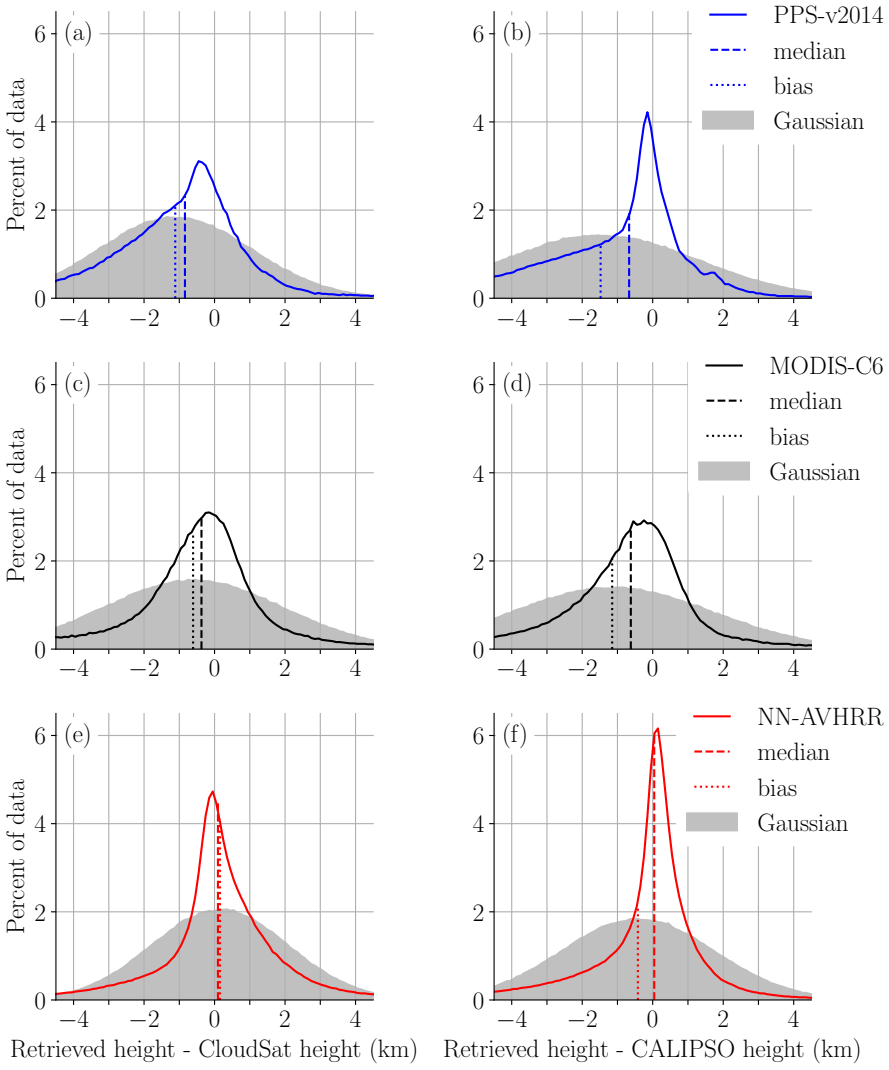

Figure A1. Error distribution compared to CPR (CloudSat) (left) and CALIOP (right) with biases and medians marked. In grey the Gaussian distribution with the same bias and standard deviation as the error distribution is shown. The percent of data is calculated in 0.1km bins. All height classes (low, medium and high) are included. 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 + 0.1km. This figure is suggested to be included in the main article. The x-axis is cut at 4.5km and some part of the ditribution is not shown. For PPS-v2014 92% is visible for CloudSat and 86% is visible for CALIPSO. The corresponding numbers are for MODIS-C6 89% and 90% and for NN-AVHRR 94% and 96%.

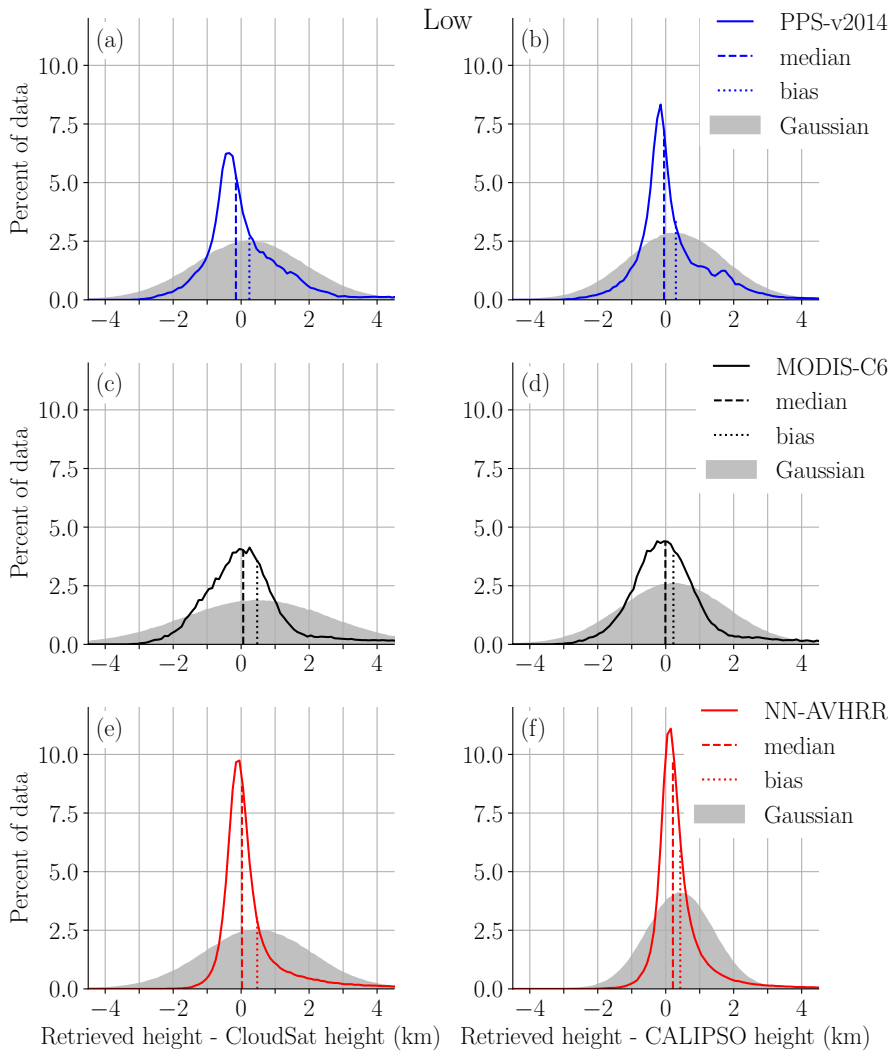

Figure A2. Error distribution compared to CPR (CloudSat) (left) and CALIOP (right) with biases and medians marked. The percent of data is calculated in 0.1km bins. Only the CALIOP low class is included. In grey the Gaussian distribution with the same bias and standard deviation as the error distribution is shown.

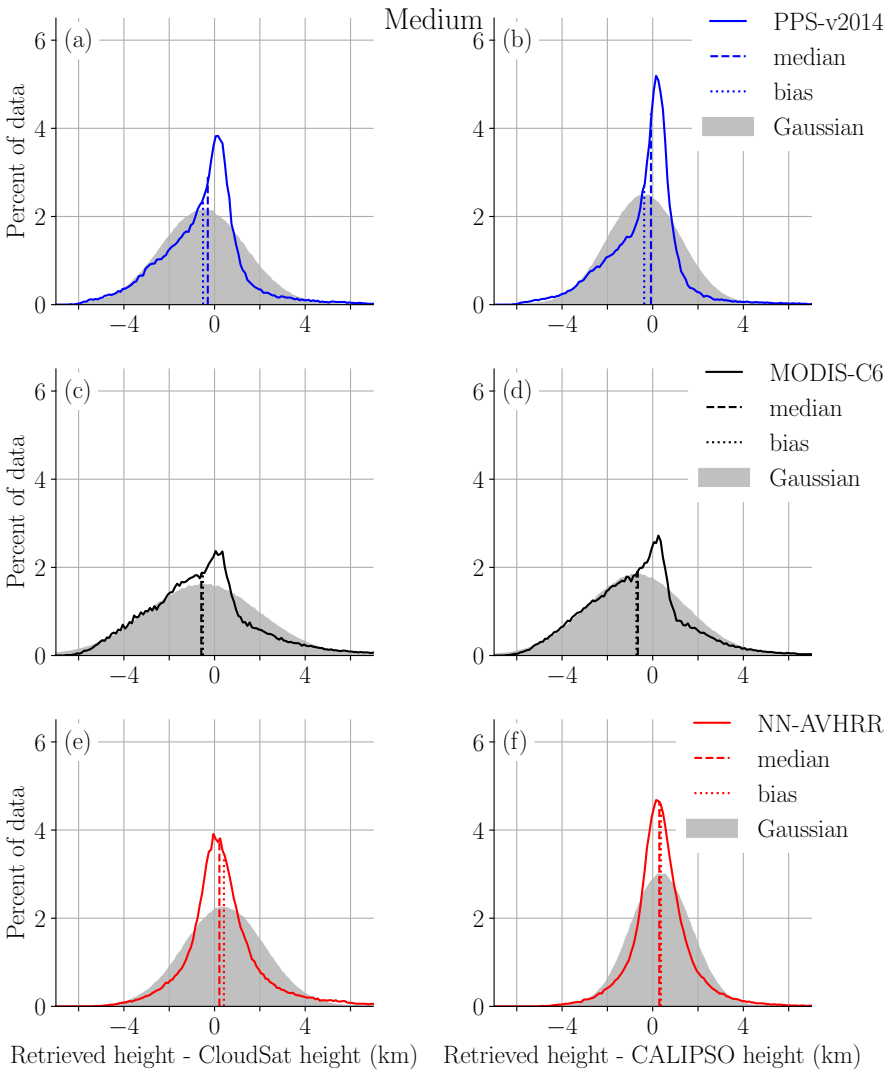

Figure A3. Error distribution compared to CPR (CloudSat) (left) and CALIOP (right) with biases and medians marked. The percent of data is calculated in 0.1km bins. Only the CALIOP medium class is included. In grey the Gaussian distribution with the same bias and standard deviation as the error distribution is shown.

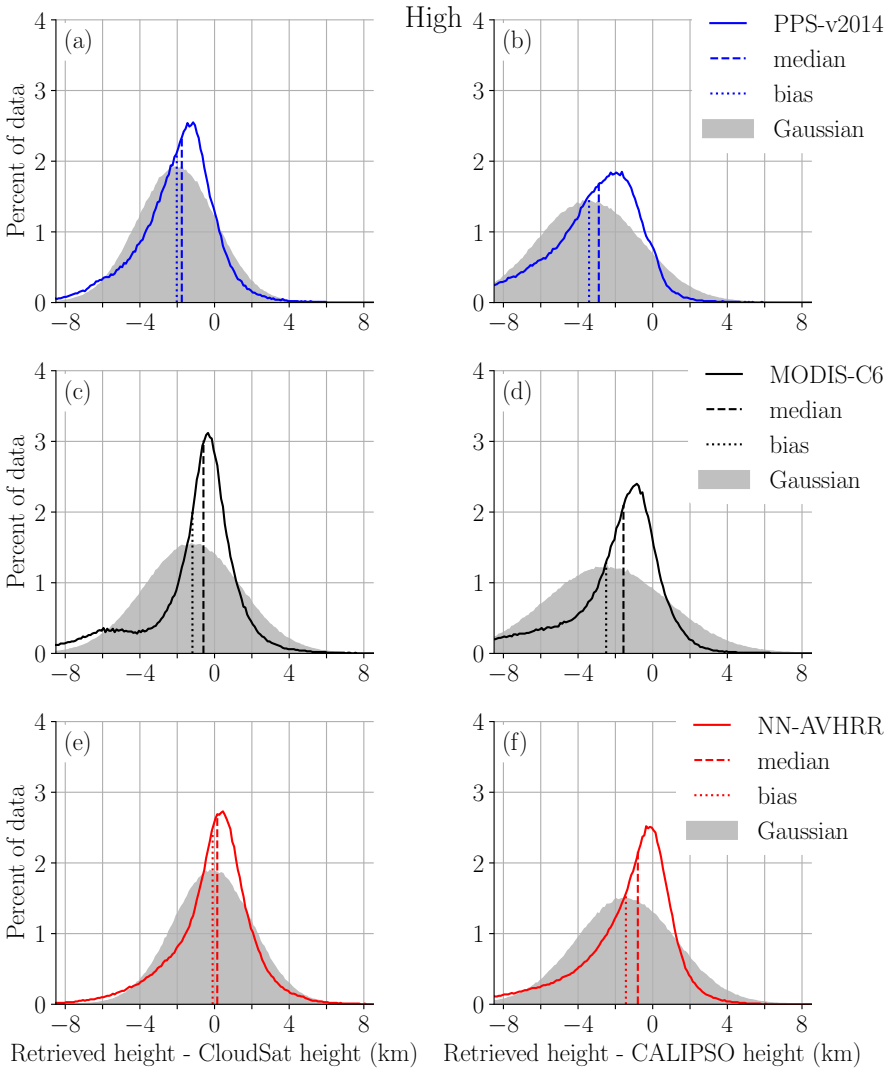

Figure A4. Error distribution compared to CPR (CloudSat) (left) and CALIOP (right) with biases and medians marked. The percent of data is calculated in 0.1km bins. Only the CALIOP high class is included. In grey the Gaussian distribution with the same bias and standard deviation as the error distribution is shown.