Final reply to Referee 1's review of the AMTD paper

" Detailed characterisation of AVHRR global cloud detection performance of the CM SAF CLARA-A2 climate data record based on CALIPSO-CALIOP cloud information"
by
Karl-Göran Karlsson and Nina Håkansson, SMHI

Note: All line numbers referred to below are relevant for the revised manuscript version written in Word change track mode and named "CLARA_A2_validation_AMT_2017_version2_tracked_changes".

Repeating general comment, part 1:

This manuscript evaluates the cloud mask of the CLARA-A2 climate data record (based on passive imagery from AVHRR polar orbiters) with collocated active cloud detections (CALIOP). Another, more general, paper has been published in ACP this year, and this AMT paper focuses exclusively on the cloud mask. This approach is sufficiently well justified, but the paper under review relies too much on the earlier publication (Karlsson et al., 2017; also to some extent on Karlsson et al., 2013) to explain the background. In order to qualify for publication in AMT, revisions need to be made to ensure that it can stand on its on, while not replicating too many of the science results.

Reply:

The two referred papers (especially Karlsson et al., 2013) are important papers in that they set the stage and define the framework for how to perform the matchups between AVHRR and CALIOP data. We get the feeling from some of the comments that we need to clarify the framework even further (i.e., that it is not enough to just provide the references). Thus, we have included a short summary of the most important points (new section 3.2) concerning the basic matching or collocation methodology (see also reply to **general comment, part 4** further below).

Regarding the justification of this paper and the question whether it adds anything new compared to the paper by Karlsson et al. (2017), we claim that one important objective of this study was to investigate the impact of upgrading the

results by using the new CALIPSO-CALIOP version 4 dataset (which indeed is the main topic of the Special AMT Issue too which the paper was submitted). The previously mentioned validation efforts were all based on CALIPSO-CALIOP version 3 data. We have now added this objective in the Introduction section (lines 82-83).
However, another more important objective was to show that the CALIPSO-CALIOP dataset can be used to investigate much more in detail the cloud detection limitations of one particular cloud screening method (like the one used for CLARA-A2) than what has been presented before. The concept of Cloud Detection Sensitivity (as illustrated by results in the original Figure 11, now Figure 12 in the revised manuscript) is a new approach which we hope can become a standard tool for a more objective evaluation of cloud climate data records in the future. Its main advantage is that it can be considered as a universal method, not depending specifically on the actually studied AVHRR dataset. It is a method based on a special organization of the CALIOP cloud dataset by use of estimated cloud optical thickness sub-categories. These results, being organized in cloud optical thickness sub-categories, can be compared to any other collocated satellite-based dataset.

We have emphasized better the objectives of the study in the Introduction section (lines 76-91) and highlight better the results and the potential of the derived Cloud Detection Sensitivities in current and future studies in the Conclusions section (lines 720-724, 765-772 and 787-799).

Repeating general comment, part 2:

In its current state, the paper is hard to review because some of the concepts are not explained sufficient well (specific examples are given below), and because details are left out. In addition, the manuscript is unnecessarily wordy in some places and has basic deficiencies with English/Grammar (for example, “were” is used instead of “where” throughout the manuscript; there are many run-on sentences; punctuation is used too sparingly; use of slang words such as “punish” for a statistical approach that are frequently used by the community, but should be used only where absolutely necessary). Before going into the copy/edit process at AMT, a native speaker should be consulted to ensure logical flow and readability of the manuscript overall.

Reply:

As being non-native English authors, we admit limitations in the ability to produce perfect quality (scientific) English text. We thank the reviewer for pointing out the most common errors and we have done our best in eliminating them. We have also consulted a native speaker before submitting the revised version of the manuscript.

Repeating general comment, part 3:

Despite the criticism of the presentation quality, the content is interesting in that the cloud detection capability is studied as a function of optical thickness and region. Obviously, the POD (probability of cloud detection) depends on surface albedo and emissivity, mechanisms that are identified by the authors. Two comments here:

1) It should be stated more clearly where such findings have been made previously. The author make a point that the regional assessment is new, but there have been previous studies that focused on some of the problematic regions specifically in the Arctic with CALIOP that are not cited here (for example, studies by Gettelman, Kay, L'Ecuyer and a few others).

Reply:

The knowledge of the dependency on surface characteristics (e.g. albedo or emissivity) for the possibility to separate clouds from Earth surfaces in satellite imagery is nothing fundamentally new. Rather, it is a well-established and well-known fact in the satellite user community. The reason is obvious: All cloud screening methods depend on the ability to find enough of contrast between clouds and underlying surfaces in the investigated images. This is valid for all spectral regions - be it visible, near-infrared, short-wave infrared or infrared. Multi-spectral methods will have the best capability since the use of many spectral channels increases the probability that at least one spectral channel will offer enough of contrast between clouds and Earth surfaces. This explains e.g. the high quality of cloud datasets from MODIS (with access to up to 36 useful spectral channels).

The challenges here are naturally largest at high latitudes and near the poles where we have both bright Earth surfaces (snow, ice) and very cold surface temperatures (very similar or even colder than clouds which normally are warmer than clouds in other regions). This explains the special interest here (as exemplified by the mentioned papers).

We have added some of these references (lines 760-763) since we agree that they absolutely need to be mentioned in this context. However, the most important thing is that we even stronger have emphasized that the proposed method offers a universal method (which could become a standard method) to monitor these problems globally and not just in specific regions (see reply above to **general comment, part 1**). This is the big advantage of the method. Thus, our statement about the novelty of the regional assessment should be interpreted as that the method offers both a monitoring of mean global conditions but also a regional monitoring including all regions on Earth and not just some selected ones.

Repeating general comment, part 4:

2) It remains unclear (partially because of the structural problems of the manuscript pointed out above) why there are some regions where cloud cover is overestimated by the passive imagers. One possible explanation is not sufficiently investigated: sub-grid resolution clouds that could be picked up by passive imagers but not by active imagers (if they are outside the FOV). There is some discussion of it, but it remains superficial.
Also, active observations are portrayed as the ultimate "judge" for the performance of the cloud mask derived from passive observations, and they shouldn't be. As pointed out by the authors, active observations have their own limitations (sensitivity, FOV, day-vs-night contrasts). The truth is that active cloud observations afford a different perspective on clouds that happens to be less sensitive to the surface reflectivity and emissivity than that of passive observations. This distinction (and the limitations of both approaches) should be made clear by the authors.

Reply:

We agree that we could have been clearer in the discussion of aspects that are related to the different FOVs of AVHRR and CALIOP. Some discussion is included on page 10 (lines 366-404) and on page 12 (lines 465-471) but this can be improved. Since other reviewers also have pointed out more or less the same thing we have done the following:

1. We introduced a short summary of the underlying basic method of how we matched AVHRR and CALIPSO data (Section 3.2). It seems the current referencing to the original paper by Karlsson and Johansson (2013) (which describes the matching method) is not enough for a full understanding. We need to recapitulate the method's most important aspects also in this paper.

2. We added a clear illustration (new Figure 1 in Section 3.2) of how matched high-resolution AVHRR FOVs relate to the CALIPSO-CALIOP FOVs within a nominal AVHRR GAC pixel. This would help understanding the problem.

3. We expanded the discussion of these results in a new Discussion section (lines 642-695). Thus, the previous Discussion section is now split into one separate Discussion section and one final Conclusion section. The problem of inter-comparing CALIOP data with other satellite data in cases of highly scattered and fractioned cloudiness is now discussed in more depth in the new Discussion section. In our opinion this aspect has been largely overlooked in many previous papers using CALIPSO-CALIOP data as the main validation source.

The question on why there seems to be regions where cloudiness is overestimated is interesting but the reasons behind this is at least partly beyond the scope of this paper. We do express some qualified guesses about the reasons for some of the found deviations in the study but, basically, this is really up to the algorithm originators to further analyze and explain in subsequent studies. However, we think that it cannot really be related solely to "sub-grid resolution clouds that could be picked up by passive imagers but not by active imagers (if they are outside the FOV)". This mismatch can definitely occur for individual AVHRR GAC pixels and for individual orbits but when summed up in a climatology based on thousands of orbits such biases will end up to be either very low or non-existing. Simply since the opposite case (clouds picked up by active sensors and not by passive sensors) is just as likely to occur. We have explained that in relation to the illustration envisaged in point 2 above (Figure 1, Section 3.2 and lines 642-695 in the Discussion section). But what is important is that the precision (variance) of the estimated mean cloud cover will degrade (i.e., higher RMS errors) when this occurs and this is emphasized in our discussion.
Another important thing is that the indicated overestimation may actually be caused by an inappropriate value of the global mean cloud detection sensitivity (i.e., minimum cloud optical thickness) for regions were cloud detection is very efficient. This is discussed in lines 627-640 in the Discussion section.

Regarding the choice of CALIPSO-CALIOP data as the "ultimate" judge, we both agree and disagree with the Reviewer's opinions. Admittedly, active data has its limitation where the FOV representability in relation to the AVHRR GAC FOV is perhaps one of the largest (as discussed above). However, for clouds with scales larger than the AVHRR GAC FOV (5 km) we still claim that no other observation reference can provide a better estimation of global cloud presence and distribution than the CALIPSO-CALIOP dataset. The big advantage with the CALIOP information is that we measure the lidar reflection from real cloud particles (in the CALIPSO-CALIOP version 4 dataset also quite confidently separated from aerosol particles) and from the backscatter energy we can also for the thinnest clouds estimate with high accuracy the optical thickness of the cloud layers (up to a certain maximum value). No other sensor can provide the same. MODIS data is an alternative but in our opinion the MODIS dataset share many of the problems experienced by dataset produces from most multispectral passive sensors (AVHRR, SEVIRI, VIIRS, ABI, etc.) and this is basically explained by the fact that the measurement always contain a mix of contribution from clouds, the atmosphere and the surface (especially in the cases of thin clouds). We cannot be sure that we only measure the impact of the cloud itself. For an active sensor we do not have the same problem. However, most important in this context is that for a study like this it is very important to have access to very accurate estimations of cloud optical depth for the very thinnest clouds in order to carry out a sensitivity study like this. Here the CALIPSO-CALIOP measurement is quite superior to MODIS. For the latter sensor, estimations of cloud optical thickness of the thinnest clouds have high uncertainties due to the strong dependency on radiance contributions from the underlying surface and atmosphere. This is the main reason for using CALIOP data instead of MODIS data. We have explained the importance of having access to accurate estimations of cloud optical thicknesses (lines 87-89, 146-154, 195-200 and the entire section 3.5) in order to carry out our study. With this information and background we think there is no need to discuss why we have chosen CALIOP instead of MODIS in this paper.

In the following we will address selected short comments (which are not simply editorial):

p2,L60: Why is CALIOP singled out as important for cloud observations, where in fact MODIS is flown in the A-Train as well. Wouldn't the MODIS observational record, in conjunction with CALIOP, lend itself to a similar study as the one presented here? Of course, its data record is much shorted, but on the other hand, MODIS and CALIOP are collocated all the time, by design.

Reply:

We just gave some arguments in the reply above to **general comment, part 4**. It is our opinion that CALIOP data is a better reference in the sense that the measurement information is free from surface (and atmospheric water vapour and aerosol) dependence. However, even more important is that we cannot use MODIS data for the cloud detection sensitivity study since the cloud detection sensitivity of MODIS is probably not very different from AVHRR. More clearly, we repeat that we need access to very accurate cloud optical thickness estimations for very thin clouds in order to make such a study. The uncertainty of the MODIS-derived optical thickness in this optical thickness interval (values less than 1.0) is too high and at least much higher than for CALIOP-derived optical thickness. We have pointed out the requirement of very accurate optical thickness information for very thin clouds (see reply to previous comment) which we think is enough for justifying the choice of CALIOP as our reference.

It would actually be very interesting to do a similar study of the MODIS C6 cloud detection sensitivity with the same method as presented here. We would expect some improvements compared to CLARA-A2. Figure 6d in the CLARA-A2 paper in ACP indicates an almost constant bias in cloud cover of about 5 % for MODIS C6 over all latitudes (more clouds observed by MODIS). The question is then if the global distribution of the cloud detection sensitivity (=minimum detected cloud optical thickness) will decrease with a constant value everywhere or are there possibly regional differences?

p2,L60: The limitations of CALIOP (e.g., day time vs. night time detection, noise etc., strategies for thin cloud detection) should be discussed here.

Reply:

We have been using CALIPSO-CALIOP cloud information in cloud validation activities ever since 2007 (shortly after data was made available). For our applications, we have not encountered or noticed any specific problems regarding the efficiency in CALIOP cloud detection between day and night. CALIOP daytime results are a bit more noisy due to some reflected sunlight contaminating the signal but we believe that the enhanced noise is mostly relevant and serious for studies of very weak signals, e.g. from very thin aerosols. In the CALIOP version 4 dataset, a better Cloud and Aerosol discrimination method was introduced and the previous problems in misclassifying heavy aerosols as clouds over specific regions of the world has been taken care of (see lines 170-174). Consequently, we see no reason to add any deeper discussion on the quality of the CALIOP measurements. We think that the representativeness issues (i.e., that AVHRR and CALIOP probes different parts of the GAC FOV) as discussed in Section 2 and extensively discussed in the Discussion section is actually more serious than actual uncertainties of the CALIOP measurement.

p2,L70: The earlier study by Karlsson is cited here. It should be summarized in at least one paragraph since this paper needs to stand on its own. What was the scope of that manuscript? The extension by CALIOP, on the other hand, are well explained (with the caveats pointed out above).

Reply:

Done, see point 1 above in the reply to **general comment, part 4**.

p3,L79-87: This paragraph should be completely rewritten. The explanation of the field of view of the passive vs. the active instrument is vital for understanding this manuscript, yet it is incomplete. What is the GAC FOV vs. the FOV(passive) vs. the FOV(active, at native vs. aggregated resolution)? What data specifically are dropped?

The best way to explain this would be through a simple illustration of the AVHRR pixels vs. the CALIOP FOV of single shots, as well as the aggregation of individual pixels/ shots in the various products used in this study. Without this added figure, it will be hard to retrace the steps that were taken in this manuscript.

Reply:

Done, see point 2 above in the reply to **general comment, part 4**.

p3,L90: Which parameter retrievals? How is the radiance inter-calibration and data record homogenization done? Simply referencing Heidinger will not do because the specifics are missing. One of the clear requirements of AMT publications is that anybody reading the paper needs to be able to retrace the steps of a study from the original data to the findings. There is not sufficient detail provided here (or in other parts of the manuscript) to do that.

Reply:

The parameters we mention concern the different variables included in the entire CLARA-A2 dataset and we have clarified this (lines 113-123). Apart from cloud amount there are 7 different cloud properties, a surface albedo estimation and an estimation of surface radiation budget parameters.

It is true that there is still no follow-on paper to Heidinger et al. (2010) describing the upgraded calibration equations. But there is a recent publication in the GSICS Newsletter describing the associated PyGAC preprocessing software (ftp://ftp.library.noaa.gov/noaa_documents.lib/NESDIS/GSICS_quarterly/v11_no2_2017.pdf) . PyGAC contains the final calibration which was used for the

CLARA-A2 processing and is available as an open source package. We have added this reference to the manuscript (line 113 and lines 837-840).

p4,l118-127: See comment above. These sections cannot be understood without better explanations of the FOVs, data aggregation and homogenization.

Reply:

See the reply to **general comment, part 4**.

p4,l129-134: Provide description of specific NOAA orbits that were included (vs. those that were not). Also, why were MODIS observations NOT used? The minimum information for the NOAA observations are: (a) instrument/satellite names and short description; (b) orbit inclination and equator crossing time; (c) life time of satellite; (d) orbital shifts over time

Reply:

See the reply to **general comment, part 4**. We have tried to cover all those aspects. The MODIS question has already been dealt with in the reply to the comment for **p2,L60.**

p4,l148: The theoretical deliberations on cloud mask/cover are insufficiently backed by literature. The paper that comes to mind when talking about the meaning of a "small" or "thin" cloud is that by Koren ("How small is a small cloud"). A short literature study on the topic would be advisable, given that it is the main topic of this article.

Reply:

Thanks for this advice. We have taken a closer look and added some adequate literature references (lines 177-185). What is clear, though, is that there are several aspects of this topic. The paper by Koren et al. (2008) discusses primarily the impact of varying sizes of small Cumulus clouds in fine resolution satellite imagery (e.g. Landsat) and this is perhaps not directly applicable to AVHRR data in the comparably much coarser GAC resolution. For GAC data, we are perhaps more interested in when large scale (in contrast to small cumulus) clouds become so optically and geometrically thin that they are not detectable any longer. This is the probably the most important aspect for GAC data. Nevertheless, also when cloud elements begin to approach a scale that is much finer than the GAC resolution (analogous to the cumulus case described by Koren et al. (2008)) we will also lose detectability. This is a very important aspect when trying to understand the implications of the matching geometry depicted in the new Figure 1. Consequently, we have expanded our discussion here on those aspects (Section 3.2 and lines 642-695 in Section 5).

p5,l184: “possibly punish AVHRR-based methods in an unfortunate and undeserved way: : :”: three words (punish, undeserved, unfortunate) are inappropriate for a scientific publications. There are multiple occurrences of such “personalized” or “humanized” comments, which should all be translated into objective, rather than “punitive” language.

Reply:

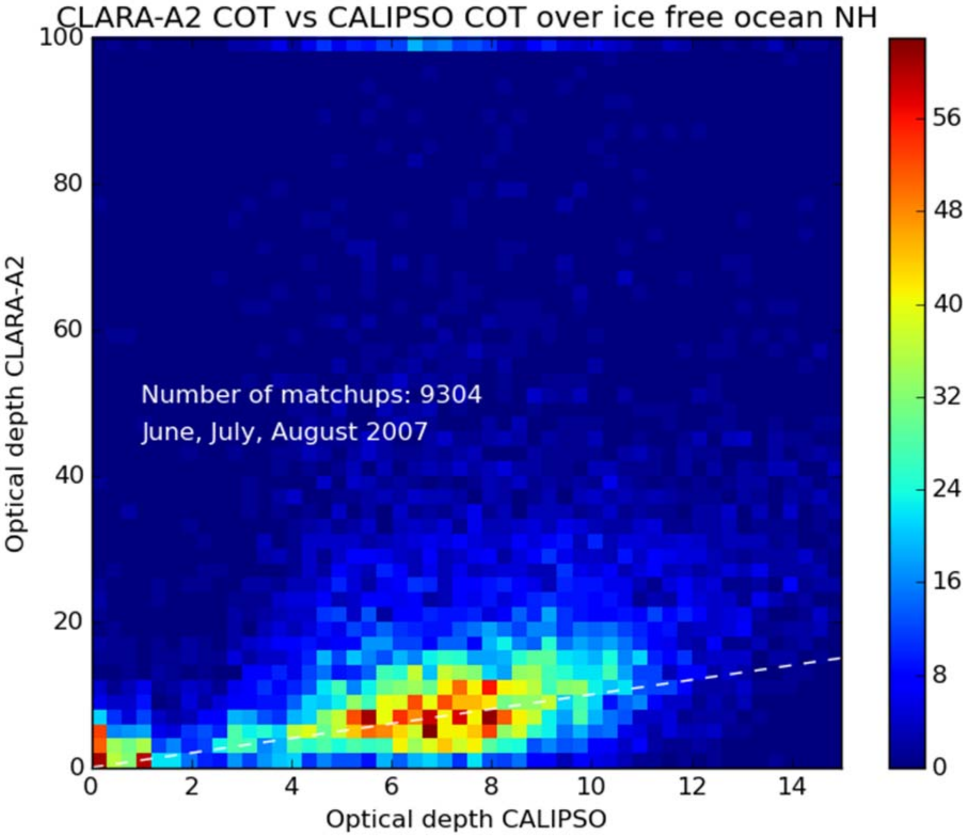
We have removed the use of non-scientific terminology.

p5, l187: The optical thickness threshold of 5 for CALIPSO is higher than usually assumed. If it is necessary for this study to work with such a high threshold, it should be justified, and it should be explained how this is possible (referring to literature where this has been done, or with a dedicated sub-section in this manuscript where it is shown that the lidar does, in fact, allow to go to COD 5, and under which circumstances).

Reply:

We admit that we do not have good support in the literature for stretching the useful upper limit of CALIPSO-derived COD to 5. However, in the description of the upgrade to CALIPSO-CALIOP version 4 it is also emphasized that previous cloud optical thicknesses in version 3 were generally underestimated. This is also clearly indicated in Figure 2 (new Figure 3) in the manuscript. Whether this increase entirely justifies moving the upper limit to 5 is still not clear.

We do have more indications from our own investigations that an adjustment of the upper limit seems possible. In a study related to a paper by Riihelä et al (2017) we investigated the correlation between CALIPSO-estimated and CLARA-A2 estimated CODs over various surfaces (with snow surfaces over Greenland as the main target). However, when isolating the collocated results over ice free ocean surfaces at high latitudes (noting that over a dark surface also the AVHRR-based estimations should be more accurate), we could clearly see a good correlation between the two estimations up to about COD=5 (see figure below):



Although this is not a perfect illustration (not included in Riihelä et al, 2017, but maybe considered for a follow-up paper) it shows how CLARA-A2-estimated optical depths compare to CALIOP-estimated optical depths in the range 0-15. Over a dark ocean surface the majority of values agree pretty well but what is clear is that an increasing number of cases (for higher optical depths) CLARA-A2 values saturates at 100 for CALIOP-values exceeding approximately 4 (noticeable at top of the figure). This reflects the inability of CALIOP to provide reasonable optical thicknesses for optically thick clouds. But, we made the conclusion that values compare pretty well even up to an optical thickness of 4-5 and this was one of the reasons why we decided to use the CALIOP interval 0-5 for this particular study (for AMT).

It is this finding that made us to use the maximum limit of 5 in this particular study. Unfortunately, in the end, we did not include this part of the inter-comparison in the finally published paper by Riihelä et al. (2017).

We propose that we keep the original maximum value of 5 in our plots but add a remark that values near this upper end are uncertain (lines 146-154). The upper limit is not crucial for the findings of our study since in most cases the cloud detection sensitivity is considerably lower than 5. Only for some positions over Greenland and Antarctica we approach these high values but whether the value is 3 or 5 here does not really matter since it deviates anyhow very much from the values found on other places (which is the main message).

The mentioned reference is the following:

Riihelä, A., Key, J. R., Meirink, J. F., Munneke, P. K., Palo, T., & Karlsson, K.-G. (2017). An intercomparison and validation of satellite-based surface radiative energy flux estimates over the Arctic. Journal of Geophysical Research - Atmospheres, 122(9), 4829–4848. https://doi.org/10.1002/2016JD026443

p6: There are multiple gaps on this page: The notion of "scores" (and different kinds) are used without sufficient (or any) explanation in this section, or in section 3.3. Too many questions remain, for example, which parameter of what satellite is validated with which other parameter, and how exactly as "score" (of any kind) is established.
How is the aggregation done? Why are scores only plotted as a function of COD up to 1, where in fact CODs up to 5 are advertised? What is the "improvement"? If the figures are insufficiently explained, it is not possible to understand. What has been "transformed from cloudy to clear cases" (l212), and how is that done? What is the role of Kuiper vs. hit rate (should be spelled "hit rate", not "hitrate"). Each of the bulleted items of the list on p6/p7 need to be explained and supported with formulae where appropriate. Here again, terms such as "punishing" should be avoided if at all possible.
After this paragraph, the reviewer was unable to give this a thorough review because the basics for understanding the remainder of the manuscript were not established.
The reviewer is willing to review another version of the manuscript where this has been fixed.

Reply:

We have improved the description here to improve the understanding of the method and the results. We had these questions in mind when dealing with point 1 in the reply to **general comment, part 4**. However, regarding the exact definition of the used validation scores (bullets on pages 6-7) we insist on that the reference to the paper by Karlsson and Johansson should be enough (although we have also added some clarifying comments on lines 338-346). This previous paper defines all these scores with illustrations and formulas. The current revised manuscript is already extended substantially as a consequence of

all the requests from reviewers and we think that further extensions shall be avoided where it is possible. Furthermore, all scores except the Kuipers score are standard scores provided with short text descriptions. Regarding the Kuipers score we have added a comment on how values should be interpreted (lines 345-346).
We are certainly grateful for the reviewer's willingness to check the revised manuscript.

p7,l265: This question is a great one, and at the center of this manuscript. However, the method description below is insufficient. Terms from machine learning ("overtrained") are evoked without explanation how they relate to the manuscript content. Also, here again, CALIOP is represented as the "objective" instrument that AVHRR is validated by where possible – where in fact the two instrument just assess different aspects of a cloud (see comment above).

Reply:

Yes, this is the core topic of the paper. In our opinion, the method of determining the Cloud Detection Sensitivity is a way of utilizing the sensitivity difference between the two sensors in the most optimal way. Regarding the use of terms "overtrained" and "overfitted" we actually insist on keeping them here. Bullet number 2 is important and expresses a general problem of cloud screening methods (not particularly the CLARA-A2 method) and how to train them (especially statistical regression methods and artificial neural networks). Again, we repeat that we think that the described method can be applied to investigate any cloud screening method and not just the one used in CLARA-A2. For that reason, this bullet is important.
We will improve the description here (the mentioned aspects have already been commented in replies to similar comments above).

p7, l278-l304: This seems to wordy and hard to follow since some of the concepts were not introduced.

p8, l306: Now some of the orbits are introduced, but that is too late in the manuscript. In addition to NOAA-18 and NOAA-19, did other data go into the CDR under investigation?

Reply:

To be dealt with as indicated in the reply to **general comment, part 4**. The used NOAA-18 and NOAA-19 data (being matched with CALIPSO) is exactly the same dataset as was used for the evaluation in the CLARA-A2 paper by

Karlsson et al, 2017). However, in that study also results for morning satellite data (NOAA-17, METOP-A, METOP-B) were presented, although only valid over a small latitude band around 70 degree latitude. Since this study focus on global conditions we excluded the morning satellite part of the dataset. Nevertheless, we keep some discussion on morning satellites since they are important for CLARA-A2 (almost 40 % of the CLARA-A2 data consists of morning satellite data). The matching with morning satellites will introduce a new type of matching problems. This will be discussed in the revised manuscript in connection with the discussion of the figure to be introduced in point 2 (new Figure 1 in Section 3.2) of the reply to **general comment, part 4**. We have also kept a discussion on how to deal with the validation of morning satellite data in the new Discussion section (lines 697-710).

p9,L350: insufficient introduction how systematic and random errors were establish make it hard to understand Figure 7.

Reply:

We think that the understanding and statistical definition of systematic and random errors should be well known in the scientific community. Systematic errors concern the error of the mean value (normally described as “bias”) while random errors define the typical variations (variability or variance) around a particular mean value. We insist on keeping the current formulations but we have introduced the terms “systematic” and “random” already in the descriptions of the used validation scores in Section 3.4 (lines 338-339).

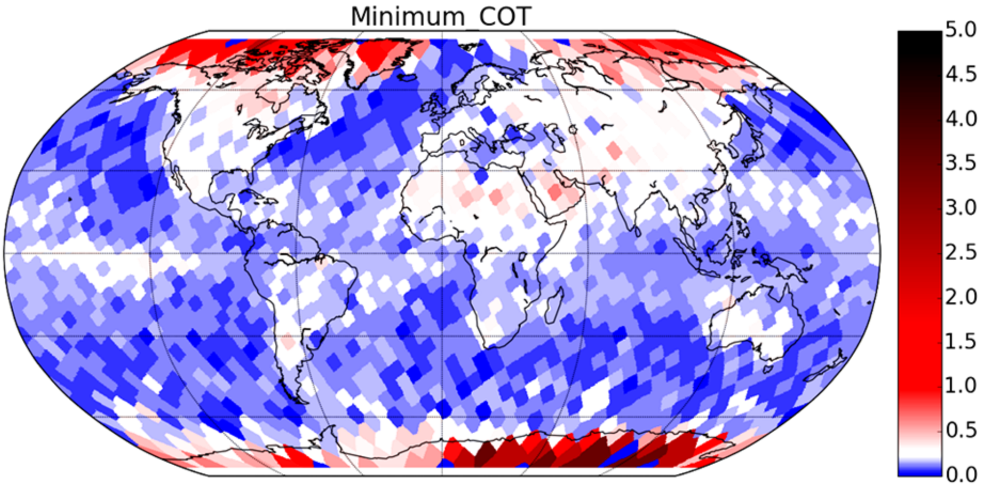
p9,l355-359: Add explanation why AVHRR gives higher cloud cover. It is easy to imagine a scenario where small cumulus clouds would be picked up by AVHRR (even if below its spatial resolution), but not by CALIOP products (for physical reasons). The statistical explanation given here does not seem to be complete and is hard to follow.

Reply:

We claim that the opposite situation (i.e., clouds picked up by CALIOP but not by AVHRR) is also very likely (see reply to reply to **general comment, part 4**.). Thus, it is not obvious that one can use this explanation to explain higher AVHRR cloud cover.

Again, it cannot be the objective or task of this paper to explain why we see the deviations we but rather to provide an as sensible and trustworthy validation result as possible. The reasons for the deviations have to be explained by those being responsible for the actual algorithms. We do give some suggestions but in the end this has to be verified by the algorithm developers.

However, the results in Figure 7 (new Figure 8) illustrate a more general problem. One has to remember that all results in Figures 6-10 (new Figures 7-11) are based on comparison with a CALIOP cloud mask filtered at cloud optical thickness 0.225 (the latter being the global mean cloud detection sensitivity). But regionally, the value of the cloud detection sensitivity varies a lot (see Figure 11 or new Figure 12). We have also changed the colour scheme in Figure 11 (new Figure 12) so that the relation to the mean value of 0.225 is made clearer:



In the new Figure 12 all values below the mean value 0.225 are plotted in blue colours and values above 0.225 in red colours. This colour representation also indicates better the highest values in the polar areas (not properly visualized in the old Figure 11).

Notice here that most oceanic areas are coloured blue and that the positive bias in Figure 7 (new Figure 8) is also mostly occurring over oceanic areas. By using a CALIOP cloud mask with cloud optical thickness being cut at 0.225 as our validation reference, we are then ignoring a substantial fraction of originally detected clouds below this cloud optical thickness limit over ocean surfaces. But these clouds are to a large extent actually detected in the CLARA-A2 results. Thus, it leads to an apparent (but largely false) overestimation of cloudiness over ocean in Figure 7 (new Figure 8 - notice that we are filtering CALIOP data and not CLARA-A2 data). We have added a discussion of this aspect in the Discussion section (lines 627-640).

This illustrates how difficult the estimation of general validation scores really is. More clearly, regardless of using a filtered CALIOP cloud mask or not when validating, there are always disadvantages. In that sense, the results expressed by the globally resolved cloud detection sensitivity is a much more objective visualization of the cloud detection performance provided that also a separate evaluation of false alarm rates are made. We repeat the following statement from section 3.4 (lines 295-300): "An important additional or complementary parameter in this context would be the false alarm rate in the unfiltered case (FARcloudy(tau=0)) since this parameter is not depending on any filtering of thin clouds". We have added this as an important result and recommendation in the Conclusions section (lines 408-414).

Finally, the most correct way of calculating and plotting the Bias in Figure 7 (new Figure 8) would have been to actually use the derived grid-resolved Cloud Detection Sensitivities in Figure 11 (new Figure 12) as representing the most appropriate CALIOP cloud mask (i.e., the filtered cloud optical thickness) for validation. We had this option in mind but we realized that this probably needs a much larger sample dataset in order to calculate stable statistics (since it requires calculation based on only those samples existing for every single grid point). Figure 12 (new Figure 13) illustrates that the available number of samples in each grid point is still rather small which makes the estimation of statistical parameters on this scale rather uncertain. But it can be considered for the future if the time series of CALIPSO collocations can be extended with several more years. The existing undersampling at grid point level (especially at low latitudes) is commented on lines 615-618 and on lines 637-640.

p10,369-371: What is Kuiper's score, what's the dominating mode in which case? At this point, some examples that help understanding one score vs. another are provided which is helpful, but that should be done (more systematically) earlier in the manuscript.

Reply:

See reply to comment for **p6** above.

p11: "The cloud detection sensitivity is here as high as 1.5"; "all optically thick clouds": : : Define what "high" and "thick" means (earlier in the manuscript).

Reply:

The cloud detection sensitivity is clearly defined earlier in the manuscript (lines 396-399). However, since it represents an optical depth we have generally replaced the word "high" with "large" (and 'small' for the opposite case) throughout the text.

p13, L495-500: Since specific orbits and satellites were not clarified, there's confusion here as to what was actually compared/validated. If it was equally applied to the morning and afternoon orbits (the wording leaves this open), one has to wonder how this would work because CALIOP operated in the afternoon orbit. How can morning cloud cover be "compared" to afternoon cloud cover, considering the significant diurnal cycle of clouds in most regions?

Reply:

We have clarified this in the revised manuscript. This study only dealt with comparisons with afternoon satellites. This was clearly stated on page 8 lines 306-308 and the reason for restricting it to afternoon satellites have been mentioned several times above in various replies to comments and questions. Still, we have discussed also the morning satellite case in the Discussion section (lines 697-710) since CLARA-A2 is based on both afternoon and morning satellite data.

The reason that we still brought up the case of morning satellites at page 13 is explained by the fact that readers of this paper (and reviewers!) would most likely start wondering how these results will relate to morning satellite data (representing almost 40 % of the entire CLARA-A2 data record). Matchups between CALIPSO and morning satellites are possible but only near the high latitude of 70 degrees on both hemispheres where the orbital tracks crosses between the two satellites (as explained on lines 220-224 in the revised manuscript).
What maybe confuses the Reviewer is that we state that some comparisons had been done also for morning satellites. However, this relates to the results in the standard CLARA-A2 validation report (of which some were presented in the CLARA-A2 paper by Karlsson et al., 2017) and not to this particular study. We have added a discussion about the morning satellite matchups in Section 5 (lines 697-710) where we have also clearly explained that the discussed results for morning satellites was published in earlier papers. But we wanted to mention these results in relation to this discussion and especially pointing out the need for addressing this issue later (lines 705-710).

The last question here is very relevant and interesting. The diurnal cycle of cloudiness is of course leading to differences which makes a direct comparison of results difficult (even at the latitude band around 70 degrees where we have matchups from both afternoon and morning satellites). Anyhow, we can first conclude that in the night part of an afternoon orbit and in the corresponding night part of morning orbits we have exactly the same type of AVHRR

measurements (same spectral channels used). Thus, the only additional difference expected might come from diurnal cycle effects which probably are quite small for the dark part of the day at these high latitudes.
The largest differences are instead expected in the illuminated part of the day since we will then use AVHRR channel 3a (at 1.6 microns) for the morning satellites while AVHRR channel 3b (at 3.7 microns) is still used for the afternoon satellites. The comment on line 497 stating that we have seen good agreement here means that even for the illuminated case we have good correspondence between afternoon and morning satellites. For the region covered by morning matchups we do not see large differences with corresponding results from afternoon satellites. This is encouraging and it indicates that the two spectral channels provide more or less the same cloud screening information (while for cloud property estimations, like optical thickness, we expect much larger differences). We also think that at these high latitudes we will probably not be very much affected by the diurnal cycle in cloudiness. A short comment on this has been added on line 702.

The statement about the good agreement were not meant to be general in the global sense but just saying that, where we can inter-compare the data from the two orbit constellations, the agreement appears to be good. Additional studies are however needed to evaluate the global performance of morning satellites and we clearly indicate a way forward here (by using CATS data – see lines 705-710 in the revised manuscript).

Language comments:

p1: “considerably” -> “considerable” Corrected

p1,l20: “were” -> “where”: multiple occurrences throughout the manuscript Comment: We found 2 occurrences of this error (typos) in the manuscript. The reviewer is too critical here, in our opinion. The typos have been corrected.

p1: “geographically higher” -> use “surface elevation” instead? Reformulated.

p1,l23-l25: run-on sentence (multiple occurrences); at the very least use punctuation (in this case, a comma after “CDRs”) to break it up. Better still, re-write. Reformulated (broken into two sentences)

p1: “sensor families”: a bit unusual for science manuscript, consider revising “family”

Comment: OK, we have removed this term (despite often being used in the remote sensing community)

p1: add comma before “which” (in most cases; multiple occurrences) Comment: We tried our best to improve here but it is not always obvious where to use a comma.

p1: “four decades: : :” -> “four decades, which qualifies them to be used in climate: : :” Corrected

p2,L51-53: “Linked to this: : :” Unclear: efforts by whom? stringent with regard to? Reformulated

p2: The A-Train stands for “Afternoon constellation”, not “Aqua Train” Corrected

p2: “a project being a part of” -> fix language Corrected
p2,l70-73: run-on sentence Corrected
p2,L91: MODIS: Introduce upon first occurrence Corrected
p3,l117: “including” -> “detecting” Corrected
p4, l121: “notice” -> “note” Corrected
p4,l148-165 and following: Avoid “you”: Not only is this inconsistent with the style of this manuscript, but it is also not advisable for a science manuscript in general. This sounds more like a seminar or talk than a paper at this point in the manuscript. I recommend a complete re-write of this section, as well as a thorough discussion of the meaning of a “cloud mask” (see comment above).

Reply: The use of “you” is removed and sentences are rephrased. Section three has been thoroughly redesigned (see reply to **general comment, part 4**, especially points 1 and 2 which affects this section.

p4,l148: “areal extension” -> “areal extent” Corrected
p5,l177-179: Revise English, hard to understand Rephrased
p5,l184: “possibly punish AVHRR-based methods in an unfortunate and undeserved way: : :” - see comment above Rephrased

p5,L199: “of which single shots that were removed: : :” fix English Rephrased

p8,L314: “navigation” -> “geolocation”? Corrected
p10: “Validation results are probably underestimated” -> What does that mean?

Reply: We have rephrased this sentence (lines 686-689) and also related this to the new Figure 1 explaining the matching geometry (see **general comment, part 4**, point 2).

p10: "compared to if only showing results based" -> fix English Rephrased

FINAL REMARKS:

- We are extremely grateful for the suggested editorial, syntax and language improvements. These are invaluable for non-native English writers like us!

- We also express our appreciation of the reviewer's large effort leading to this very detailed review.