Final reply to Referee 2'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 comments:

The paper presents an unprecedented evaluation of satellite-based cloud climatology (CMSAF's CLARA-A2) against CALIPSO/CALIOP performed at the global scale. Despite some limitations of CALIOP dataset discussed in the paper, it is the only currently considerable reference for cloud retrievals covering oceans, polar regions and other areas of very sparse cloud observations and measurements. Such evaluation has become possible with the sufficiently long CALIOP dataset. The authors also present an analysis of the CLARA-A2 cloud detection sensitivity, i.e. the threshold in the cloud optical thickness (COT) above which the cloud detection algorithm detects more than 50% of clouds. Screening the CALIOP data with COT below the globally-averaged detection sensitivity allows for "more realistic" evaluation, i.e. taking into account the difference between the sensitivity of CALIOP (active sensor) and AVHRR (passive sensor). Therefore, the paper will be an important first step towards proposing described validation methodology for the list of standard validation activities performed before releases of new cloud climate data records.

While the content of the paper is novel, valuable and appropriate for the publication in AMT, the paper structure should be significantly improved. Finally, the paper has some grammar and language issues, which should be addressed. They are mostly related to the syntax, i.e. sentence length and inappropriate word order. Some examples are indicated in the following, but the whole manuscript should be revised.

Reply:

We thank the reviewer for this positive evaluation. We notice the request for a reorganization of the paper (also demanded by other reviewers) and we have done our best to accomplish this. We reply to all specific comments below.

Repeating specific comment 1:

The title of the paper is a bit misleading. "Detailed characterization" suggests that the evaluation of the CDR is more detailed than the standard one, e.g. provided in CLARA-A2 validation report. However, the collocations of AVHRR and CALIOP are limited to NOAA-18 and NOAA -19, afternoon orbits and 10-year period only (from 30y+ of the CDR). Taking into account that one of the challenges in deriving CDR is stable performance in time, the evaluation presented in the manuscript cannot serve as an evaluation of CLARA-A2 CDR.

Reply:

Yes, we understand this remark and we agree that the validation presented here cannot be fully representative of a validation of the entire 34-year CLARA-A2 data record. But we still argue that the validation presented here is improved and more detailed than the validation (i.e., the CALIPSO-CALIOP part) presented in the CLARA-A2 validation report. The reason is the use of CALIPSO version 4 datasets (version 3 was used in the CLARA-A2 validation report) and the introduction of the new concept evaluating the cloud detection sensitivity which is the core topic of this paper. So we are quite confident that this is the best validation effort that can be done from existing reference data (lines 85-87), at least if requiring global coverage. The validation based on SYNOP data in the CLARA-A2 validation report indeed covers the full 34-year period but it cannot present a result that is globally valid in the same sense as the CALIPSO-CALIOP validation. We have emphasized this situation on lines 50-58.

As regards the collocations with NOAA-18 and NOAA-19, these are exactly the same as for the standard CLARA-A2 validation (i.e., same number of collocations, about 5000 orbits). However, in this study we exclude collocations with the morning orbits of NOAA-17, Metop-A and Metop-B since these are only possible over a narrow latitude band close to 70 degrees. Thus, we want to focus on the global performance and that can best be studied based on afternoon orbit data. The exact content of the entire validation dataset is now described in the new section 3.6.

The point about the necessity to evaluate the stability of a long-term data record is indeed an important aspect but also one of the most difficult ones to deal with. How can we find a suitable reference dataset of cloud observations with global coverage to perform this stability analysis? To be honest, there is no such reference dataset offering the required length and coverage of observations. The only candidate is surface (SYNOP) observations of cloudiness but they cannot fulfill the requirement of global coverage (e.g. oceanic and polar regions are largely not covered) as mentioned on lines 50-58. They also have their own quality problems (e.g., lack of knowledge of the thinnest cloud being observed, low quality at night-time and also hampered by being subjective in their character in that different observers have different opinions on how to interpret clouds and their coverage). Furthermore, the surface observation network has undergone rapid changes during the last decades due to automatization and this has caused problems in maintaining stable observation quality over time. With this background, we are of the opinion that there is no better reference than the 10-year CALIPSO dataset for evaluating the CLARA-A2 (and similar) satellite-derived data records, despite the fact that it only covers about one third of the CLARA-A2 observation period. It offers the global coverage (only excluding some areas in close proximity to the poles) and a high and stable quality of observations. Estimating the stability is still a challenge but we hope that on a longer term also this aspect will be properly dealt with assuming that the era of active cloud lidar observations from space can continue (e.g., with new data from EarthCARE and CATS replacing CALIPSO and hopefully also data from new lidar missions beyond the lifetime of EarthCARE). This aspect is mentioned at the end on lines 801-809.

Finally, we have also changed the title to the following:

“Characterization of AVHRR global cloud detection sensitivity based on CALIPSO-CALIOP cloud optical thickness information: Demonstration of results based on the CM SAF CLARA-A2 climate data record”

Repeating specific comment 2:

Objectives of the study should be described better in the Introduction. In relation to (1), it should be clear if the aim is to present new methodology using a subset of CLARA-A2 as an example or to evaluate CLARA-A2.

Reply:

Yes, we have done that on lines 76-89 (see also the reply to 1). The new title also emphasizes the presentation of a new methodology more than the presentation of new CLARA-A2 validation results.

The study intends to provide revised or upgraded validation results (compared to the validation reports from the standard CLARA-A2 validation) with some extended or additional features (like the Cloud Detection Sensitivity). The revision is partly required by the upgrade of the available CALIPSO-CALIOP datasets and the results of the impact of this change are also included as one separate (or preparatory) objective of the study (described in sections 3.3 and 4.1).

Repeating specific comment 3:

The current discussion section is a mix of discussion remarks and conclusions. I recommend to separate the two. In the results' section, there are also interpretations, which are hypothetical (they often start with "we believe", "we claim") and should be moved to the discussion. Otherwise it is often difficult to judge which statements are really supported by the results achieved in this study.

Reply:

Yes, we admit this weakness of the current manuscript. We have followed the recommendation and included both a Discussion section (section 5) and a Conclusion section (section 6).

Repeating specific comment 4:

The analysis of detection sensitivity reveals some interesting non-expected results. One is that CLARA performance is not better at dark and warm ocean surfaces (L374-375). The hypothesis this is due to sampling and geometry of AVHRR and CALIOP FOVs needs more explanation. The problem was detected here, because it leads to unexpected results. However, how to measure a possible effect of this issue on results in other situations, regions, etc.? I would consider a separate section (or paragraph) in the discussion..

Reply:

Yes, we admit that this result deserves more attention. We also got a similar remark from the other reviewers. We have improved the description in three ways:

1. We introduced a short summary (first part of Section 3.2) of the underlying basic method of matching AVHRR and CALIPSO data. It seems the current referencing to the original paper by Karlsson and Johansson (2013) (which introduces 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 an illustration (new Figure 1) of how matched high-resolution AVHRR FOVs relate to the CALIPSO-CALIOP FOVs within a nominal AVHRR GAC pixel. The consequences for the matching of the two datasets are described in the second part of Section 3.2.

3. We expanded the discussion of these results in the new Discussion section (Section 5, lines 642-695). However, we believe that further studies on the full (global and local) impact of the differences of matched AVHRR and CALIOP FOVs could indeed deserve a paper on its own. Thus, we cannot dwell too much on this seemingly unexpected result since this would risk leading to a much too long paper. We only want to highlight the existence of this problem which has (in our view) been largely overlooked in many previous papers using CALIPSO-CALIOP data as the main validation source.

Repeating specific comment 5:

Is the cloud detection sensitivity a measure of CDR performance itself? There is no discussion if 0.225 signifies good or bad CLARA performance. One can imagine the same analysis (i.e. evaluation against screened CALIOP data), but with the estimated cloud detection sensitivity of, say, 0.5. Please elaborate on that. In addition, since the authors recommend the methodology to be widely used (e.g. in CFMIP), more detailed guidelines would be appreciated. For instance, when applied to different passive-sensor-based CDRs, should the cloud detection sensitivity be always recalculated?

Reply:

Yes, even if it only concerns cloud detection performance, we believe that it is at least one very important piece of information for characterizing the entire CDR performance. Despite of the fact that it only deals with the cloud masking quality and not specifically with the quality of other parameters of CLARA-A2 (e.g. other cloud properties, surface albedo and surface radiation budget parameters), we also know that errors in cloud masking definitely will affect the

quality of other parameters derived further down-stream in the processing of a data record. For example, incorrect cloud screening (missed clouds) over dark surfaces will inevitably lead to an overestimation of surface albedos. Exactly how the uncertainty in cloud masking is propagating into the uncertainty of other parameters is yet to be determined in more details than what is done today. However, to better describe this is one of the challenges in the CM SAF project when preparing the next version of the CLARA dataset (CLARA-A3). But for the current CLARA-A2 dataset (and which could also relevant for other similar type of datasets), this new description of the cloud detection performance can be seen as one important step towards a better uncertainty description.

The question whether the average cloud detection sensitivity at (cloud optical thickness) 0.225 represents a good or a bad performance has no clear answer. This is because this study is the first of its kind proposing such a measure defined in exactly this way (as described in the paper). However, one indication that it is probably not too bad is that the COSP (Cloud Feedback Model Intercomparison Project (CFMIP) Observation Simulator Package) satellite simulator for ISCCP uses a global cloud optical depth threshold of 0.3 to describe the cloud detection ability of the ISCCP dataset.
However, this quantity can only be evaluated when and if it is later put in relation to corresponding values (computed in the same way) for other datasets (like datasets from MODIS Collection 6, PATMOS-X, ISCCP or ESA-CLOUD-CCI). We encourage such studies since we think that this measure of performance is a universal one which has nothing to do with AVHRR data in particular. Instead, it should be applicable to any other global cloud dataset based on passive satellite imagery. And, yes, it should always be recalculated for every new dataset to be evaluated (answer to last question). These cloud detection sensitivities could then be inter-compared between different data records. This is the main point in promoting this method as a universal method.

The value 0.225 is only a global average calculated for CLARA-A2 (or to be strictly correct, for the 2006-2015 period of CLARA-A2) and it should only be inter-compared and evaluated with corresponding global averages derived for other cloud datasets. In that sense, the question about what happens if using the value 0.5 is not relevant. More interesting would rather be to compare the results of the global distribution of the cloud detection sensitivity (new Figure 12) with corresponding distributions for other cloud datasets. This would be the most interesting aspect for use in a wider context since this would be able to reveal global differences (at a rather fine resolution) in performance for different algorithms and data records. Examples of such inter-comparisons are still rather few (with the GEWEX inter-comparison study by Stubenrauch et al. in BAMS July 2013 as the best example). A tentative repeated GEWEX inter-comparison study in the future could be imagined to include such global performance and

difference maps valid for the entire period of CALIPSO data. That would really show how all these data records perform if using CALIPSO-CALIOP as representing the truth.

We have included some of these clarifications and proposals/suggestions in the new Discussion and Conclusion sections (e.g., lines 627-640, lines 787-799 etc).

Reply to short comments and editorial remarks:

L50, "be very accurate to be able.." - please be more specific, e.g. referring to GCOS recommendations

Reply: We are of the opinion that the reference Ohring et al. (2004) explains exactly what "very accurate" means. Their discussion also involves references to GCOS recommendations. We don't want to expand the discussion further here, especially when considering the need to expand other sections as a consequence of other more serious requests from reviewers.

L82, "FOV resolution" - field of view does not have a resolution, I would keep FOV and remove 'resolution' (or 'size' in other places in the manuscript)

Reply: OK, we may have used the wrong terminology here. The field of view (or sometimes being denoted "Instantaneous Field of View) can be defined as "***The area on the ground that is viewed by the instrument from a given altitude at any time.***" So, yes, this area is not equivalent to a resolution. The resolution we are thinking of is rather linked to the diameter of the FOV (assumed to be circular or elliptic in shape). This diameter, in turn, is then often used as the resolution of the image grid or image matrix defining a satellite image. In that sense, there is often some sort of relation between the FOV (diameter) and an image resolution.
However, to just remove resolution (or size) does not solve the problem here. For example, the sentence
"*AVHRR is measuring in five spectral channels (two visible and three infrared channels) with an original horizontal field of view (FOV) resolution at nadir of 1.1 km.*" cannot be written as
"*AVHRR is measuring in five spectral channels (two visible and three infrared channels) with an original horizontal field of view (FOV) at nadir of 1.1 km*".

From the definition, FOV is an area and the modified sentence is therefore still wrong.

We propose kind of a compromise here so that we do not have to change too much of the text. We propose to use the expression "FOV size" to denote the approximate diameter of the FOV area. This requires that we explain this interpretation the first time we use it. Thus, we have added the following lines 99-100 after the introduction of AVHRR measurements:

"The size is defined in this context as the approximate diameter (assuming a circular or elliptic shape) of the FOV and this definition will be used throughout this paper."

We hope that this explanation will be enough for the reader to understand when we talk about the different FOV sizes (e.g., 70 m, 330 m, 1 km and 5 km) in the remainder of the paper.

L92, 'various parameters retrieval' - be more precise

Reply: CLARA-A2 contains more than just cloud parameters. There are also surface radiation and surface albedo products. The description is expanded slightly to explain this (lines 113-121).

L117-119, "Thus CALIOP products..." - please provide a reference for this statement

Reply: This is also described in the earlier mentioned reference Vaughan et al., (2009). Thus, we repeat it here (line 141).

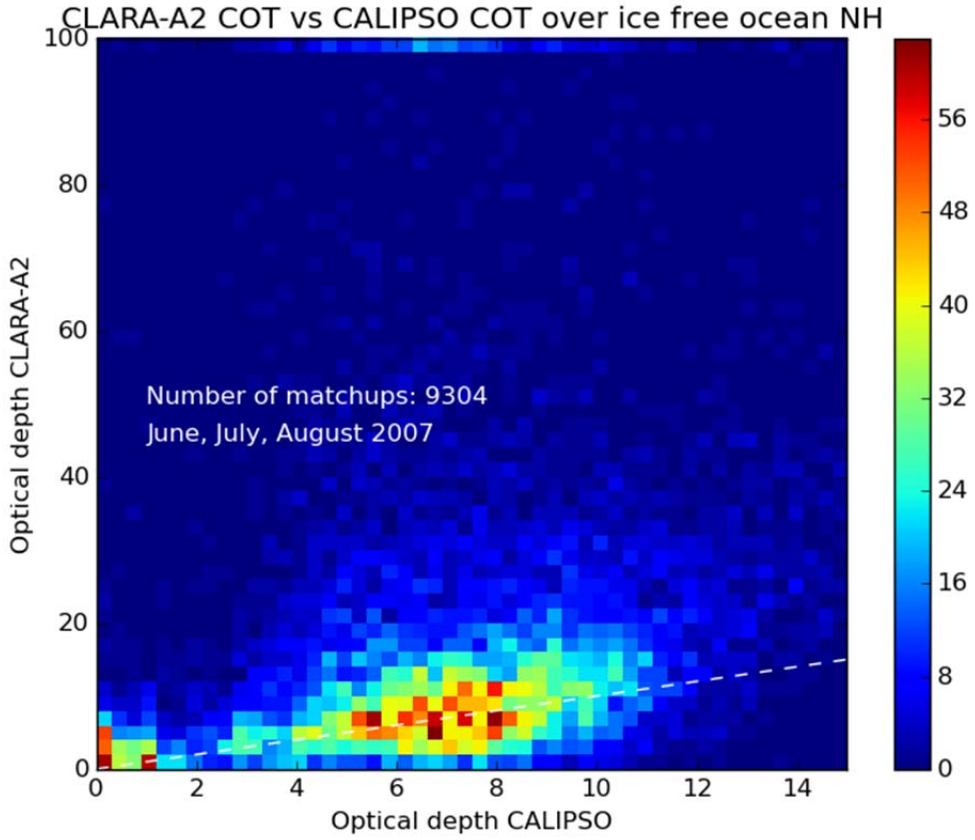
L126-127, "...claiming that useful...seems to be available" - based on which results?

Reply: We also got a question on this from another reviewer. We repeat the reply to that question below:

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

L140-L145, If these improvements are relevant for the study, please explain them better

Reply: We are of the opinion that the three selected changes are obviously important for this study and that no further comments are needed. Full information about all changes is given by the link given before on line 138. Here we only highlight three selected changes which we think are most important.

The first selected change (line 170) points at a general improvement of the fundamental cloud-aerosol-discrimination method. This method is, of course, crucial for the quality of CALIOP cloud information.

The second selected change (line 171) points at a special problem that previously was noted for cloud-aerosol discrimination over certain regions. This is also crucial for our validation study since it reduces the risks that regional features in our validation results are due to weaknesses of the underlying CALIOP data.

The third change (line 173) is important in that it offers an alternative method to take into account some of the inconsistencies between fine resolution and low resolution CALIOP datasets. This is discussed more in detail in Section 3.2 (lines 250-280) and in Section 3.3.

Thus, we keep the text as it is. In our opinion, to add extended text is more important for more serious review points.

L150, "..how thin or thick..." - do you mean optically, in height?

Reply: We mean optically thin or thick. We have added this for clarity on line 134 and on several other places in the manuscript.

L151, "The second aspect..." - something is wrong with the syntax, please rephrase

Reply: We have rephrased the text considerably (lines 177-185).

L192, The investigation if the method used by Karlsson and Johansson (2013) is still applicable to the new CLAY version should be listed as one of the paper objectives (i.e. already in the introduction). The results (L206-223) should be moved from this paragraph to the Section 4.

Reply: Yes, we agree. We made the following changes:

1. A short sentence on the upgrade to CALIPSO-CALIOP version 4 and the impact of this change is added to the Introduction (lines 82-83).
2. We added a sentence (lines 297-298) explaining that the results of the preparatory study are given in (new) section 4.1.
3. The current description of results of the preparatory study is moved to (new) Section 4.1.

L249, why 'CLARA-A2 cloud masks', i.e. in plural?

Reply: Rephrased as follows (line 348-349):

"The results are computed by treating both CLARA-A2 and CALIOP cloud masks as binary values, ….."

L250, "This approximation is acceptable.." - provide a reference

Reply: Well, the simple answer is that there is no estimation of sub-pixel cloudiness in the CLARA-A2 case. Thus, we actually have no other choice. We have removed this sentence to avoid any confusion.

L288, Why 50% is an appropriate threshold for the cloud detection probability?

Reply: We do discuss this in the text (in the sub-sequent sentences after L288, which are lines 395-404 in the revised manuscript). The argument is that above this threshold, by definition we detect more clouds than we miss (in the statistical sense). A cloud detection scheme that misses more clouds than it detects is not an efficient scheme. So, a minimum requirement should be that it at least should detect 50 %. This is our point. If this is not a satisfying answer we wonder: How would you otherwise describe or define a measure or the cloud detection sensitivity? A threshold anywhere below the 50 % level can be questioned since the scheme then would generally fail here by missing more clouds than it detects. So, in our opinion, the 50 % level is the most sensible choice.

L326, "..but we still believe..." - what if the authors are wrong?

Reply: It is difficult to answer this question. In the ideal world you would always have an infinite number of samples to make the perfect statistical estimation. But in reality there are always limitations. The best thing to do here is probably to remove this rather speculative sentence and instead highlight that there might still be locations where estimations are uncertain. We reformulate the sentence in the following way (lines 432-433):

"…with only a few exceptions mainly located over the Pacific Ocean. In these locations the uncertainty in the results might be expected to be larger than for the rest of the globe."

L328 and L349, Please consider giving different section names. These two are not very informative.

Reply: OK, we suggest the following:

4.2 Results based on original CALIOP cloud masks compared to results excluding contributions from very thin clouds

4.3 Additional validation scores

L369, "This contributes..." - it's not clear what is meant. Please rephrase.

Reply: We suggest the following (line 644-645, also adjusting to new Figure numbers to reflect the new Figure 1):

“This explains to a large extent the fairly low values of the Kuipers’ score over these regions (Figure 10) leading to a slightly different distribution of results in comparison to the Hitrate (Fig. 7).”

L361-404 – It would be easier to follow the text divided in paragraphs

Reply: OK, we have sub-divided the text into several paragraphs.

L381, “We first conclude...” - is it based on actual results or it is a hypothesis?

Reply: This follows from the actual geometries of the matched AVHRR GAC and CALIOP FOVs. We have commented this further in relation to discussion of the additional figure demonstrating the matching geometry (see point 2 in the reply to **specific comment 4**).

L406-407, Wrong syntax, please rephrase

Reply: Rephrased sentence (lines 577-579):

“We have here presented validation results after having ‘removed’ (in the sense of interpreting them as cloud-free cases) all clouds with smaller optical depths than the cloud detection sensitivity parameter. This leads undoubtedly to a clear improvement of results compared to if only showing results based on the original CALIOP cloud mask (i.e., comparing Figs. 5 and 7).”

L407, “...is undoubtedly a clear improvement”, please explain why?

Reply: We think this is rather obvious when comparing results from the unfiltered (old Figure 4, new Figure 5) and the filtered case (old Figure 6, new Figure 7). Hitrates are considerably higher which is emphasized in section 4.3. The problem with the unfiltered case is highlighted in lines 332-334 in the original manuscript. Since CALIOP is a much more sensitive sensor than AVHRR there should be a certain fraction of clouds that are detectable by CALIOP but which never will be detected by any AVHRR-based method. The filtering approach is one way of trying to compensate for this.
We think we can rely on the current text and discussion here. No changes are made.

L436-438, Please explain better, preferably in a separate paragraph in the Discussion

Reply: We have done that (please see point 3 in the reply to **specific comment 4**).

Figure 11, it would be useful to have a different color scale (e.g. as in previous figures), with a shift between colours at 0.225. Otherwise it is difficult to see the 'edge' at 0.225

Reply: We definitely agree. This was one of the changes we had planned even before achieving review comments to the discussion paper. Here is our proposed new Figure 1 with blue colours denoting places where the detection sensitivity value is lower than the average value of 0.225 and where red colours show places where values are higher than the average. This new plot is also better in showing the high values over the poles (which were just masked out in grey colours in the previous figure).

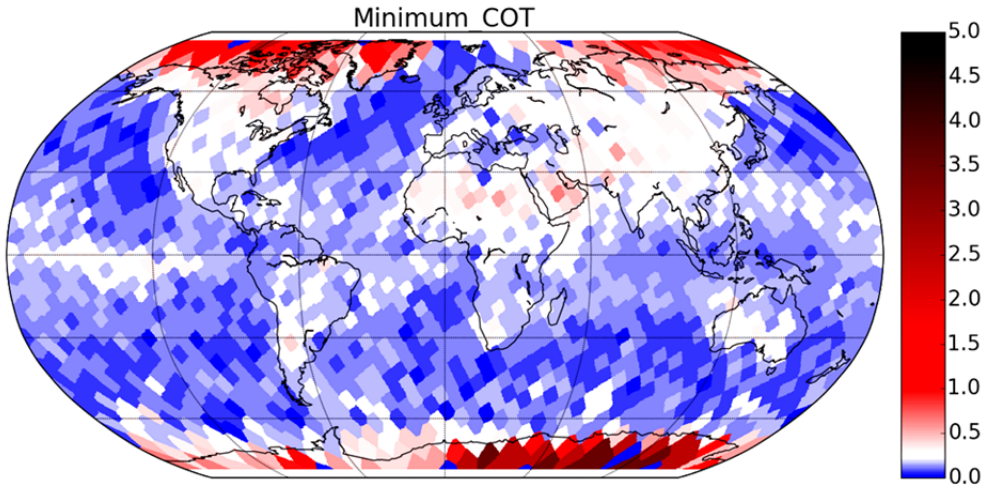


Figure 12, it would be useful to add FAR or KSS here. POD alone does not reveal the true performance of the cloud detection, as it gives no information about false alarms.

Reply: In principle we agree with this opinion but we have also argued in the text that this figure is really resulting from the stretching of our results to the very limit of what can be safely presented. This is because we have a limited number of available samples for individual grid points, especially at low latitudes. More clearly, the "true" POD curve is theoretically expected to show a continuous increase with increasing cloud layer optical thickness (if having

access to an unlimited number of samples). Thus, the variation we see here with some unexpected oscillation (e.g. near COT=0.8 for the Sahel curve) is a clear sign of that we still need more samples to make a very confident estimation of these POD curves. In that sense, this figure serves more like an appetizer for what we can do in the future with an even more extended CALIOP dataset (hoping for a long CALIOP lifetime). Still, the curves illustrate very well how the probability of detection of cloud layers varies for different geographical locations. So, despite of limitations, these are unprecedented results that we for the first time are capable of (almost realistically) depicting. In conclusion, we stick to the visualization of exclusively the POD variable for individual grid points. This should be seen more as a feasibility demonstration of what can be achieved in the future when having access to a much larger AVHRR-CALIOP matchup dataset.
We have added some further arguments and discussion on these under-sampling aspects (lines 615-618 and 635-640).

Technical (editorial) corrections:

many times in the manuscript, use a lower case after using a colon in the sentence **Checked and corrected**
L11, should be “sensitivity of the detection” **Corrected**
L14, results of? Please rephrase. **Done**
L16, “portions” looks weird in this context **Replaced with “parts”.**
L23, use elevation or altitude instead of “highest” **Corrected**
L66, remove “ **Done**
L132, 70 N/S **Corrected**
L200, remove second “be” **Done**
L230, should be ‘where’ not ‘were’ **Corrected**
L237, give colon after ‘namely’ **Done**
L317, “...a minimum of the number or matchups” should be “a minimum number of matchups” **Corrected**
L371, should be “Kuipers score” **Corrected**
L570, incorrect order of references **Corrected**