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.

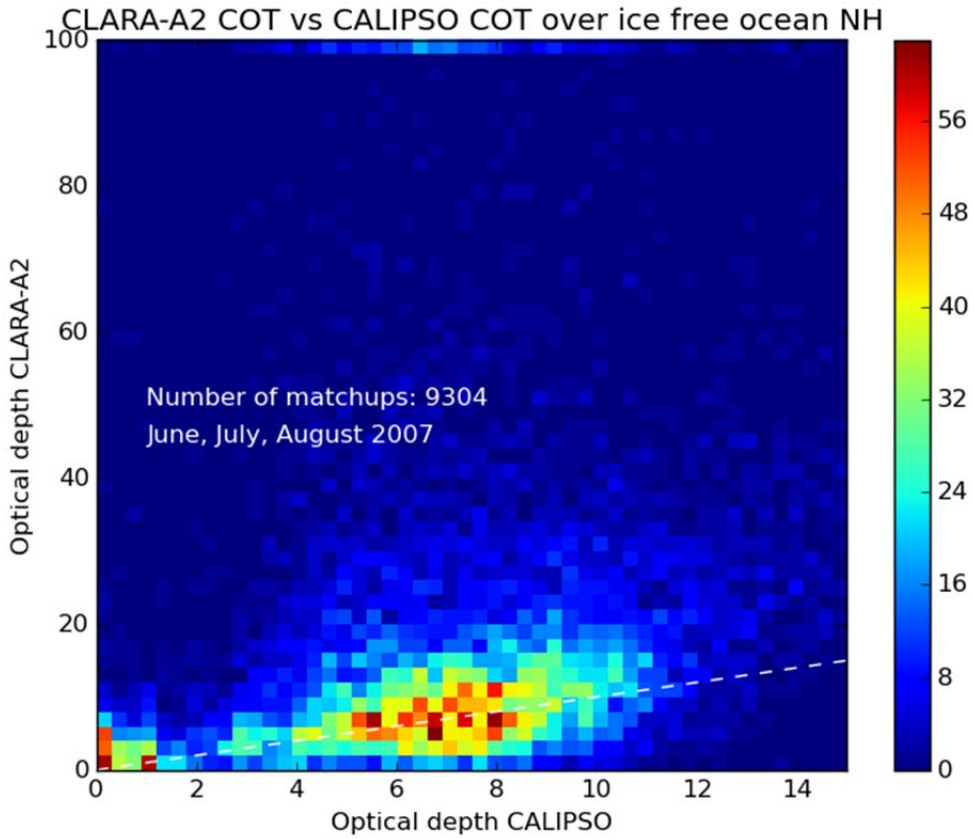
p5, l187: The optical thickness threshold of 5 for CALIPSO is higher than usually assumed. If it is necessary for this study to work with such a high threshold, it should be justified, and it should be explained how this is possible (referring to literature where this has been done, or with a dedicated 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.

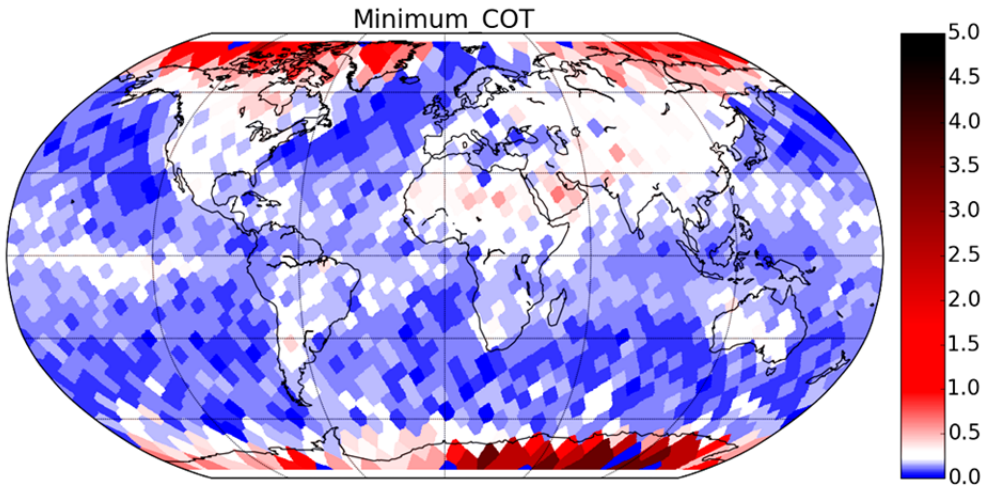
Reply:

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.

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).

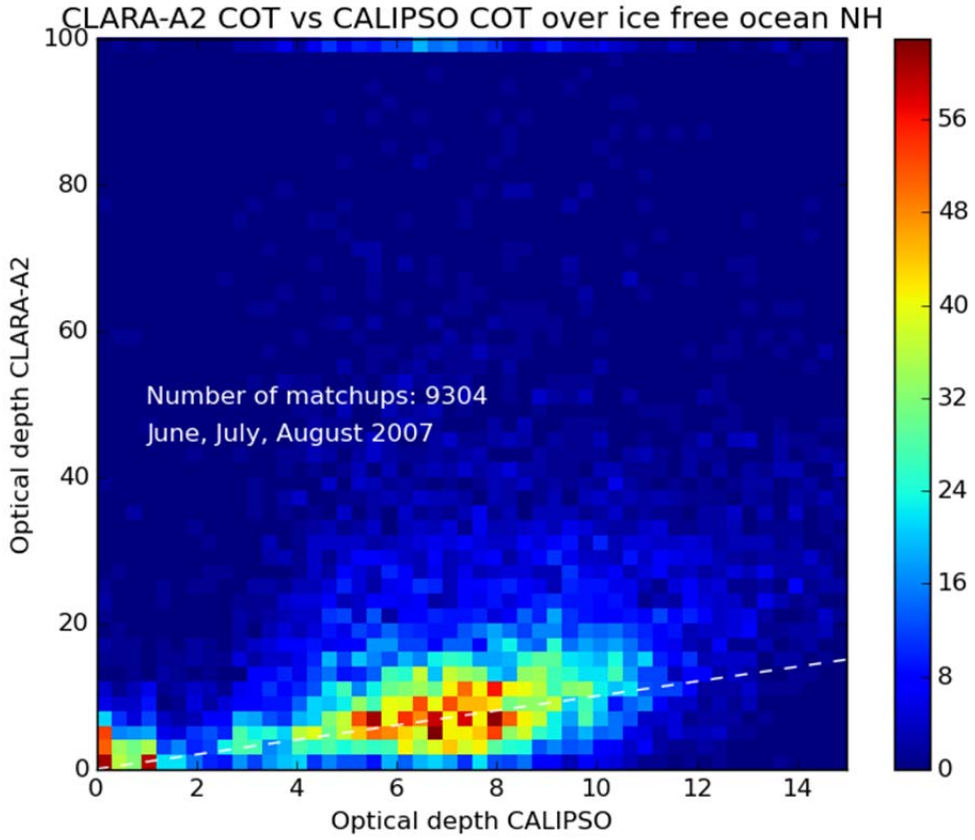
L126-127, "...claiming that useful...seems to be available" - based on which results?

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).

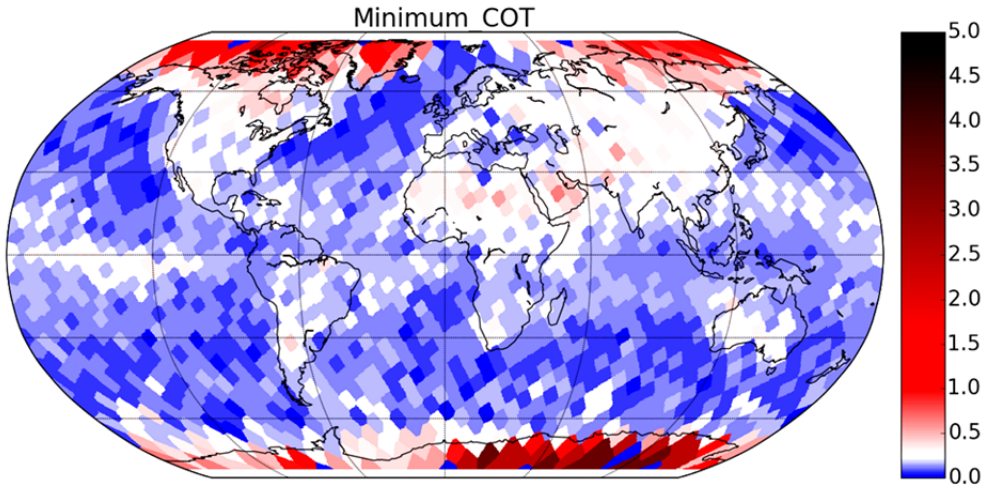


Figure 12, it would be useful to add FAR or KSS here. POD alone does not reveal the true performance of the cloud detection, as it gives no information about false alarms.

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**

Final reply to Hartwig Deneke'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 manuscript provides an in-depth investigation of the cloud detection performance of the algorithm employed in the CLARA climate data record, utilizing CALIOP lidar observation as reference. The topic of the paper is interesting, presents novel results, and the approach is scientifically sound, hence I do recommend the paper for publication in AMT.
There are however a number of general comments/concerns which I'd like to see addressed/at least discussed in the manuscript before publication, which will further clarify the relevance of the results for readers. I also added a number of specific minor points/language corrections below, which is likely incomplete. I do recommend proofreading of the manuscript by a native English speaker.

Reply: Thanks for this positive evaluation. We will address all points in the following. The final manuscript has been checked by a native English speaker.

General comment 1:
- Title: "Detailed characterisation" => from my point of view, the term "characterisation" mainly refers to a characterisation of performance in terms of CALIOP cloud optical thickness, I'd recommend adding COT to the title (e.g. "based on CALIPSO-CALIOP cloud optical thickness"), this is more specific than "cloud information" (what other information do you use?). I would also prefer the term "sensitivity" over "Performance",
but that is definitely a matter of taste. Hence please consider modifying the title, taking these points into account.

Reply: We got similar remarks from other reviewers. We have changed the title as follows:

"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"

General comment 2:

- a) The authors should describe in more detail the cloud detection scheme and the changes between the CLARA-A1 and A2 data records, in particular with respect to cloud masking. The short paragraphs at the end of Section 2.1. seem somewhat too brief, considering that the aim of the paper is to characterize the performance of that scheme, and the findings might be different for other cloud screening methods. Has the cloud mask algorithm been changed/improved between the two versions of CLARA?

b) Are changes in cloud detection performance expected, is it possible to quantify such changes using the validation approach?

c) Do the calibration updates affect the cloud mask performance?

d) Has the analysis of Karlsson et al.,2013, helped to improve the algorithm, i.e. have you been able to tune the algorithm based on the results of the previous validation study?

e) Do you expect that your results are specific to this cloud masking method, or do you expect them to be linked to fundamental characteristics of the AVHRR observations you are using, so your findings would apply similarly to other AVHRR-based cloud detection algorithms? If the latter, how would this translate to other sensors as e.g. MODIS/SUOMI NPP/geostationary observations?

Reply:

a) We disagree here in the sense that the CLARA-A2 paper by Karlsson et al., (2017) does exactly what is asked for here, i.e., it explains what has been done to algorithms (not only cloud retrievals) and calibration methods for the upgrade to the CLARA-A2 data record. We cannot repeat this here considering the length of the paper and the need to dwell deeper on other more serious subjects brought up by reviewers. However, we added a statement making it more clear where descriptions of algorithm changes can be found (lines 126-129).

b) Definitely. The paper by Karlsson et al. (2017) gives already some validation results (e.g. comparisons with MODIS Collection 6 results in Figure 6d in that paper). It also refers to the weaknesses of the CLARA-A1 cloud detection which largely have been solved by the new methods in CLARA-A2. However, the purpose of this paper is not to evaluate the improvement in the cloud detection algorithm from CLARA-A1 to CLARA-A2. Rather it introduces a method for a more detailed characterization of cloud detection sensitivity.

c) Yes. The cloud screening methods use fixed or pre-calculated thresholds which mean that if calibration drifts (i.e., visible reflectances changes) cloud detection results will also change. However, the used cloud detection scheme uses thresholds in the short-wave infrared and infrared regions with a higher priority than the visible thresholds. In that sense the sensitivity to visible thresholds is small (but not negligible).

d) Absolutely! It helped in finding the largest weaknesses of the cloud screening algorithm (e.g. the problems found over semi-arid regions) and the validation method has been heavily used to evaluate the impact of subsequent and final algorithm changes. We consider it as maybe the most important tool in the development work. But, of course, the CALIOP data itself (i.e., the access to almost one full decade of CALIOP data) is the most important aspect here.

e) Of course, these presented results are specific to the cloud screening method used for CLARA-A2. However, we believe that the evaluation method itself is universal and not specifically linked to AVHRR data or AVHRR-based methods. We state this very clearly in the Conclusions section on lines 720-724 and on lines 765-772. All satellite observations/retrievals which can be matched/collocated with CALIOP data can be evaluated in the same way. We think it is a strong point to suggest the use of one such universal method for determining the cloud detection sensitivity. It can facilitate how to inter-compare results from different methods and different satellite sensors.

Regarding the mentioned sensors (MODIS/SUOMI NPP/geostationary) we see no particular problem in trying to repeat the same kind of study. In fact, we are planning to do it ourselves in the near future, with the highest priority on evaluating measurements from the Suomi-NPP and NOAA-20 VIIRS sensors.

General comment 3:

-In general, I find the approach of looking at the COT regardless of observing conditions somewhat too simple. I expect the detection performance to be very different during daylight/nighttime conditions, and also depend on cloud type/phase (viewing angle might be another important influencing factor). Additionally, the cloud detection scheme relies on a combination of tests, which will show different sensitivities to thin/thick/low/high clouds (it might be interesting to look at the sensitivity for each individual test separately). While it is nice to quantify the geographic variation of detection performance, what are the dominating factors for those variations (I guess surface albedo, cloud type?). Here, I urge the authors to discuss their results with more focus on the underlying physical effects (suggested plot: using a global surface albedo map e.g. from MODIS, show an x-y plot of threshold COT vs. surface albedo), and at least discuss if considering day/night different cloud types separately would add new insights.

Reply: We definitely agree with the reviewer here regarding the potential for deeper and more detailed studies. But we have to stress (which is mentioned several times in the paper, e.g. on lines 637-640), that for doing this we need to have a more extensive dataset. Already with the present dataset we have identified problems in getting enough of samples to get statistically reliable results at the individual gridpoint level (here, we use 300 km resolution grid points). See for example the discussion about the results of Figure 13 in the revised manuscript (lines 615-618). The sparseness of data is mostly found at low latitudes which can be explained by the way samples are collected and the used polar orbits. To further sub-divide our dataset, e.g., into daytime and night-time portions, will probably lead to extended areas with lack of collocations.

Furthermore, we don't think it is really our job to explain why we have these validation results in terms of the cloud screening algorithm details. This is up to the development team of each investigated algorithm to discuss and understand. This study is mainly a validation study which may highlight algorithm weaknesses but it can neither explain the weaknesses nor provide solutions to overcome them.

In conclusion: More detailed studies may come later after receiving a longer time period of data and possibly if using less stringent matching criteria (i.e., allowing a temporal difference of 10 minutes instead of 3 minutes). But here, we prefer to stay with the current approach of making a first attempt to derive global results as a demonstration of the potential and only give a few examples of more local results (Figure 13).

General comment 4:

-Due to GAC sampling, the comparability of CALIOP and AVHRR observations likely suffers. Can you quantify this effect using spatially complete data, e.g. by use of MODIS data to simulate GAC sub-sampling, in particular for those regions where clouds with significant small-scale variability are expected (i.e. the sub-tropical ocean). Even an analysis on limited data might shed some more insights in the context of the rather speculative disuccsion on page 10 ("We believe"...).

Reply: We got similar questions from the other reviewers. We concluded that we need to improve our description and discussion of the matching methodology and better illustrate the geometrical aspects and consequences of matching the AVHRR GAC and CALIOP FOV observations. We have done that in three ways:

1. We introduced a short summary of the underlying basic method of how we matched AVHRR and CALIPSO data (first part of 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 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). Thus, the current Discussion section will be split into one separate Discussion section (Section 5) and one final Conclusion section (Section 6). The problem of inter-comparing CALIOP data with other satellite data in cases of highly scattered and fractioned cloudiness needs to be discussed. In our opinion this aspect has been largely overlooked in many previous papers using CALIPSO-CALIOP data as the main validation source.

General comment 5:

-In the conclusions, the author's stress that long-term availability of active observations from space would be benefical in the conclusions. While I generally support this point, due to the inherent value of active observations, I am not convinced that this indeed adds value to the aims of this paper. Do the authors expect the performance of the cloud mask to change over time? If so, what factors could change? Why is not a once-only characterization sufficient?

Reply: Yes, in principle a once-only characterization is probably OK for an individual data record like CLARA-A2. But for its evolution over time (i.e., upcoming new versions of CLARA, like the currently planned CLARA-A3 to be released in 2021-2022) there is a need for new evaluations. Especially, future versions of CLARA will have to be transformed into an AVHRR-heritage type of data record since the AVHRR instrument itself will soon be missing on upcoming satellites. The last AVHRR will be launched on METOP-C (scheduled for 2019) which effectively means that no AVHRR measurements can be expected beyond the 2025-2030 time frames. However, AVHRR-heritage datasets are still possible if utilizing AVHRR-like spectral channels on other sensors, e.g. the VIIRS sensor of the JPSS satellites. But to evaluate and get a smooth transition of the data record in this way we need to repeat studies like this with the existing data from active (lidar) measurements. We have added a comment on this (lines 806-809).

However, there is also a very important aspect in that we currently lack good reference data to estimate the stability of data records (mentioned on lines 804-806). An extension of missions with active lidar instruments in space will eventually allow more accurate estimations of the data records stability over time.

General comment 5:

-Finally, I do think that the language/wording of the article can be significantly improved, both in terms of English language use and in terms of being stricter/more consistent in terminology (some examples: use of terms "parameters" vs. "scores", "performance" vs. "sensitivity", "cloud screening" vs. "cloud detection" vs. "cloud masking", using the abstract term "detection sensitivity" instead of COT). Please do revise the paper once more carefully with respect to this points.

Reply: Certainly, we are aware of language limitations and mistakes in the manuscript. We have taken these aspects into account and also in the end we

used native English speaking people for a final check of the manuscript. We are grateful for all language comments and suggestions in the following.

Detailed/language comments (disclaimer: I am not a native speaker myself...):

-L10 : "including their global distribution" => "regional variation"(?) (results is unspecific,so it remains unclear what a "distribution" of results actually refers to) Rephrased (lines 13-15)

-L11 "sensitivity of the results" => which results? This opens up the possiblity for misunderstanding, please change "the results" to "the cloud detection performance" or name the statistical score you are referring to.
Rephrased (lines 16-17)

-L 11: "cloud optical thicknesses" => "thickness" Corrected (line 19)

-L 21: "sensitivities : : : were larger than 0.2" => please make it clear that COT is used as measure for sensitivity, and hence 0.2 is value of COT!
The quantity "cloud detection sensitivity" is clearly defined in the text (lines 16-17) as a COT value. No change.

-L22 "over Sahara" => "over the Sahara" Corrected.

-L23-L24: "The validation method', "validation results are proposed". This is fairly unspecific. Why not mention exlicitely "It is suggested to also quantify the detection performance of other CDRs in terms of a sensitivity threshold of cloud optical thickness which can be estimated using active lidar observations" Adopted.

-L28: "appear increasingly important", do not use "appear", or do the author's doubt the value of their own work?
"appear" is replaced with "are".

-L29: "cloud description and : : : feedback processes" => suggested re-phrasing "the parametrization of cloud processes and cloud-aerosol interactions including related climate feedbacks." Adopted.

-L37: I suggest to drop the part "in combination with ...", I do think satellite observations have sufficient value even without complementary ground-based observations **Adopted.**

-L41: "the global view" => "their global coverage" **Corrected.**

-L57: "Aqua train" => I have never heard this term, all references I can come up with translate A-Train to "Afternoon train"
Corrected (lines 67-68).

-L162: "A very strict definition" => I do not think this is a definition, but a characterization (this point also applies to other similar uses later in the manuscript) **Rephrased (lines 197-198).**

-L235: "behave in a strange way" => maybe "introduce distortions" **Adopted (lines 332).**

-L341/342: places=> regions/locations **Corrected.**

-L442: performance parameters => be more consistent in terminology, do you mean skill scores, or the threshold in COT? **Rephrased (line 714).**

-L448: "The method : : : is not : : : valid for the CLARA-2 : : : method": from my reading, this statement seems to invalidate the whole paper, and does not make sense. Do the authors mean: "The method of using CALIOP data as reference is applicable" **Adding the word "exclusively" after "valid" (line 720-721) clarifies that we (of course) don't want to invalidate the whole paper.**

-L449-450: "Because of this...": I do not understand the meaning of this sentence, please clarify it.
Reformulated (lines 723-724) and adding reference to Stubenrauch et al., 2013.

-L495: "A specific problem with the current method": its not an inherent problem of the method, but of data availability of active observations, I would thus suggest to use a different wording.
Rephrased (lines 697-710).

~~Detailed~~ 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 ~~characterisation of AVHRR global cloud detection performance of the CM SAF CLARA-A2 climate data record based on CALIPSO-CALIOP cloud information~~

Karl-Göran Karlsson[1], Nina Håkansson[1]

[1]Swedish Meteorological and Hydrological Institute, Folkborgsvägen 17, 601 76 Norrköping, Sweden

Correspondence to: Karl-Göran Karlsson (Karl-Goran.Karlsson@smhi.se)

Abstract. The ~~cloud detection performance~~ sensitivity in detecting thin clouds of the cloud ~~mask~~ screening method being used in the CM SAF cloud, albedo and surface radiation dataset from AVHRR data (CLARA-A2) cloud climate data record (CDR) has been evaluated ~~in detail~~ using cloud information from the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) onboard the CALIPSO satellite. ~~Validation results~~The sensitivity, including ~~their~~ its global ~~distribution~~variation, ~~have~~ has been ~~calculated~~ studied ~~from~~ based on collocations of AVHRR and CALIOP measurements over a ten-year period (2006-2015). ~~The sensitivity of the results to the cloud optical thicknesses of CALIOP-observed clouds were studied leading to the conclusion that~~ T~~t~~he ~~global~~ cloud detection sensitivity has been ~~(~~defined as the minimum cloud optical thickness for which 50 % of clouds could be detected~~,)~~ ~~was estimated~~with the global average sensitivity estimated to be 0.225. After ~~applying this optical thickness threshold to~~ using this value to reduce the CALIOP cloud mask (i.e., clouds with optical thickness below this threshold were interpreted as cloud-free cases), cloudiness results were found to be basically unbiased over most of the globe except over the polar regions where a considerable~~y~~ underestimation of cloudiness could be seen during the polar winter. The overall probability of detecting clouds in the polar winter could be as low as 50 % over the highest and coldest ~~portions~~ parts of Greenland and Antarctica, showing that also a large fraction of optically thick clouds remains undetected here. The study included an in-depth analysis of the probability of detecting a cloud as a function of the vertically integrated cloud optical thickness as well as of the cloud's geographical position. Best results were achieved over oceanic surfaces at mid-to-high latitudes where at least 50 % of all clouds with an optical thickness down to a value of 0.075 were detected. Corresponding cloud detection sensitivities over land surfaces outside of the polar regions were generally larger than 0.2 with maximum values of approximately 0.5 over the Sahara desert and the Arabian Peninsula. For polar land surfaces the values were close to 1 or higher with maximum values of 4.5 ~~over~~ for the ~~geographically highest parts~~ parts with the highest altitudes over ~~of~~ Greenland and Antarctica. It is suggested to also quantify the detection performance of other CDRs in terms of a sensitivity threshold of cloud optical thickness which can be estimated using active lidar observations~~The validation method is suggested to be applied also to other satellite-based CDRs. and~~ Vvalidation results are also proposed to be used in Cloud Feedback Model Intercomparison Project (CFMIP) Observation Simulation Package (COSP) simulators for cloud detection characterisation of various cloud CDRs from passive imagery.

1 Introduction

Monitoring the global amount, ~~and~~ distribution ~~of clouds as well as assessing the~~and optical properties of clouds ~~appear~~ is increasingly important as a result of the increasing evidence ~~following the growing insight~~ that the parametrization of cloud processes and cloud-aerosol interactions including related climate feedbacks, ~~cloud description and cloud-aerosol feedback~~

Formaterat: Understruken

~~processes stand out as key~~are critical contributors to the uncertainty ~~factors~~ in climate change analysis and in ~~climate~~ predictions from climate models (Stocker et al., 2013). However, it is encouraging in this ~~aspect~~ respect to note~~is~~ the steadily increasing amount of observations from spaceborne ~~from~~ passive and active sensors (an excellent overview is available at https://www.wmo-sat.info/oscar/) and the ~~continuous prolongation~~prolonged growth of the observational records from ~~for some~~ the initial satellite sensors ~~families since the time of introducing reliable and sustainable satellite observation systems back~~launched in the 1970's. These early satellite observations, ~~basically~~ which consist~~ing~~ of spectral radiance measurements, can be used to retrieve information about~~on~~ clouds and other relevant Earth-Atmosphere parameters. Most importantly ~~is that~~ they have now evolved into time series of observations with lengths approaching four decades, which qualifies them for use as climate data records (CDRs) ~~in combination with other Earth surface-based climate observations~~. Examples of CDRs built upon such observations are described by Rossow and Schiffer (1999), Karlsson et al. (2017), Heidinger et al. (2014) and Stengel et al. (2017).

~~From the climate analysis perspective, t~~The advantage of using satellite-based observations for climate analysis is ~~naturally the~~ their global ~~view~~coverage. A similar ~~view~~ coverage is very difficult to achieve ~~from~~ with surface-based observations alone because of the ~~inhomogeneous and sometimes lacking coverage~~sparsity of the surface-based observational network. ~~Similar to many other kinds of observations, this concerns also~~This is particularly true for observations of cloudiness and ~~the information on~~ cloud properties, where large parts of the Earth, ~~(e.g.~~ especially oceanic and polar regions,~~)~~ are still poorly covered. ~~However, t~~The different observation capabilities and conditions for space-based sensors and ~~Earth~~ surface-based observations also leads to problems when trying to characterise the accuracy of space-based CDRs. Although the quality of observations may be estimated for selected Earth positions or for small~~er~~ regions with dense surface networks, it is very difficult to achieve a representative and homogenous view of the accuracy over the entire globe using surface observations. The ~~importance of the CDR~~ quality of CDRs ~~aspect reflects the fact that~~ is especially important as observations used for climate monitoring must be very accurate to ~~be able to~~allow the reliable~~y~~ estimation of~~e~~ potential climate change signals (Ohring et al., 2004), which is a central aspect in the planning and definition of the global climate observing system (Dowell et al., 2013). ~~Linked to this are also recent efforts for~~For this reason, there is also a ~~becoming~~ need to become more stringent in the description of the uncertainty of CDRs by following international metrological norms (Merchant et al., 2017).

One solution for achieving both the global coverage and an improved ~~better prospect for~~ quality description is to ~~introduce and~~ make use of high-quality reference measurements from space-borne platforms (Dowell et al., 2013). This has already been successfully demonstrated by utilizing data delivered by the A-Train satellites, ~~(~~i.e., ~~Aqua Train~~Afternoon Satellite Constellation or sometimes referred to as the Afternoon Train) ~~concept.~~; This is a system of satellites operating in the same orbit configuration and ~~with~~ having close to simultaneous observation times (Stephens et al., 2002). ~~Particularly~~ The most important satellite in the A-train for the ~~cloud observation~~ detection of clouds ~~topic has been one of the satellites in the A-Train:~~is the CALIPSO satellite, which has ~~with~~ the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) onboard (Winker et al., 2009). The sensitivity of CALIOP to clouds in the atmosphere is much higher than for other space-based sensors and this makes it a natural reference for evaluating the cloud detection efficiency in data records compiled from passive sensor data (e.g., as demonstrated by Heidinger et al., 2016).

This paper presents a detailed CALIOP-based evaluation of the cloud detection efficiency and the uncertainty of the cloudiness information provided by the CLARA-A2 (The CM SAF cloud, albedo and surface radiation dataset from AVHRR data" - second edition) CDR (Karlsson et al. 2017). This CDR was released in 2017 by the Climate Monitoring Satellite Application Facility (CM SAF); a project ~~being a part of~~belonging to the satellite ground segment of the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT, Schulz et al., 2009). The evaluation ~~of this~~

Formaterat: Understruken
Formaterat: Understruken
Formaterat: Understruken
Formaterat: Understruken
Formaterat: Understruken

~~CDR~~presented is based on an original validation method described by Karlsson and Johansson (2013) ~~but now~~ which has been extended with several new features. The method ~~has been~~was first updated to use the latest revision of the CALIPSO-CALIOP dataset (Version 4) and results showing the impact of this change are presented. The study~~and it is also taking~~ then takes advantage of the greatly extended CALIOP observation period (here covering almost 10 years) to ~~allowing the~~ monitor~~ing of~~ globally averaged cloud conditions ~~with~~ in ~~fine~~ unprecedented detail~~s~~. The achieved validation results, which cover approximately one third of the CLARA-A2 observation period, can be considered to be the best currently available characterisation of the global quality of the CLARA-A2 cloud data record. A specific enhancement of the original validation method is the estimation of the geographical distribution of cloud detection probability as a function of cloud layer optical thickness. Section 2 describes the CLARA-A2 and CALIPSO datasets, Section 3 outlines the extended validation method and the compiled validation dataset and is followed by results in Section 4. ~~Finally,~~ Section 5 ~~summarizes~~ discusses the results and ~~discusses~~ Section 6 provides conclusions and proposes potential future applications.

2 Data

2.1 The CLARA-A2 climate data record.

CLARA-A2 is ~~based on~~constructed from historic measurements of the Advanced Very High Resolution Radiometer (AVHRR) operated onboard polar orbiting NOAA satellites ~~as well as onboard~~and the Metop polar orbiters operated by EUMETSAT since 2006. AVHRR ~~is~~ measur~~esing in~~ radiation in five spectral channels (two visible and three infrared channels) with an original horizontal field of view (FOV) ~~resolution~~ size at nadir of 1.1 km~~, although t~~. The data used in CLARA-A2 is a resampled version of these measurements at a reduced resolution of~~(~~5 km~~,)~~ defined as global area coverage (GAC)~~ resampled version of these 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. Only resampled GAC data is available~~ globally~~ (i.e., being archived) globally over the full period since the introduction of the AVHRR sensor in space. The resampling of original data into GAC representation means that four out of five original FOVs are selected for the first scan line while the next two scan lines are ignored. Radiances for these four selected FOVs are then averaged and ~~then~~ used to represent the GAC FOV consisting of 15 original full resolution FOVs. Thus, only about 25 % of the nominal GAC FOV is actually observed (see also visualization in Figure 1 in Section 3.2).

~~This second CLARA edition~~CLARA-A2 ~~is an~~ improve~~d~~ and extend~~ed follow up of~~ the first version of the data record released in 2012 (Karlsson et al., 2013) and ~~is~~ now cover~~ing~~ a 34-year time period (1982-2015). Original visible radiances were inter-calibrated and homogenised, using MODIS (Moderate Resolution Imaging Spectroradiometer) data as a reference, before ~~applying the various parameter retrievals~~generating each component of the CLARA-A2 product portfolio. The inter-calibration was based on the~~an original~~ method introduced by Heidinger et al. (2010), which has now ~~has~~ been updated (MODIS Collection 6) and extended (six years have been added). This updated calibration is described by Devasthale et al. (2017). CLARA-A2 features ~~a range of~~of the following cloud products: cloud mask/cloud amount, cloud top temperature/pressure/height, cloud thermodynamic phase, and (for liquid and ice clouds separately) cloud optical thickness, particle effective radius and cloud water path. These c~~C~~loud products are available as monthly and daily averages in a 0.25 by 0.25 degree latitude-longitude grid and also as daily resampled global products (Level 2b) in a 0.05 by 0.05 degree latitude-longitude grid. The daily resampled products are valid per satellite and orbit node (ascending or descending)~~.~~ while ~~T~~ the daily average product is an average of all available daily resampled products and the monthly products are the averages of all the daily average products. Cloud parameter results are also presented as multi-parameter distributions (i.e., joint frequency histograms of cloud optical thickness, cloud top pressure and cloud phase) for daytime conditions. ~~Besides~~

As well as cloud products CLARA-A2 also includes surface radiation budget and surface albedo products. E and examples of the CLARA-A2 products can be found in Karlsson et al. (2017).

In this study, we focus exclusively on the quality of the original AVHRR GAC cloud mask because of its central importance for to the quality of all other CLARA-A2 products. Validation results for other CLARA-A2 products can be found in Karlsson et al. (2017) and in CM SAF 1 (2017). The method for generating the CLARA-A2 cloud mask originates fromis generated using an improved and extended version of the method first proposed by Dybbroe et al. (2005) but significant improvements and adaptations have been made forwhich enablesing reliable processing of the historic AVHRR GAC record. These improvements are described in detail in Karlsson et al. (2017) and in (CM SAF 2, 2017).

2.2 The CALIPSO-CALIOP cloud information.

An extensive description of the existing CALIPSO-CALIOP cloud and aerosol datasets can be found in Vaughan et al. (2009). In short, the used Cloud Layer product from CALIOP (denoted CLAY) used in this study provides information onf up to 10 individual vertically displaced cloud layers in the vertical. However,As the detection of a cloud layer requires that all layers above a giventhat layer are optically thin enough to allow the lidar signal to penetrate down to that particular layer., there can be a bias Thus, in reality the number of layers observedmay be higher if overlaying clouds are optically thick. The CLAY products are is provided in three different horizontal resolutions (along track): 333 meter ("single shot"), 1 km and 5 km. The resolutions cCoarser resolutions than 333 meters are constructed through averaging over several single shots. This is done to increase the signal to noise ratio for to allowing detection of thinner clouds than what can could be achieved at the original single shot resolution. Thus, CALIOP products at coarser resolution will be capable of including detecting more clouds than at finer resolutions and, in particular, it is preferable that studies of thin Cirrus clouds should preferably be based on products in the coarsest resolution at 5 km (Vaughan et al., 2009). Notcice that the nominal single shot resolution FOV size does not correspond to the true lidar FOV size but rather to the along-track sampling distance. Consequently, withAs the true lidar FOV size is only of 70 meters (Winker et al., 2007), less than 5 % of the nominal single shot FOV size is actually observed (see also Fig. 1 in Section 3.2).

An estimation of the cloud optical thickness of each layer is also provided but only for a FOV resolution size of 5 km. To bear in mind here is that,However these values are only reliable for clouds with relatively low optical thickness (below approximately 3), because of signal saturation in optically thick clouds, the cloud optical thickness values are only reliable for relatively thin clouds, i.e., with cloud optical thickness values below approximately 3 (Vaughan et al., 2009 and Sassen and Cho, 1992). In this study we have used the optical thickness interval 0-5 , claiming that useful information also slightly above the suggested value of 3 seems to be available. because the new CALIPSO CLAY dataset version 4.10 provides slightly increased cloud optical thickness values compared to previous versions. We interpret this change to represent underestimation in previous values. Despite this change there is still a high degree of uncertainty in values near the upper end of these limits and these may, in reality, include some clouds which are optically thicker.

The CALIPSO satellite follows the A-Train track in a sun-synchronous orbit with an equator-crossing local time of 01:30. It This means that observations from the NOAA satellites can be matched to CALIPSO-CALIOP data in near-nadir conditions for a full orbit if being they are in an orbit with the same or very close to the same equator-crossing time. For all other NOAA satellite orbits (and now also including the MetopETOP satellites), matchups are only possible at high latitudes close to +/- 70 degrees N/S. Since CALIPSO is operated in a slightly lower and faster orbit than the NOAA/MetopETOP satellites (i.e., orbital period of CALIPSO is 98.5 minutes while NOAA/Metop period is 102 minutes), close matchups in time (i.e.,

~~with observation time differences less than 5 minutes)~~ ~~can~~ are ~~only be~~ found with ~~an interval~~ a recurrence of ~~3-5~~approximately 2 days.

In this study, we have used the fourth reprocessed version of the CALIOP CLAY datasets (version 4.10), which was released in 2016. The main features of this updated version are described at https://www-calipso.larc.nasa.gov/resources/calipso_users_guide/qs/cal_lid_l2_all_v4-10.php.

Regarding the basic CALIOP cloud mask, the most relevant changes affecting this study are

1. Revised and improved basic cloud-aerosol-discrimination method
2. Removal of mis-classifications of aerosols and dust as clouds at certain locations at high latitudes (as discussed by Jin et al., 2014)
3. Inclusion of information on single shot cloud detection in the 5 km dataset, ~~(~~ the implications of which are~~to be~~ discussed ~~further~~ in Section 3.~~3~~2).

3 Validation analysis methods and datasets ~~and analysis methods~~

3.1 Some theoretical considerations about clouds

Cloudiness is not an absolute well-defined quantity like other cloud properties or most other geophysical parameters. Firstly, it depends on the scale of interest, i.e., ~~you need to specify~~ the ~~aerial~~ areal ~~extension~~ extent over which cloud cover has to be calculated needs to be specified. Secondly, and perhaps more~~st~~ importantly, ~~you need to define what you mean~~a definition of what is meant by a cloud is required to allow a subsequent quantitative use of the results. ~~, e.g.,~~For example, ~~h~~how optically thin or thick should a cloud be to be called a cloud? This threshold is important when studying the cloud impact on components of the radiation budget (Charlson et al., 2007 and Barja and Antuña, 2011). How to define clouds detected in satellite imagery is also related to the scale of individual clouds (Koren et al., 2008). The cloud definition aspect is often missing in studies describing various cloud data records. Typically, products and validation results are presented without any deeper discussion on for what clouds the results are really valid.~~The second aspect is in most cases not well-defined which has made the use of this quantity rather difficult.~~ A good example is found in comparison studies between satellite-derived and manual surface-observed cloudiness (e.g., Sun et al., 2015). Results from such studies are difficult to interpret because of the different observation geometries for the compared datasets and the lack of an objective and clear definition of the clouds being observed in ~~the surface reference~~either of the two datasets. Because of this ambiguity it has often been recommended to use ~~other~~ parameters other than cloudiness or cloud cover (as mentioned in WMO 1, 2012) to instead describe the effect of clouds (e.g. “cloud albedo”, “effective cloud cover” or “joint histograms of cloud top pressure and cloud optical thickness”) in climate analysis and climate model evaluation studies. Nevertheless, the need to get the geographical distribution of modelled clouds correct is still a crucial requirement (as pointed out in WMO 1, 2012), ~~also bearing in mind~~particularly when considering that parameters describing the effect of clouds are still critically dependent~~ing~~ on how you define the underlying cloud or cloud mask. This calls for continued studies of cloud cover from both the observational and modelling perspective. We claim here that the access to high-quality reference cloud observations from CALIPSO-CALIOP may help us to take a significant step forward regarding ~~this aspect~~ the use of a more strict quantitative definition of cloudiness. A ~~very strict definition~~detailed characterization of the clouds we are observing can be made using CALIOP data. Thus, the ability to observe similar clouds in data records based on passive imagery can then be assessed, which will augment the usefulness of these data records. The following sub-sections ~~will~~ outline ~~the way forward to~~a new approach which will enhance the value of results from such cloud validation studies.

Formaterat: Teckensnitt:(Standard) +Rubriker (Times New Roman)

3.2 Basic CALIOP matching method and matching geometry~~adaptation to version 4 CLAY products~~

The underlying method for matching the two cloud datasets is described in detail by Karlsson and Johansson (2013). However, because of the importance for the understanding of method extensions and the achieved results in this study, we repeat here the most important aspects:

1. Positions where the orbital tracks cross are identified for the orbits of the two datasets to be collocated.
2. If the time difference of the two observations at the crossing point is within a certain maximum time difference $T_{diffmax}$ the observations at this position are denoted Simultaneous Nadir Observations (SNOs). Only orbits with SNOs satisfying the maximum $T_{diffmax}$ criterion are selected for further collocation studies. A $T_{diffmax}$ value of 45 seconds has been used in this study. As a consequence of a slightly shorter orbital period for the CALIPSO satellite, collocations could then be made with an approximate two-day repeat cycle.
3. For NOAA satellites flying in an afternoon orbit (which is similar or almost similar to the orbit of CALIPSO), it is possible to compare observations also before and after the SNO point since both satellites continue to observe the same points on Earth close in time. For example, if using a maximum observation time difference of 3 minutes, almost all observations during an entire orbit along the CALIPSO track can be inter-compared. Not all observations from the NOAA satellite afternoon orbits will be made in nadir conditions but relatively close to nadir (i.e., within 15 degrees). The current study has used afternoon orbit data with an observation time difference to CALIOP of 3 minutes to ensure global coverage.
4. For NOAA and METOP satellites flying in a morning orbit, the orbital tracks will cross almost perpendicularly and SNOs will then only occur at high latitudes (near 70 degrees N/S). A consequence of this is that collocations can only be made over distances limited by the AVHRR swath width. Furthermore, all individual collocations will then have varying AVHRR viewing angles along the matched track. Matchups with morning satellite data are not included in this study because of the limited geographical coverage.

In order to better understand the effects of different sensor sampling conditions and the collocation geometry, Fig. 1 shows an idealised representation of CALIOP collocations with AVHRR GAC data for both afternoon and morning orbits. The figure is idealised in the sense that it shows the perfect collocation, i.e., a collocation where the centre positions of both GAC and CALIOP FOVs are perfectly matched. We repeat that the AVHRR GAC sampling means that four out of five original FOVs are selected for the first scan line (marked as blue filled circular FOVs in Fig. 1) while the next two scan lines are ignored (empty blue boxes in Fig. 1). Radiances for these four selected FOVs are averaged and then used to represent the entire GAC FOV consisting of 15 original full resolution FOVs (schematically described as 3x5 blue boxes in Fig. 1). The 5 km CALIOP FOV observation is represented as an array of 15 original 333 m resolution red boxes in Fig. 1. Notice that the true FOV of the CALIOP sensor is smaller in size. In Fig. 1 they are represented as red filled circles with 70 m size and separated by 333 m distances. The 5 km CALIOP cloud observation is composed through averaging over the 15 original measurements but also from averaging over measurements outside of the nominal 5 km distance. This is done to detect optically very thin clouds (cirrus clouds) which could not be detected solely from data within the nominal 5 km FOV (as described by Vaughan et al., 2009).

The different panels for afternoon and morning orbit collocations in Fig. 1 are meant to illustrate how collocation conditions change from the along-track collocation mode for afternoon orbits to the across-track collocation mode for morning orbits. As a contrast to afternoon satellites, the orbital tracks crosses then almost perpendicularly between CALIPSO and morning orbit satellites, explaining the shift to a horizontal instead of a vertical orientation of the array of CALIOP measurements in Fig. 1. The effects of the limited coverage of true AVHRR observations within the nominal GAC FOV and the different

Formaterat: Liststycke, Numrerad lista + Nivå: 1 + Numreringsformat: 1, 2, 3, ... + Starta vid: 1 + Justering: Vänster + Justerad vid: 0,63 cm + Indrag vid: 1,27 cm

Formaterat: Liststycke

orientations of the array of CALIOP FOVs for morning and afternoon satellites can be ignored if cloud elements have scales larger than 5 km. However, for cases with smaller scale (sub-pixel) cloud elements or cases with cloud edges within the GAC FOV, we can expect differences between AVHRR and CALIOP observations. The implications because of this for the collocation and validation results will be discussed further in Section 5.

Formaterat: Liststycke

As explained by Karlsson and Johansson (2013), binary cloud masks for 5 km FOVs from AVHRR and CALIOP are inter-compared and evaluated using a range of standard validation scores. ~~-~~However, prior to comparison, the content of the original 5 km CALIOP FOV observation is adjusted to be consistent with the corresponding cloud mask defined at 1 km resolution. ~~It uses a combination of the CALIOP CLAY products at 1 km and 5 km resolutions. The reason for combining the two CLAY datasets with different resolutions, rather than using exclusively the CLAY 5 km version with the same nominal resolution as AVHRR GAC data, was the observation~~This check was introduced after noting that global CALIOP-estimated cloudiness ~~estimated~~ for ~~each~~ individual orbits ~~wa~~is not always increasing when switching from the 1 km resolution dataset to the 5 km resolution dataset. Conceptually, cloudiness should increase for the 5 km datasets as it is better able to detect also the optically thinnest cloud layers in addition to those cloud layers detected at finer resolutions (Vaughan et al., 2009). However, a~~A~~ non-negligible fraction of cases (~ 3-5 %~~)~~ of all investigated cases in a preparatory study) actually showed lower cloud amounts for the 5 km resolution. ~~Ideally, there should be an increase since CALIOP products with coarser resolution should contain more thin clouds which are not seen in the higher resolution products.~~ This inconsistency comes as a side effect of the actual method used for creating the coarser resolution CALIOP datasets (Vaughan et al., 2009 and David Winker, CALIPSO Science Team, 2017, pers. comm.). Prior to performing the horizontal averaging of the CALIOP scattering signal over several single shots, some single shot views are excluded from the analysis if they contain~~ing~~ strongly reflecting boundary layer clouds or aerosols. In the vast majority of cases, the number of these removed single shots is less than 50 % of all single shot measurements within the 5 km ~~-~~FOV. ~~which~~ Considering the official 5 km FOV CALIOP cloud mask, this procedure would then still justify labelling of the 5 km FOV as cloud free if no other cloud layers are detected. However, in some areas the frequency of small-scale convective clouds may be high and for these cases this could lead to underestimated cloudiness in the 5 km products. Another important aspect is that strongly reflecting clouds on the sub-pixel scale of AVHRR GAC data may still be detectable because of non-linear radiance contributions (with similarities to the "hot spot" effect from fires) in the short-wave infrared channel at 3.7 µm (Saunders and Grey, 1985, and Saunders, 1986). Thus, to not include these clouds in the CALIOP datasets ~~would possibly punish AVHRR based methods in an unfortunate and undeserved way in the validation process~~might lead to too low or non-representative validation scores for some of the investigated cases. Karlsson and Johansson (2013) showed ~~also~~ that validation scores also improved for AVHRR-based cloud products when adding clouds from the 1 km datasets if 3 or more of the 1 km FOVs within the 5 km FOV were cloudy in cases when the original 5 km products were deemed cloud-free. For these added clouds from 1 km data, the 5 km cloud optical thickness (not estimated in CLAY 1 km data) was set to 5, i.e., at the maximum upper end of realistically estimated cloud optical thicknesses. This is a justifiable approach ~~should be reasonable since~~as these clouds are by definition strongly reflecting and in most ~~of the~~ cases ~~it~~ would lead to effective cloud optical thicknesses close to or above 5.

Formaterat: Rubrik 2

3.3 Adaptation to CALIPSO version 4 CLAY products

An important ~~preparatory~~ objective of ~~step in~~ this study was to ~~check~~ verify ~~if~~ that the method used by Karlsson and Johansson (2013) would still be applicable to the new version 4 ~~release~~ of the CALIOP CLAY product released in 2016 and ~~if~~ to investigate whether the validation results changed in any systematic way. Despite the implemented modifications

(mentioned at the end of section 2.2).~~Of importance here is that~~ the fundamental retrieval method for the CALIOP CLAY product has ~~basically~~ remained the same~~ despite the implemented modifications mentioned at the end of section 2.2~~. Consequently, the above mentioned inconsistencies between fine and coarse resolution CALIOP datasets are likely to remain and would need a similar post-processing adjustment as for previous version 3 products. However, the new version of the 5 km CALIOP cloud product (i.e., in this study we have used the standard CLAY product version 4.10) has been expanded to include full information ~~of which~~ on the single shots removed ~~that were removed~~ during the averaging process. Thus, the previous use of 1 km data in the method by Karlsson and Johansson (2013) could in principle be abandoned and ~~be~~ replaced by the direct use of this single shot removal information (the latter method to be called "modified method" in the following). Another ~~additional~~ improvement found in the ~~used~~ version 4.10 dataset is that the removed single shot FOVs have also been labelled as being either cloudy or filled with thick aerosols. This separation was not available in version 3 where all removed single shot FOVs were assumed to be cloudy. An inter-comparison of version 3 and version 4 products is presented in section 4.1.

~~The modified method was compared against the old method for a limited test dataset of 80 NOAA-18 matched orbits between October and December 2006 and the results are presented in Figs. 1 and 2. Figure 1 shows validation results for the two different approaches based exclusively on CALIOP CLAY products version 4.10. We are here using the same visualisation of the results for two validation scores (Hitrate and Kuipers score, see also discussion and definition in section 3.3) as in Karlsson and Johansson (2013). Results of using the original CALIOP cloud mask is given by the leftmost value with a filtered cloud optical thickness of 0.0. The curves are constructed by validating against a successively reduced CALIOP cloud mask were clouds being optically thinner than the values at the x-axis have been transformed from cloudy to clear cases. In this way we can estimate for which CALIOP cloud mask (i.e., for which filtered cloud optical thickness) we get the highest scores. Figure 1 shows practically identical results for the two methods or actually slightly improved results for the method using the single shot information. The improvement may come from the improved cloud aerosol labelling of removed single shots. Figure 2 shows the overall effect of introducing the new matching method and the new version 4 dataset compared to the results achieved using the former version 3 dataset and the previous matching method. We notice a small increase in the overall results (maximum scores) and a progression of the maximum values towards larger optical depths. We believe that the improvement in results reflects an improved CALIOP product and that the shifting of peak score values towards larger filtered cloud optical depths results from more realistic and larger optical depths in CALIOP version 4.10 data (as confirmed by David Winker, CALIPSO Science Team, 2017, pers. comm.). This is quite in line with expectations and we conclude that the modified method is an appropriate basis for further validation studies based on the updated CALIOP CLAY dataset.~~

3.~~4~~3 Applied validation concept and validation scores

~~An important difference in this study compared to the conditions prevailing for the study by~~Compared to the previous study by Karlsson and Johansson (2013) this study has ~~is the~~ access to CALIOP data for a much longer validation period; almost 10 years (2006-2015). This means that ~~that not only overall mean conditions can be approximated but also~~it is now possible to calculate the geographical distribution of validation results~~.~~, in addition to global mean conditions. ~~More clearly, after ten years we have compiled~~ Due to a sufficiently large amount of AVHRR-matched nadir looking CALIOP observations ~~to estimate cloud detection conditions over all locations. Thus, for~~ it is possible, for the first time, to ~~we can~~ evaluate the quality of a cloud CDR in a (close to) homogeneous way over almost the entire globe with the only exception being close ~~in the near vicinity~~ to the poles where CALIOP measurements are not available. Consequently, ~~we will~~the validation results ~~here focus on presenting our results~~calculated in this paper are presented as ~~in~~ global maps rather than as tables and figures with global mean values. For the plotting of these global maps ~~we have~~the results have been rearranged and calculated using ~~the results~~

~~in a global equal-area grid. We have used~~ a Fibonacci grid with 28878 grid points evenly spread out around the Earth approximately 75 km apart. The resulting grid has almost equal area and almost equal shape of all grid cells. ~~Fibonacci grids behave the same near the poles as at the equator,~~making it preferable to ~~compared to~~ traditional latitude-longitude grids which often ~~behave in a strange way~~introduce distortions near the poles. For further details on Fibonacci grids, see González (2009) and Swinbank and Purser (2006).

We ~~will estimate~~have used the same set of validation scores as those described and defined by Karlsson and Johansson (2013), namely:

- Mean error (bias) of cloud amount (%), describing the systematic error of the mean
- Bias-corrected Root Mean Square Error (RMS) of cloud amount (%), describing the random error of the mean
- Probability of Detection ($0 \leq POD \leq 1$) for both cloudy and cloud-free conditions relative to all observed cloudy or clear cases
- False Alarm Rate ($0 \leq FAR) \leq 1$) for both cloudy and cloud-free conditions relative to all predicted cloudy and clear cases
- Hitrate: Frequency (value between 0 and 1) of correct cloudy and clear predictions relative to all cases
- Kuiper's skill score ($-1 \leq KSS \leq 1$) where value 1 means perfect agreement, value 0 means uncorrelated (random) results and value -1 means consistently opposite results (see Karlsson and Johansson for the exact definition).

~~Observe~~ The results are computed by~~that we will~~ treating both CLARA-A2 ~~cloud masks~~ and CALIOP cloud masks as binary values, i.e., each FOV is considered as either fully cloudy or cloud free. ~~This approximation is acceptable for the estimation of cloud amount when we have a large number of matched observations as in this case.~~ The Kuiper's skill score can be used ~~for~~ to better identify~~ication~~ cases of mis-classifications ~~in cases~~ when one of the categories is dominating. ~~It punishes~~ The KSS is sensitive to misclassifications even if they ~~are~~ occur in only a small minority of ~~all~~ the studied cases. The KSS score ~~tries~~ aims to answer the question of how well the estimation separated ~~the~~ cloudy events from ~~the~~ cloud-free events. ~~A value of 1.0 is in this respect describing the situation of a perfect discrimination while the value -1.0 describes a complete discrimination failure.~~

~~According to, e.g. Merchant et al. (2017), a~~A minimum requirement for describing the accuracy of a parameter is to estimate the mean error or bias (giving the systematic error) and the variance of the error (giving the random error or dispersion) (Merchant et al. 2017). However, ~~for~~ to enable the identification of specific problems with cloud identification it is ~~also useful~~ necessary to look at the ~~other quantities~~additional scores mentioned above, particularly ~~and especially~~ in cases when one of the two categories ("cloudy" or "clear") is dominant~~ting~~. ~~The latter~~This is motivated by the fact that any cloud contamination (even if it is just a few cases) can have serious implications for parameter retrievals further downstream in the processing. ~~Thus, the use of a rather full toolbox of~~Therefore multiple validation scores ~~can be~~are needed to correctly identify all problematic and critical cases.

3.~~5~~4 Extension of the original validation method: ~~e~~Enhanced analysis and introduction of cloud layer detection probability

The use of the CALIOP cloud mask for validation of cloud masking methods based on passive imagery is rewarding but also challenging. ~~Especially, we~~It is known ~~that a comparison of~~from previous results ~~with those from~~which used the original CALIOP cloud mask ~~means~~ that ~~we compare results from a high sensitive sensor to results from sensors with a lower sensitivity.~~ there is a large difference in sensitivity between CALIOP (high sensitivity) and passive sensors (moderate to low

sensitivity) which leads to the ~~The~~ question ~~is~~: h~~H~~ow ~~shall we handle~~can this sensitivity difference be managed to ensure the generation of~~and still get~~ useful results?

~~Of importance here is that t~~There are two major risks ~~with~~ when comparing ~~results~~ cloud masks retrieved from passive sensors to the original CALIOP cloud mask:

1. The CALIOP dataset will include sub-visible clouds (Martins et al., 2011) which are not possible to detect in passive imagery.
2. In areas where sub-visible clouds exist in abundance, a method may have been 'overtrained' or 'overfitted' (e.g., if trained with CALIOP data by statistical regression methods) to always predict clouds since this gives the best overall validation scores.

These two problems can be handled by ~~zooming in~~focusing on what happens for clouds ~~having~~ that have different vertically integrated optical thicknesses as provided by the CALIOP 5 km cloud product. ~~We recall that the use of~~By applying successively reduced CALIOP cloud masks ~~(as applied in Figs 1 and 2) means that~~in the validation exercise we may exclude the thinnest clouds from the analysis by transforming them ~~to be interpreted as~~into cloud-free FOVs. This also means that ~~if~~ we can isolate clouds within finite cloud optical thickness intervals (i.e., ~~resulting from~~by subtracting two adjacent restricted CALIOP cloud masks with different filtered cloud optical thickness) ~~we can~~in order to calculate validation results exclusively for ~~those~~ this sub-set of clouds. If the cloud optical thickness interval is sufficiently small and the number of samples within ~~this particular~~ each interval is sufficiently high we may then estimate the method's efficiency in detecting a cloud (i.e., the cloud layer detection probability $POD_{cloudy}(\tau)$ where τ is the mean optical thickness or depth in the given interval) with this particular cloud optical thickness ~~given by the mean cloud optical thickness in this interval~~. We may then expect to see low detection scores for ~~low~~ small optical thicknesses ~~but~~ with scores ~~that will increase with~~ improving as ~~increasing~~ cloud optical thickness values increase. We argue that a special situation occurs when this cloud layer detection probability for the first time exceeds 50 % for increasing cloud optical thicknesses. ~~It~~ This marks an important performance point which could be seen as a minimum performance requirement: a~~A~~t this cloud optical thickness we detect at least 50 % of all clouds. In the following we will denote this value of the filtered cloud optical thickness as the method's ***cloud detection sensitivity***. ~~It is also clear that we should get~~There should also be a peak in the Hitrate parameter at exactly this point. For ~~lower~~ small optical thicknesses, scores would improve if we filter out thin clouds, while for ~~higher~~ larger optical thicknesses scores start to decrease ~~since we then transform~~as too many correctly detected clouds are transformed to the cloud-free case. We ~~argue~~ maintain that the best way ~~of~~ to evaluate~~ing~~ a cloud masking method ~~would be~~is to estimate this cloud sensitivity parameter and to re-compute all validation scores after applying optical thickness filtering ~~with~~ using exactly this value. This ~~would~~ describe~~s~~ a method's optimal performance when using CALIOP cloud masks as the reference. The cloud detection sensitivity parameter ~~would~~ define~~s~~ the method's cloud detection capability in terms of the thinnest cloud that can confidently be~~ing~~ detected ~~with confidence~~. Furthermore, ~~and~~ the validation scores computed at this ~~particular~~ value of the filtered optical thickness ~~would~~ then define the method's optimal performance (in terms of the Hitrate) taking into account also false classifications. An important ~~additional or~~ complementary parameter in this context ~~would be~~is the false alarm rate in the unfiltered case ($FAR_{cloudy}(\tau=0)$) since this parameter ~~is~~ does not depend~~ing~~ on any filtering of optically thin clouds. $FAR_{cloudy}(\tau=0)$~~This parameter could~~ can ~~preferably~~ be used to investigate the degree of overtraining of a method (according to second bullet above). In the following Section 4, we ~~will~~ present results of the cloud detection sensitivity and~~,~~ a range of validation scores computed at the point of the cloud detection sensitivity (i.e., using a CALIOP

cloud mask filtered for thin clouds using the cloud detection sensitivity parameter as the optical thickness threshold) ~~and $FAR_{cloudy}(\tau=0)$.~~ Most of these results ~~will be~~are presented as global maps.

3.6 The final compiled validation dataset

We have matched a total number of 5747 global afternoon orbits of the NOAA-18 and NOAA-19 satellites with corresponding CALIPSO-CALIOP data in the time period October 2006 to December 2015. Due to increasing orbital drift of the NOAA-18 satellite after 2010 (with resulting deviation from the A-Train orbit and increasing off-nadir viewing angles for matchups), the matchup dataset contains a small fraction of observations with higher satellite zenith angles. The observation time difference is limited to 3 minutes and the spatial matchup error was maximised to 2.5 km (as a consequence of using the nearest neighbouring technique and after assuming negligible geolocation errors). This resulted in more than 23 million global matchups. The distribution of the matchups is shown in Fig. 2 using a Fibonacci grid resolution of 75 km.

Figure 2 shows a large variation in coverage as a function of latitude with a minimum number of matchups occurring at low latitudes and a maximum of matchups for the highest latitudes. Although the likelihood for a valid matchup to occur is the same everywhere on a particular matched orbit, the pattern of the matchup numbers is explained by the converging orbital tracks towards the poles. Furthermore, the large variation with some distinct features (e.g., over the Pacific Ocean) shows that it was not possible to extract all theoretically available matching cases (some periods with loss of data exist for both CALIOP and AVHRR). Although there is not fully homogeneous global coverage the dataset represents the best possible effort in that direction that we can make at present. Even at low latitudes the number of matches generally exceeds 300 for a grid resolution of 75 km, 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.

Formaterat: Normal

4 Results

4.1 ~~Data coverage~~Results from inter-comparisons of validation results based on CALIPSO-CALIOP version 3 vs version 4

Results from t~~T~~he modified validation method were~~was~~ compared against ~~the~~ results from the old method for a ~~limited~~ test dataset of 80 NOAA-18 CALIPSO-matched orbits between October and December 2006. ~~and t~~These results are presented in Figs. 3~~1~~ and 4~~2~~. ~~Figure 1~~Figure 3 shows validation results for the two different approaches based exclusively on CALIOP CLAY version 4.10 products ~~version 4.10~~. ~~We are here using the same~~The visualisation used here, showing ~~of~~ the results for two validation scores (Hitrate and Kuipers score, see also discussion and definition in section 3.4~~3~~) ~~as~~is identical to the approach seen in Karlsson and Johansson (2013). Results ~~of~~ using the original CALIOP cloud mask ~~is~~ are given by the leftmost value with a filtered cloud optical thickness of 0.0. The curves ~~are constructed by~~represent validations ~~ag~~ ~~against~~which use a successively reduced CALIOP cloud mask where clouds ~~being~~ optically thinner than the values at~~on~~ the x-axis have been transformed from cloudy to clear cases. In this way we can ~~estimate~~calculate for which CALIOP cloud mask (i.e., for which filtered cloud optical thickness) we get the highest scores. Fig~~ure~~ 3~~1~~ shows ~~shows practically identical results for the two methods or actually~~ slightly improved results for the method using the single shot information, although they are practically identical. The slight improvement may ~~come from~~be attributed to the improved cloud-aerosol labelling of removed single shots. Figure 4~~2~~ shows the overall effect of introducing the new matching method and the new version 4 dataset compared to the results achieved using the former version 3 dataset and the previous matching method. ~~We notice~~There is a small increase in the overall results (maximum scores) and a progression of the maximum values towards

larger optical depths. ~~We believe that the~~The improvement in results ~~reflects~~indicates an improved CALIOP product and ~~that~~ the shifting of peak score values towards larger filtered cloud optical depths ~~results from~~is indicative of more realistic and larger optical depths in CALIOP version 4.10 data (as confirmed by David Winker, CALIPSO Science Team, 2017, pers. comm.). ~~This~~ These results are ~~is quite~~ in line with expectations and ~~we conclude~~demonstrate that the modified method is an appropriate basis for further validation studies based on the updated CALIOP CLAY dataset.

Formaterat: Normal

~~We have matched a total number of 5747 global afternoon orbits of the NOAA-18 and NOAA-19 satellites with corresponding CALIPSO-CALIOP data in the time period October 2006 and December 2015. The study does not include results from satellites in morning orbit since these can only be matched with CALIOP data at high latitudes (further discussed in Section 6). Due to increasing orbital drift of the NOAA-18 satellite after 2010 (with resulting deviation from the A-Train orbit and increasing off-nadir viewing angles for matchups), the matchup dataset is unfortunately not exclusively based on AVHRR near-nadir observations even if they dominate. However, all NOAA-18 data were included here since we wanted a representative evaluation of the AVHRR CDR for the entire studied period. The observation time difference was limited to 3 minutes and the spatial matchup error was maximised to 2.5 km (as a consequence of using the nearest neighbouring technique and after assuming negligible navigation errors). This resulted in a total number of more than 23 million global matchups. The distribution of the matchups is visualized in Fig. 3 using a Fibonacci grid resolution of 75 km (which is also used in the following figures for the subsequent plotting of most of the results).~~

~~Figure 3 shows a quite varying degree of coverage as a function of latitude with a minimum of the number or matchups occurring at low latitudes and a maximum of matchups for the highest latitudes. Although the likelihood for a valid matchup to occur is the same everywhere on a particular matched orbit, the pattern of the matchup numbers is explained by the converging orbital tracks towards the poles. Furthermore, the large variation with some typical geographical features and variations also in the zonal direction shows that we have not been able to extract 100 % of all theoretically available matching cases (some periods with loss of data exist for both CALIOP and AVHRR). This means that the ambition of getting a homogeneous global coverage cannot be perfectly met but it is still the best effort in that direction that we can make. Even at low latitudes the number of matches generally exceeds 300 for a grid resolution of 75 km. Some exceptions can be seen, particularly over the Pacific Ocean, but we still believe that the number of samples is sufficient for obtaining a fair estimation of the cloud screening performance.~~

4.2 Results based on original CALIOP cloud masks compared to results excluding contributions from very thin clouds ~~Results based on one original and one restricted CALIOP cloud mask~~

Figure 54 shows the ~~achieved~~ global distribution of the Hitrate parameter when comparing to the original CALIOP cloud mask. Results indicate a fairly good ~~performance~~ cloud screening capability over mid- to high latitudes (especially over oceans) but degraded results ~~over~~ at most low latitudes and over the polar regions. ~~with t~~The poorest results ~~occur~~occur~~ing~~ over Greenland and Antarctica.

Further analysis of results is complicated by the fact that the original CALIOP cloud mask includes all CALIOP-detected clouds as explained in Section 3.54. In particular, we suspect that the rather poor results in Fig. 54 in the tropical region may be significantly influenced by the presence of sub-visible clouds.

~~If~~ By using all available matchups, we can calculate $POD_{cloudy}(\tau)$ ~~and plot results~~ for all values of τ (Fig. 65) using the method outlined in Section 3.5. Calculations have been based on optical thickness intervals of 0.05 in the range $0.0<\tau<0.5$, intervals of 0.1 in the range $0.5<\tau<1.0$ and intervals of 1.0 in the range $-1.0<\tau<5.0$ (results from the latter ~~results~~ interval are not shown in Fig. 65.). ~~From Fig.~~Figure 65 ~~we deduce~~shows that the cloud detection sensitivity (i.e., where a probability of

50 % is reached) ~~can be estimated to~~is 0.225 for the investigated AVHRR-based results. Consequently, we will use this value to ~~represent~~ indicate the optimal Hitrate results, with the ~~and we then get the~~ global distribution of these results ~~as~~ presented in Fig ~~ure~~ 7~~6~~. As expected, the results improve considerably for most ~~places~~ locations compared to Fig. 5~~4~~, ~~and~~ especially over low latitudes. ~~In most places,~~ Hitrates above 80 % are now achieved over most regions. The polar regions (at least the snow- and ice-covered parts) stand out as regions of poor quality with the worst ~~conditions~~ results seen over central Greenland and Antarctica. There is also s~~S~~ome degradation in the results ~~is still seen~~ over some regions at low-to-middle latitudes ~~and this will be analysed further in the next section~~.

~~We claim that T~~the results in Fig. 7~~6~~ give a much ~~better idea~~clearer measure of the ~~performance~~ cloud detection capability of the CLARA-A2 cloud ~~mask~~ screening method than ~~what was~~those shown in Fig. 5~~4~~, ~~especially since it is~~because they are now linked to a well-defined description of the involved clouds. We will ~~use~~ apply the same filtering approach ~~for~~ to obtain the results ~~to be~~ shown in the next sub-section.

4.~~3~~3 Additional validation scores ~~Complementary results based on a restricted CALIOP cloud mask~~

Figure 8~~7~~ presents results for the systematic (bias) and random errors (bias-corrected RMS) of the CLARA-A2 cloud amounts. It is clear that the cloud detection problems over the polar regions, as indicated by the Hitrate parameter in Fig. 7~~6~~, lead~~s~~ to a ~~massive~~ significant underestimation of cloud amounts, especially over ~~the parts being~~those areas normally covered with snow or ice. However, ~~notice that~~ this is an overall mean (close to an annual mean) and ~~that~~ the underlying results may be seasonally varying. For example, cloud detection in the polar summer season is considerably better than during the polar winter (as shown by Fig. 6~~5~~ in Karlsson et al., 2017). The ~~most unbiased~~ results with least bias are found over mid-to-high latitudes while some overestimation ~~ed cloud amounts are~~is seen over lower latitudes, particularly over oceanic surfaces. RMS values are ~~naturally also~~ high in the polar regions ~~but also~~and over what can be described as oceanic sub-tropical high regions. This agrees ~~also~~ well with the corresponding Hitrate results seen in Fig. 6~~5~~. RMS values are ~~also~~ low over dry desert regions but mostly as a consequence of the general lack of cloudy situations here.

~~In a further search of the~~To further investigate areas where ~~we have~~there is significant misclassification~~s~~ of cloudy and clear conditions we can study results of probability of detection of the ~~two~~ cloudy and clear categories in Figure 9~~8~~. For the cloudy category results are ~~mostly already~~consistent with those deduced from previous figures with the exception of ~~possibly for~~ the low probabilities of cloud detection over northern Africa and the Arabian Peninsula. For the clear category we note high values over predominantly dry land portions of the world while low values are seen over the tropical region and over oceanic storm track regions at high latitudes.

Results for the Kuipers score are shown in Fig. 10. This score does not show as much regional variability as the Hitrate score. Again, we note low score values over the snow-covered polar regions and over some desert regions. The largest difference to the Hitrate is seen over high-latitude oceanic regions where the Kuipers score show rather modest values while Hitrate showed relatively high score values.

Figure 11 show the corresponding false alarm rates for cloudy and clear conditions. We note high false alarm rates for cloudy conditions over tropical and sub-tropical regions (with some dominance for oceanic regions) while for clear conditions the largest false alarm rates are found in the polar regions.

~~It means that in this particular area, where cloudiness is generally low, we still find particular problems in detecting the few occurring cases of clouds. The reasons for this have to be investigated further but they are likely to be linked to remaining~~

~~uncertainties in the used surface emissivities over these dry and desert-like surfaces. The two maps in Figure 8 reveal another interesting feature: In areas where cloudiness is low (e.g., over sub-tropical ocean and land regions) POD_{cloudy} is low and where cloudiness is high (e.g., mid-latitude storm tracks and ITCZ near the equator) POD_{clear} is low. This contributes to give fairly low values of the Kuipers' score over these regions (Figure 9) leading to a slightly different distribution of results in comparison to the Hitrate (Fig. 6). However, we must remember that Hitrate is dominated by results for the dominating mode (cloudy or clear) while Kuipers punishes particularly the existence of misclassifications of the minority mode. Figures 8 and 9 reveal that even if the dominantly cloudy and clear regions are generally captured very well the few cases of the opposing mode have a high frequency of misclassifications. This is difficult to understand from the perspective of long term experience of AVHRR cloud screening. More clearly, cloud screening is generally understood to work best over dark and warm ocean surfaces in good illumination. So, why are results not better here (e.g., over oceanic sub-tropical high regions)? We believe that this unexpected behaviour is a consequence of the limitations of both AVHRR GAC data and CALIPSO-CALIOP data when it comes to the sampling of the true conditions within the nominal 5 km FOV. It was already mentioned in Section 2 that only about 25 % of the nominal AVHRR GAC FOV of 5 km is actually observed and that the corresponding figure for CALIOP single shot nominal FOV of 330 meters is as low as 5 %. Notice that the latter means that CALIOP is only able to cover about 0.3 % of the nominal 5 km FOV. This has important consequences for all cases when we have cloud elements present which are smaller in size than the nominal 5 km FOV. We first conclude that only in the case of having cloud elements larger than the nominal 5 km FOV we can be confident in getting the same results from AVHRR and CALIOP observations. For all other cloud situations involving clouds being smaller in size than 5 km the two data sources will give different results since the sensors will probe different parts of the 5 km FOV. The situation is made even worse by the fact that the AVHRR scan lines are perpendicular to the CALIPSO track when matching the two datasets in the near-nadir mode. It means that the CALIOP sensor will consistently probe a different part of the nominal 5 km FOV than AVHRR. Theoretically, a maximum of 3 CALIOP single shot measurements out of totally 15 would actually be able to measure the same spot on Earth as the AVHRR GAC measurement within the FOV of 5 km. A consequence of this must be that in the case of dominating fractional cloudiness with cloud size modes below the 5 km scale the random errors and the false-alarm rates should increase even if the bias could remain small. This is exactly what is observed over the oceanic sub-tropical high regions (Figure 7 and Figure 10, upper panel) also explaining the degraded overall scores in this region (in particular the POD_{cloudy} score in Fig. 8). These regions have a reduced total cloud amount in the annual mean (e.g., see Fig. 6 in Karlsson et. al., 2017), mainly because of generally more stable conditions here with prevailing large-scale subsidence (poleward parts of the Hadley cell) suppressing cloudiness in mid- to high layers and basically only allowing convective and stratiform boundary layer cloudiness to form. This boundary layer cloudiness consists to a large degree of scattered small-scale cumulus and stratocumulus clouds, i.e., typically the kind of clouds for which we would expect enhanced disagreeing results for the AVHRR and CALIOP datasets following the above reasoning. It is interesting to notice that not only oceanic areas show this feature. Also some eastern parts of continents show similar results, e.g. easternmost part of South America and Africa. It could mean that scattered cumulus cloudiness is the dominant mode of cloudiness also here. Finally, notice also that we can see exactly the same effect for fractional clear areas, e.g. over northern and southern hemisphere stormtracks at mid- to high latitudes as shown by the large FAR_{clear} values here in Fig. 10. We conclude that, because of the problems to correctly representing cases of both small-scale cloudiness and small-scale holes in cloud decks in the two datasets, validation results are probably underestimated (i.e., giving too low scores) over these dominantly cloudy or dominantly clear regions of the globe.~~

4.~~4~~3 Estimating the global variability of cloud detection limitations

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 to a clear improvement in the

results when compared to the original CALIOP cloud mask (i.e., comparing Figs. 5 and 7).~~The concept of presenting validation results after having removed all clouds with smaller optical depths than the cloud detection sensitivity parameter is undoubtedly a clear improvement compared to if only showing results based on the original CALIOP cloud mask.~~ However, ~~we still have the problem that the currently used~~the cloud detection sensitivity value currently applied is a global average ~~meaning that we can still have~~which could contribute to the large geographical variations in the ~~performance~~results. To investigate how serious this simplification is, we can plot the results of τ_{min}(POD>50) calculated exclusively for every Fibonacci grid point (Fig ~~ure~~ 12~~1~~). To reduce the uncertainty in this calculation due to ~~spuriously occurring~~ low number of samples per grid point as indicated in Fig. 2~~3~~ for low latitudes, we have ~~here~~ increased the radius of the Fibonacci grid from 75 km to 300 km. ~~We notice~~Figure 12 shows a considerable variation in cloud detection sensitivity over the globe ~~in Fig. 11~~. ~~Especially we notice~~It is clear that the cloud detection sensitivity is ~~generally~~ considerably lower than the global average value of 0.225 over ~~large parts of the~~most oceanic areas as well as over tropical land areas. On the other hand, values are generally larger than 0.225 over dry and desert-like regions and over high-latitude and polar land areas. For the polar land areas the cloud detection sensitivity frequently exceeds 1 and for some grid points even reaches values close to 5. These values~~, when put in relation to~~ contrast with the global average value of 0.225, ~~tell us~~indicating that more representative (and most likely higher) ~~overall~~ validation scores could have been achieved if ~~re-calculating validation scores per Fibonacci grid point using these~~ globally resolved cloud detection sensitivity values were used to re-calculate each of the validation scores. However, we have not taken this step here because of the relatively low number of samples in some grid points (even at the 300 km scale).

We ~~may~~ can also visualise the variable cloud detection sensitivity by plotting the same kind of cloud layer probability curves as in Fig. 6~~5~~ for a selection of individual grid points~~ in Fig. 11~~. Figure 13~~2~~ shows ~~such~~ these curves for the three locations marked ~~out~~ in Fig. 12~~2~~. ~~The three points describe three typical but also extreme situations.~~ The blue curve in Fig. 13~~2~~ shows cloud layer detection probabilities for a distant (from land) point in the North Atlantic Ocean. It marks a position where cloud detection is clearly ~~works the best~~most effective compared to ~~in~~ the global ~~perspective~~average. The cloud detection sensitivity value is ~~as low as~~ 0.075 ~~here~~ at this location demonstrating ~~meaning~~ that even very thin clouds are well detected there. The c~~C~~loud detection ~~performance~~ capability~~is~~ also reach~~es~~~~ing~~ a maximum value of approximately 95 % ~~already at~~by τ = 0.5. This is ~~probably~~ considered to be as high as can be reached because of the limitations of the datasets, for instance the remaining and unavoidable AVHRR-CALIOP mis-location and matching problems (both in time and space). As a contrast, a grid point located in the Sahel region (green curve in Fig. 13~~2~~) shows ~~less good~~worse results with a cloud detection sensitivity of 0.375 and ~~it barely reaches~~ maximum cloud detection ~~performance~~ capability only observed at τ = 3.5 and higher. However, ~~even worse results are recorded for the~~a more extreme case is the location ~~point~~ over central Greenland (red curve in Fig. 13~~2~~). The cloud detection sensitivity ~~is~~ here is as ~~high~~ large as 1.5 and ~~it is clear that not~~ even at a maximum τ value of 4.5 we can not come close to achieving an optimal cloud detection ~~performance~~capability. Thus, over a snow-covered and often extremely cold location we cannot even detect all optically thick clouds ~~(~~which is ~~in line~~consistent with the low POD_{cloudy} results seen over Greenland and Antarctica in Fig. 9~~8~~, upper panel).

The ~~visualisation of the~~ results in Fig. 13~~2~~ ~~reveals~~ again indicate that ~~we are~~the validation matchup dataset ~~probably~~ slightly undersampl~~es~~~~ing~~ the true conditions ~~at some individual~~for a limited number of grid points. This is indicated by the unexpected decrease in POD at some points for increasing τ values. Theoretically, one would expect a steady increase in POD as a function of τ.

5 Discussion

There are several features of the results depicted in Figs 7-11 which warrant further attention and discussion. One of these is the reduction in performance observed over areas which are known to be dry and mostly cloud-free. The POD_{cloudy} results in Fig. 9 show particularly low values over the Sahara Desert and the Arabian Peninsula. This indicates that in these particular areas, where cloudiness is generally low, CLARA-A2 still has difficulty detecting the few cloudy cases which occur. The exact reasons for this have to be investigated further but are likely linked to remaining uncertainties in the surface emissivities used over these semi-arid regions and deserts.

Another feature to discuss is the overestimation of cloudiness over low and medium latitudes (especially over oceans) seen in the Bias plot in Fig. 8. This feature illustrates how it is difficult to find a simple representative way of evaluating results while also taking into account the existence of sub-visible clouds. The method applied in Fig. 8 (and in all Figs 7-11) is to ignore cloud contributions in the CALIOP dataset for clouds having an optical thickness less than 0.225. But, as already mentioned in Section 4.4, the latter value is a global mean value and in many places on Earth clouds with smaller optical thicknesses are actually detected confidently. This is clearly demonstrated in Fig. 12 where the cloud detection sensitivity over oceanic surfaces is noticeably better (smaller) than the global mean of 0.225. This means that by applying the global value 0.225 as the filtering threshold of CALIOP-detected clouds, many clouds which were originally correctly detected in CLARA-A2 will now be treated as being falsely detected. If a locally representative value of the cloud detection sensitivity (as shown in Fig. 12) is used for the CALIOP filtering procedure, this apparent overestimation of clouds would largely disappear. However, to confidently apply such localised filtering a larger set of collocated observations is required to remove the sensitivity to low numbers of samples in individual grid points. Such a study will be possible in a few more years once an even larger matchup dataset has been collected. An extended dataset could also allow a further sub-division of the dataset to study the diurnal and seasonal variation of the validation results.

A more interesting and general feature is shown in Fig. 9: In areas where cloudiness is low (e.g., over sub-tropical ocean and land regions) POD_{cloudy} is low and where cloudiness is high (e.g., over mid-latitude storm tracks and near the equator) POD_{clear} is low. This explains to a large extent the fairly low values of the Kuipers' score over these regions (Fig. 10) leading to a slightly different distribution of results in comparison to the Hitrate (Fig. 7). However, we must remember that Hitrate is dominated by results for the dominating mode (cloudy or clear) while the Kuipers score highlights more clearly the existence of misclassifications of the minority mode. Figs 9 and 10 reveal that even if the dominantly cloudy and clear regions are generally captured very well the few cases of the opposing mode have a high frequency of misclassifications. This result is difficult to understand from the perspective of long-term experience of AVHRR cloud screening, as cloud screening works best over dark and warm ocean surfaces in good illumination. So, why are results not better here (e.g., over oceanic sub-tropical high regions)? We believe that this unexpected behaviour is a consequence of the limitations of both AVHRR GAC data and CALIPSO-CALIOP data when it comes to the sampling of the true conditions within the nominal 5 km FOV.

To understand this we have to go back to Fig. 1 displaying the conditions for the matching of AVHRR GAC and CALIOP observations and the overall collocation geometry. Sections 2.1 and 3.2, together with Fig. 1, clearly describes how only about 25 % of the nominal 5 km AVHRR GAC FOV is actually observed by AVHRR and that the corresponding figure for CALIOP single shot nominal FOV of size 330 meters is as low as 5 %. Notice that the latter means that CALIOP is only able to cover about 0.3 % of the nominal 5 km FOV. This has important consequences for all cases where we have cloud elements present which are smaller in size than the nominal 5 km FOV. We can first conclude that only in those cases containing cloud elements larger than the nominal 5 km FOV can we be confident that AVHRR and CALIOP observations will be comparable. For all other cloud situations involving clouds smaller than 5 km or when a cloud edge occurs within the

GAC FOV, the two data sources will give different results since the sensors will observe different parts of the 5 km FOV. The situation is compounded by the fact that the AVHRR scan lines are perpendicular to the CALIPSO track when matching the two datasets in the near-nadir mode (Fig. 1, upper panel). This means that the CALIOP sensor consistently probes a different part of the nominal 5 km FOV to AVHRR. Theoretically, a maximum of 3 CALIOP single shot measurements (out of a total of 15) would be able to measure the same spot on Earth as the AVHRR GAC measurement within the FOV size of 5 km. However, it is clear from Fig. 1 that in a non-negligible fraction of cases, the two sensors will not even observe any common part of the nominal GAC FOV. This occurs when the nearest-neighbour matching of GAC and CALIOP FOVs places the CALIOP FOV in the rightmost part of the GAC FOV (see Fig. 1, upper panel). A direct consequence of these differences between the actual AVHRR and CALIOP measurements is that, in the case of dominating fractional cloudiness with cloud size modes below the 5 km scale, the random errors and the false-alarm rates will increase even if the overall bias remains small (assuming that the cloud element distribution within the GAC FOV is random over a long time period, i.e., as expected for climate data records). This behaviour is exactly what is observed over the oceanic sub-tropical high regions (Fig. 8 and Fig. 11, upper panel) and also explains the degraded overall scores in this region (in particular the POD_{cloudy} score in Fig. 9) relative to other surrounding regions.

These regions of interest also have a reduced total cloud amount in the annual mean (e.g., see Fig. 6 in Karlsson et. al., 2017), mainly because of the more stable atmospheric conditions here. The prevailing large-scale subsidence (poleward parts of the Hadley cell) in these locations suppresses cloudiness in mid- to high layers and is conducive only to the formation of convective and stratiform boundary layer clouds. This boundary layer cloudiness consists mainly of scattered small-scale cumulus and stratocumulus clouds, i.e., typically the kind of clouds for which we would expect enhanced disagreement between the AVHRR and CALIOP datasets as a result of variability wihin the 5 km FOV. It is interesting to note that this feature is not exclusive to oceanic areas. In addition some eastern parts of continents show similar results, e.g. easternmost part of South America and Africa. This could indicate that scattered cumulus cloudiness is also the dominant mode of cloudiness in these locations. Finally, notice also that we can see exactly the same effect for fractional clear areas, e.g. over northern and southern hemisphere stormtracks at mid- to high latitudes as shown by the large FAR_{clear} values in Fig. 11. We conclude that, because of the problems with correctly representing cases of both small-scale cloudiness and small-scale holes in cloud decks in the two datasets, the validation results could be underestimated (i.e., giving too low scores) over these dominantly cloudy or dominantly clear regions of the globe. This reduction of scores would then be largely attributed to mis-matches due to GAC FOV geolocation errors (which are not zero), matchup errors (explained by the nearest-neighbour matching of GAC and CALIOP FOVs) and to the different cloud representation in each dataset rather than to real cloud detection problems. Thus, examination of the cloud detection capability of a method should also take into account the scales of clouds being investigated. A consequence of this is that detailed studies of small-scale convective cloudiness should rather be based on original resolution AVHRR and CALIOP observations than on datasets with a coarse resolution data representation.

Finally, a specific problem with the applicability of the current method is the inability to assess the global quality of products from polar satellites in morning orbits (e.g., from the NOAA-17 and Metop satellites) as a consequence of CALIPSO following an afternoon orbit. Matchups with CALIPSO-CALIOP are consequently only possible at high latitudes leaving low-to-middle latitudes without reference observations for AVHRR products. Previous cloudiness comparisons for morning satellites at high latitudes (CM SAF 1, 2017) show good agreement with corresponding results from afternoon satellites (assuming that diurnal cycle cloud effects are small at high latitudes). Thus, for cloud amount information (in contrast to some other cloud parameters, like cloud effective radius) there is no reason to suspect large differences between morning and afternoon results even if morning orbit data is partly using measurements in another spectral band (at 1.6 μm) in the short-

wave infrared spectral region. However, this needs to be confirmed in the future through the use of reference data from the Cloud-Aerosol Transport System lidar (CATS, https://cats.gsfc.nasa.gov/) on the International Space Station or by use of data from the Earth Cloud Aerosol and Radiation Explorer (EarthCARE) mission (http://m.esa.int/Our_Activities/Observing_the_Earth/The_Living_Planet_Programme/Earth_Explorers/EarthCARE/ESA_s_cloud_aerosol_and_radiation_mission) for new afternoon satellites with two coexisting short-wave infrared channels onboard (e.g. NOAA-20).

Formaterat: Rubrik 1

6. Conclusions

We have shown that with ~~the~~ access to the latest cloud information provided by the high-sensitivity CALIPSO-CALIOP lidar (CALIOP Version 4.10 dataset, covering almost a full decade (2006-2015)~~,~~ it is possible to construct a detailed global ~~maps~~ analysis of the cloud detection ~~performance parameters~~sensitivity and other skill scores of the cloud screening method used in the AVHRR-based CLARA-A2 cloud climate data record. A wide range of validation scores, ~~several of them~~including those complementary to the essential scores describing systematic and random errors, have been used to get a very detailed picture of the cloud screening efficiency of CLARA-A2. Furthermore, by use of the CALIOP-derived information on cloud optical thickness, it has been possible to make a clear definition of ~~what~~ which clouds ~~that has~~have been observed and thus for which clouds the validation scores are valid. We believe this to be crucial ~~for allowing~~ to the further quantitative use of the results. The method ~~as such~~ is not specifically developed or valid exclusively for the CLARA-A2 cloud masking method but ~~should~~ is also~~be~~ applicable to any method utilizing CALIOP data as ~~its~~ a reference. ~~Because of this the proposed method is suggested to be used also in studies evaluating different methods in an objective way.~~Consequently, we propose that this method be used in future inter-comparisons of results from different cloud masking methods and cloud CDRs (following the example by Stubenrauch et al., 2013).

~~The necessity~~It is necessary to specify the clouds being investigated ~~is partly linked~~because ~~to the fact that~~ the CALIOP sensor is capable of detecting clouds which ~~could be considered as being~~are fundamentally "sub-visible" for passive imaging sensors. Therefore, a globally estimated minimum cloud optical thickness value (denoted "Cloud detection sensitivity"), ~~and~~ for which the majority of clouds would be detected~~,~~ was estimated to be 0.225 for the CLARA-A2 cloud masking method. This value was used to remove contributions to validation scores from thinner clouds than this minimum optical thickness, thus maximising the validation scores. For example, by utilising this definition of detectable clouds, resulting cloud amounts were found to be unbiased over most locations of the world except for a major underestimation over the polar regions. For the latter, a large part of all clouds still remain undetected during the polar night and this fraction can be as high as 50 % over the coldest and highest portions of Greenland and Antarctica. Under these conditions not even optically thick clouds may be detected due to the very similar thermal characteristics of clouds and Earth surfaces. ~~Another observed deviation is a small overestimation of cloudiness over tropical ocean areas.~~ Land-ocean differences were generally small with only results over Greenland and Antarctica standing out as clear exceptions.

The study revealed some ~~small but noticeable degraded results~~interesting reductions in performance over mainly sub-tropical ocean areas. ~~Here,~~In these locations random errors were ~~surprisingly high~~elevated indicating a decrease~~d~~ in agreement between AVHRR and CALIOP observations despite otherwise very favourable cloud detection conditions (e.g., warm ocean temperatures and good illumination conditions).We argue that this is caused by ~~insufficient and mutually~~the different sampling conditions within the studied 5 km FOV of the AVHRR and the CALIOP sensors, which is particularly evident in cases wher~~e~~n small-scale boundary layer cloudiness dominates the cloud situation. Because of this we suspect that the cloud

detection ~~performance~~ capability over these areas ~~is~~ could actually be better than that shown by these results. ~~in reality compared to what the presented results show.~~

An important novel feature of this study compared to many previous validation efforts based on CALIPSO-CALIOP data is the estimation of the probability of detecting an individual cloud as a function of its vertically integrated optical thickness and its geographical position on Earth. This was accomplished by isolating finite optical thickness intervals in the CALIOP cloud information and calculating validation scores for this subset of data in a coarse global grid. Results show a substantial variation compared to the ~~previously mentioned~~ global mean optical thickness value of 0.225 for the thinnest retained cloud in the CALIOP cloud mask to give optimal global ~~performance of~~ validation scores. The highest sensitivity to clouds in AVHRR data is generally found over mid-to-high latitude ocean surfaces. Here, clouds with cloud optical thicknesses as low as 0.075 can be detected efficiently. This ~~can be compared~~is in comparison to a value of approximately 0.2 over tropical oceans and ~~generally larger~~typically greater than 0.2 over most land surfaces. The latter value reaches 0.5 ~~at~~ over some dry and desert-like regions (e.g., the Sahara Desert and the Arabian Peninsula) and increases towards or beyond 1 over polar regions with a highest value of 4.5 found over Greenland and Antarctica. ~~The latter~~These results indicate~~s~~ that not even optically thick clouds can be confidently identified over Greenland and Antarctica during the polar winter. While these are not entirely new findings (e.g., see Karlsson and Dybbroe, 2010), this study has increased the confidence in the validation results over the polar regions. Consequently, these results could help in optimizing the combined use of passive and active cloud observations over the polar areas in specific process and radiation studies (similar to earlier work by Kay and Gettelman, 2009, and Kay and L'Ecuyer, 2013).

The presented validation method can be viewed upon as a step towards a more stringent~~et~~ and universal validation method to be used consistently for cloud climate data records generated from passive imagery (as discussed in Wu et al., 2017). The more than decadal long CALIPSO-CALIOP cloud dataset should be used for benchmarking and for evaluation of current CDRs and future revisions of them. The method presented here could be seen as one candidate method. The ~~possibility~~ ability to derive globally distributed results makes it ~~also~~ easier to define and test global quality requirements ~~on~~ for the CDRs. For example, requirements could be formulated in terms of minimum global coverage within a certain quality threshold instead of today's often ~~very generally formulated~~overly generalised global requirement which use ~~in~~ one finite value or a value range (WMO 2, 2011).

~~A specific problem with the current method is the inability to assess the global quality of products from polar satellites in morning orbits (e.g., from the NOAA-17 and Metop satellites). Matchups with CALIPSO-CALIOP are here only possible at high latitudes leaving low-to-middle latitudes without reference CALIOP observations for AVHRR products. Comparisons have been made for morning satellites at high latitudes (CM SAF 1, 2017) showing good agreement with corresponding results from afternoon satellites. Thus, for cloud amount information (in contrast to some other cloud parameters, like cloud effective radius) there is no reason to suspect large differences between morning and afternoon results even if morning orbit data is partly using measurements in another spectral band (at 1.6 µm) in the short-wave infrared spectral region. However, this has to be further confirmed in the future, e.g., by use of reference data from the Cloud-Aerosol Transport System lidar (CATS, https://cats.gsfc.nasa.gov/) on the International Space Station or by use of data from the Earth Cloud Aerosol and Radiation Explorer (EarthCARE) mission (http://m.esa.int/Our_Activities/Observing_the_Earth/The_Living_Planet_Programme/Earth_Explorers/EarthCARE/ESA_s_cloud_aerosol_and_radiation_mission) for afternoon satellites with two coexisting short-wave infrared channels onboard.~~

One particular aim of this study was to provide a strict definition of the clouds being validated ~~together with~~alongside the main validation results. This has been accomplished ~~by~~ through the use of ~~a combination of~~ the CALIOP-derived cloud mask and the CALIOP-estimated optical thickness of clouds. ~~In this way we believe that results could become~~As a result these validation results are more quantitatively useful. One obvious application would be to incorporate this information about strengths and limitations of cloud detection capabilities into the cloud dataset simulators of the Cloud Feedback Model Intercomparison Project (CFMIP) Observation Simulation Package (COSP, Bodas-Salcedo et al., 2011). Existing COSP simulators for cloud datasets generated from passive satellite imagery (e.g., ISCCP and MODIS) do not explicitly take into account these potential inherent cloud detection problems and instead~~. Instead~~, they concentrate on simulating some satellite-specific or retrieval-specific features (e.g., systematic underestimation of cloud top height of thin high clouds) leaving it to the user of the simulator to add existing knowledge on cloud detection efficiency in the final evaluation process. ~~We are of the opinion that also~~It would clearly be beneficial if aspects of cloud detection capabilities ~~need~~ were to be explicitly ~~taken into~~ accounted for in these simulators. A specific CLARA-A2 COSP simulator is therefore under development where the description of such quality aspects will be included based on the findings of this validation study.

Finally, we repeat our opinion ~~of~~ that CALIPSO-CALIOP data ~~as being~~is ~~a great~~an invaluable asset for the current and future evaluation of cloud CDRs based on passive satellite imagery. At the same time, we must express our concern about the current uncertainty regarding the long-term planning of possible replacements of both the A-Train satellites and the upcoming EarthCARE mission. Without follow-on missions it will be very difficult to assess the critical long-term stability of these CDRs, which in turn ~~means~~ increases the difficulty~~ies~~ in assessing the reliability of ~~possible~~ any climate trends deduced from these CDRs. There is also a need to slowly transform CLARA-type data records to AVHRR-heritage data records, i.e., extend the AVHRR results into the future using results from similar spectral channels existing on other sensors (e.g., the VIIRS sensor on recently launched and future polar NOAA satellites). A continued access to observations from active space-born lidar systems is essential for the development of such AVHRR-heritage data records.

Author contributions

Karl-Göran Karlsson conducted the validation study and wrote the manuscript. Nina Håkansson prepared the calculation and visualisation of the global results.

Competing interest

The authors declare that they have no conflict of interests.

Acknowledgements

CALIPSO-CALIOP datasets were obtained from the NASA Langley Research Center Atmospheric Science Data Center. The authors want to thank Dr. David Winker in the CALIPSO Science Team for valuable advice regarding the use of CALIOP cloud information. The authors are also grateful to David Dufton for constructive comments on the manuscript.

This work was funded by EUMETSAT in cooperation with the national meteorological institutes of Germany, Sweden, Finland, the Netherlands, Belgium, Switzerland and United Kingdom.

The CLARA-A2 data record is (as all CM SAF CDRs) freely available via the website https://www.cmsaf.eu.

References

Barja, B. and Antuña, J.C.: The effect of optically thin cirrus clouds on solar radiation in Camagüey, Cuba, Atmos. Chem. Phys., 11, 8625-8634, doi:10.5194/acp-11-8625-2011, 2011.

Bodas-Salcedo, A., Webb, M.J., Bony, S., Chepfer, H., Dufresne, J.-L., Klein, S.A., Zhang, Y., Marchand, R., Haynes, J.M., Pincus, R. and John, V.O.: COSP: Satellite simulation software for model assessment, Bull. Amer. Meteor. Soc., August 2011, 1023-1043, doi: 10.1175/2011BAMS2856.1 , 2011.

Charlson, R.J., Ackermann, A.S., Bender, F.A.-M., Anderson, T.L. and Liu, Z.: On the climate forcing consequences of the albedo continuum between cloudy and clear air, Tellus, 59B, 715–727, doi:10.1111/j.1600-0889.2007.00297.x., 2007.

CM SAF 1: Validation Report - CM SAF Cloud, Albedo, Radiation data record, AVHRR-based, Edition 2 (CLARA-A2) – Cloud Products, SAF/CM/DWD/VAL/GAC/CLD version 2.3, Available at http://dx.doi.org/10.5676/EUM_SAF_CM/CLARA_AVHRR/V002, 2017.

CM SAF 2: Algorithm Theoretical Basis Document - CM SAF Cloud, Albedo, Radiation data record, AVHRR-based, Edition 2 (CLARA-A2) – Cloud Fraction, SAF/CM/DWD/ATBD/CMA_AVHRR version 2.0, Available at http://dx.doi.org/10.5676/EUM_SAF_CM/CLARA_AVHRR/V002, 2017.

Devasthale, A., Raspaud, M., Schlundt, C., Hanschmann, T., Finkensieper, S., Dybbroe, A., Hörnquist, S., Håkansson, N., Stengel, M. and Karlsson, K.-G.: PyGAC: An open-source, community-driven Python interface to preprocess the nearly 40-year AVHRR Global Area Coverage (GAC) data record, GSICS Quartherly Newsletter, Summer 2017, Special Issue on Re-Processing, 11, doi: 10.7289/V5R78CFC, 2017.

Dowell, M., P. Lecomte, R. Husband, J. Schulz, T. Mohr, Y. Tahara, R. Eckman, E. Lindstrom, C. Wooldridge, S. Hilding, J. Bates, B. Ryan, J. Lafeuille, and S. Bojinski, 2013: Strategy Towards an Architecture for Climate Monitoring from Space. Pp. 39. This report is available from: www.ceos.org; www.wmo.int/sat; http://www.cgms-info.org/

Dybbroe, A., Thoss, A. and Karlsson, K.-G.: NWCSAF AVHRR cloud detection and analysis using dynamic thresholds and radiative transfer modelling – Part I: Algorithm description, J. Appl. Meteor, 44, 39-54, 2005.

González A.: Measurement of Areas on a Sphere Using Fibonacci and Latitude--Longitude Lattices. Mathematical Geosciences. 42 (1), 49-64. doi:10.1007/s11004-009-9257-x, 2009.

Heidinger, A., Foster, M., Botambekov, D., Hiley, M., Walther, A. and Li, Y.: Using the NASA EOS A-Train to Probe the Performance of the NOAA PATMOS-x Cloud Fraction CDR. Remote Sens., 8, 511, 2016.

Heidinger, A. K., Foster, M. J., Walther, A. & Zhao, Z.: The Pathfinder Atmospheres Extended (PATMOS-x) AVHRR climate data set. Bull. Am. Meteorol. Soc. 95, 909–922, 2014.

Heidinger, A.K., Straka, W.C., Molling, C.C., Sullivan, J.T. and Wu, X.Q.: Deriving an inter-sensor consistent calibration for the AVHRR solar reflectance data record. Int. J. Rem. Sens., 31(24), 6493-6517, doi: 10.1080/01431161.2010.496472, 2010.

Jin, Y., Okamoto, J and Hagihara, Y.: Improvement of CALIOP cloud masking algorithms for better estimation of dust extinction profiles, J. Meteorol. Soc. of Japan, 92, 433-455, doi: 10.2151/jmsj.2014-502. 433-455, 2014.

Karlsson, K.-G., Anttila, K., Trentmann, J., Stengel, M., Meirink, J.F., Devasthale, A., Hanschmann, T., Kothe, S., Jääskeläinen, E., Sedlar, J., Benas, N., van Zadelhoff, G.-J., Schlundt, C., Stein, D., Finkensieper, S., Håkansson, N. and Hollmann, R.: CLARA-A2: The second edition of the CM SAF cloud and radiation data record from 34 years of global AVHRR data. Atmos. Chem. Phys., 17, 5809–5828, doi: 10.5676/EUM_SAF_CM/CLARA-AVHRR/V002, 2017.

Formaterat: Teckensnitt:(Standard) +Rubriker (Times New Roman)

Formaterat: Teckensnitt:(Standard) +Brödtext (Times New Roman)

Formaterat: Engelska (USA)

Formaterat: Engelska (USA)

Formaterat: Engelska (USA)

Formaterat: Engelska (USA)

Karlsson, K.-G. and Dybbroe, A.: Evaluation of Arctic cloud products from the EUMETSAT Climate Monitoring Satellite Application Facility based on CALIPSO-CALIOP observations. Atmos. Chem. Phys., 10(4), 1789–1807, https://doi.org/10.5194/acp-10-1789-2010 , 2010.

Karlsson, K.-G., Riihelä, A., Müller, R., Meirink, J. F., Sedlar, J., Stengel, M., Lockhoff, M., Trentmann, J., Kaspar, F., Hollmann, R., and Wolters, E.: CLARA-A1: a cloud, albedo, and radiation dataset from 28 yr of global AVHRR data, Atmos. Chem. Phys., 13, 5351-5367, doi:10.5194/acp-13-5351-2013, 2013.

~~Jin, Y., Okamoto, J and Hagihara, Y.: Improvement of CALIOP cloud masking algorithms for better estimation of dust extinction profiles, J. Meteorol. Soc. of Japan, 92, 433-455, doi: 10.2151/jmsj.2014-502. 433-455, 2014.~~

Karlsson, K.-G. and Johansson, E.: On the optimal method for evaluating cloud products from passive satellite imagery using CALIPSO-CALIOP data: example investigating the CM SAF CLARA-A1 dataset. Atmos. Meas. Tech., 6, 1271–1286, www.atmos-meas-tech.net/6/1271/2013/, doi:10.5194/amt-6-1271-2013, 2013.

Kay, J.E. and Gettelman, A.: Cloud influence on and response to seasonal Arctic sea ice loss, J. Geoph. Res., 114, D18204, doi:10.1029/2009JD011773, 2009.

Kay, J.E. and L'Ecuyer, T.: Observational constraints on Arctic Ocean clouds and radiative fluxes during the early 21st century, J. Geoph. Res. Atmos., 118, 7219-7236, doi:10.1002/jgrd.50489, 2013.

Koren, I., Oreopoulos, L., Feingold, G., Remer, L.A. and Altaratz, O.: How small is a small cloud?, Atmos. Chem. Phys., 8, 3855-3864, www.atmos-chem-phys.net/8/3855/2008, 2008.

Martins, E., Noel, V. and Chepfer, H.: Properties of cirrus and subvisible cirrus from nighttime Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP), related to atmospheric dynamics and water vapor, J. Geoph. Res., 116, D02208, doi: 10.1029/2010JD014519, 2011.

Merchant, C. J., Paul, F., Popp, T., Ablain, M., Bontemps, S., Defourny, P., Hollmann, R., Lavergne, T., Laeng, A., de Leeuw, G., Mittaz, J., Poulsen, C., Povey, A. C., Reuter, M., Sathyendranath, S., Sandven, S., Sofieva, V. F. and Wagner, W.: Uncertainty information in climate data records from Earth observation, Earth Syst. Sci. Data, 9, 511-527, doi: 10.5194/essd-9-511-2017, 2017.

Ohring, G., Wielicki, B., Spencer, R., Emery, B. and Datla, R.: Satellite instrument calibration for measuring global climate change. Bulletin of the American Meteorology Society, 86, 1303–1313, doi: 10.1175/BAMS-86-9-1303, 2005.

Rossow, W.B. and Schiffer, R.A.: Advances in understanding clouds from ISCCP, Bull. Am. Meteorol. Soc., 80, 2261-2287, doi:10.1175/1520-0477(1999)080%3C2261:AIUCFI%3E2.0.CO;2, 1999.

Sassen, K. and Cho, B. S.: Subvisual-Thin Cirrus Lidar Dataset for Satellite Verification and Climatological Research, J. Appl. Meteor., 31, 1275-1285. https://doi.org/10.1175/1520-0450(1992)031%3C1275:STCLDF%3E2.0.CO;2, 1992.

Saunders, R. W.: An automated scheme for the removal of cloud contamination from AVHRR radiances over western Europe, Int. J. Rem. Sens., 7, 867, 1986.

Saunders, R. W. and Grey, D. E.: Interesting cloud features seen by NOAA-6 3.7 µm images, Met. Mag., 114, 211, 1985.

Schulz, J., Albert, P., Behr, H.-D., Caprion, D., Deneke, H., Dewitte, S., Dürr, B., Fuchs, P., Gratzki, A., Hechler, P., Hollmann, R., Johnston, S., Karlsson, K.-G., Manninen, T., Müller, R., Reuter, M., Riihelä, A., Roebeling, R., Selbach, N., Tetzlaff, A., Thomas, W., Werscheck, M., Wolters, E., and Zelenka, A.: Operational climate monitoring from space: the EUMETSAT Satellite Application Facility on Climate Monitoring (CM-SAF), Atmos. Chem. Phys., 9, 1687-1709, doi: 10.5194/acp-9-1687-2009, 2009.

Formaterat: Engelska (USA)

Formaterat: Engelska (USA)

Formaterat: Engelska (USA)

Formaterat: Engelska (USA)

Formaterat: Upphöjd

Stengel, M., Stapelberg, S., Sus, O., Schlundt, C., Poulsen, C., Thomas, G., Christensen, M., Carbajal Henken, C., Preusker, R., Fischer, J., Devasthale, A., Willén, U., Karlsson, K.-G., McGarragh, G. R., Proud, S., Povey, A. C., Grainger, D. G., Meirink, J. F., Feofilov, A., Bennartz, R., Bojanowski, J., and Hollmann, R.: Cloud property datasets retrieved from AVHRR, MODIS, AATSR and MERIS in the framework of the Cloud_cci project, Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2017-48, in review, 2017.

Stephens, G. L., Vane, D.G., Boain, R.J., Mace, G.G., Sassen, K., Wang, Z., Illingworth, A.J., O'Connor, E.J., Rossow, W.B., Durden, S.L., Miller, S.D., Austin, R.T., Benedetti, A., Mitrescu, C., and the CloudSat Science Team: The CloudSat mission and the A-Train. Bull. Amer. Meteor. Soc., 83, 1771–1790, doi: http://dx.doi.org/10.1175/BAMS-83-12-1771, 2002.

Stocker, T.F., D. Qin, G.-K. Plattner, L.V. Alexander, S.K. Allen, N.L. Bindoff, F.-M. Bréon, J.A. Church, U. Cubasch, S. Emori, P. Forster, P. Friedlingstein, N. Gillett, J.M. Gregory, D.L. Hartmann, E. Jansen, B. Kirtman, R. Knutti, K. Krishna Kumar, P. Lemke, J. Marotzke, V. Masson-Delmotte, G.A. Meehl, I.I. Mokhov, S. Piao, V. Ramaswamy, D. Randall, M. Rhein, M. Rojas, C. Sabine, D. Shindell, L.D. Talley, D.G. Vaughan and S.-P. Xie: Technical Summary. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Stubenrauch, C.J., Rossow, W.B., Kinne, S., Ackermann, S., Cesana, G., Chepfer, H., Di Girolamo, L., Getzewich, B., Guignard, A., Heidinger, A., Maddux, B.C., Menzel, W.P., Minnis, P., Pearl, C., Platnick, S., Poulsen, C., Riedi, J., Sun-Mack, S., Walther, A., Winker, D., Zeng, S. and Zhao, G.: Assessment of global cloud datasets from satellites, Bull. Amer. Meteor. Soc., July 2013, 1031-1049, https://doi.org/10.1175/BAMS-D-12-00117.1, 2013.

Sun, B., Free, M., Yoo, H. L., Foster, M. J., Heidinger, A., and Karlsson, K.-G.: Variability and trends in US cloud cover: ISCCP, PATMOS-x and CLARA-A1 compared to homogeneity-adjusted weather observations, Journal of Climate, 28, 4373–4389, doi: 10.1175/JCLI-D-14-00805.1, 2015.

Swinbank, R. and R. J. Purser.: Fibonacci grids: A novel approach to global modelling. Quarterly Journal of the Royal Meteorological Society, 132 (619), 1769–1793. doi:10.1256/qj.05.227, 2006.

Vaughan, M. A., Powell, K.A., Winker, D. M., Hostetler, C. A., Kuehn, R. A., Hunt, W. H., Getzewich, B. J., Young, S. A., Liu, Z. and McGill, M.: Fully Automated Detection of Cloud and Aerosol Layers in the CALIPSO Lidar Measurements, J. Atmos. Oceanic Technol., 26, 2034–2050, doi: 10.1175/2009JTECHA1228.1, 2009.

Winker, D. M., Hunt, W. H. and McGill, M. J.: Initial performance assessment of CALIOP, Geoph. Res. Lett., 34, L19803, doi:10.1029/2007GL030135, 2007.

Winker, D. M., Vaughan, M. A., Omar, A., Hu, Y. and Powell, K. A.: Overview of the CALIPSO mission and CALIOP data processing algorithms, J. Atmos. Oceanic. Technol., 26, 2310-2323, doi: 10.1175/2009JTECHA1281.1, 2009.

WMO 1: Recommended methods for evaluating cloud and related parameters, WWRP 2012-1, Report of the WWRP/WGNE Joint Working Group on Forecast Verification Research (JWGFVR), available at https://www.wmo.int/pages/prog/arep/wwrp/new/documents/WWRP_2012_1_web.pdf, 2012.

WMO 2: Systematic observation requirements for satellite-based data products for climate - Supplement details to the satellite-based component of the "Implementation Plan for the Global Observing System for Climate Support of the UNFCCC (2011 update), GCOS-154, available at https://library.wmo.int/opac/doc_num.php?explnum_id=3710, 2011.

Wu, D. L., Baum, B. A., Choi, Y.-S., Foster, M., Karlsson, K.-G., Heidinger, A., Poulsen, C., Pavolonis, M., Riedi, J., Roebeling, R., Sherwood, S., Thoss, A. and Watts, P.: Towards Global Harmonization of Derived Cloud Products, Bull. of Amer. Meteor. Soc., February 2017, 49-52, doi: 10.1175/BAMS-D-16-0234.1, 2017.

940

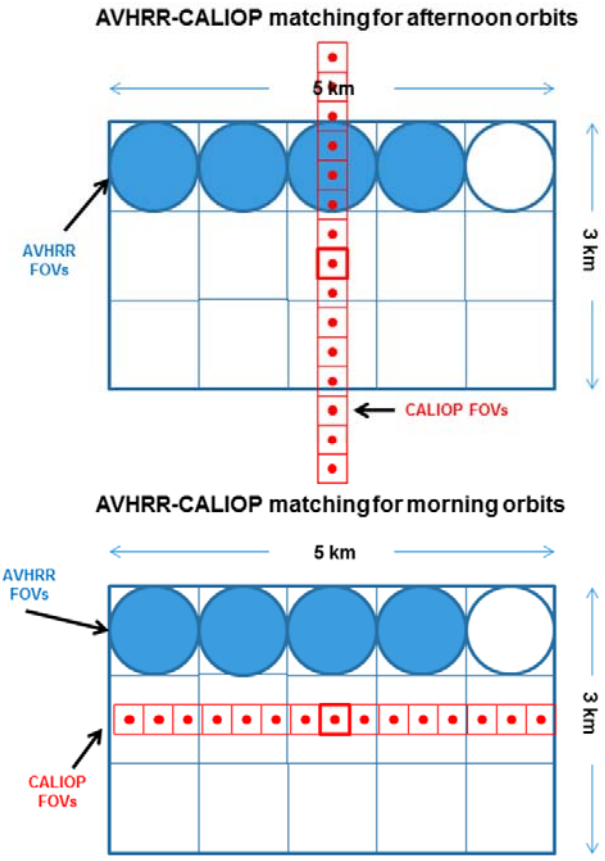


Figure 1: Matchup geometry for perfectly collocated AVHRR GAC and CALIOP FOVs for afternoon satellites (top) and morning satellites (bottom). The GAC FOV is visualized as a rectangle with sides 3 km and 5 km and with individual full resolution AVHRR FOVs represented as 1 km squares. Blue circles indicate actual (more realistic) AVHRR measurements being used. Note that only the blue filled AVHRR FOVs are averaged to represent the full GAC FOV. Red squares denote 15 original nominal 333 m CALIOP FOVs which represent the CALIOP 5 km FOV coverage. The highlighted centre FOV marks the position of the perfect match (i.e., at the center of the GAC FOV). Note that the red filled circles describe actual CALIOP measurements. See text for a more detailed explanation.

Formaterat: Centrerad

Formaterat: Teckensnitt:Fet

Formaterat: Justerat

Formaterat: Teckensnitt:Fet

Formaterat: Centrerad

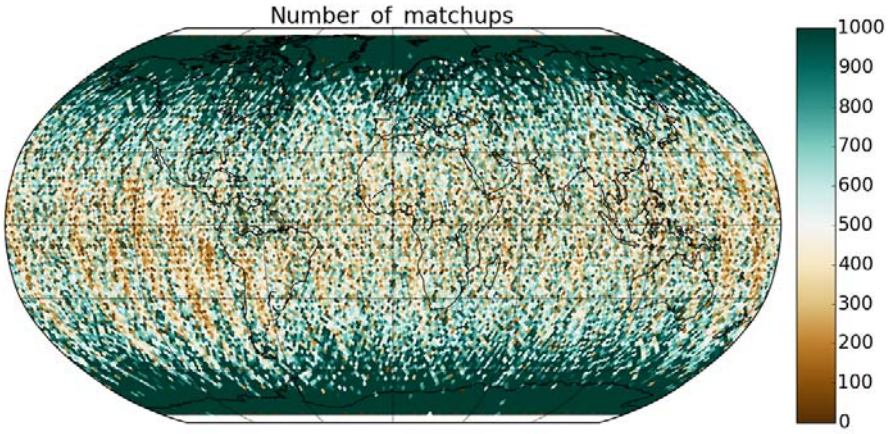
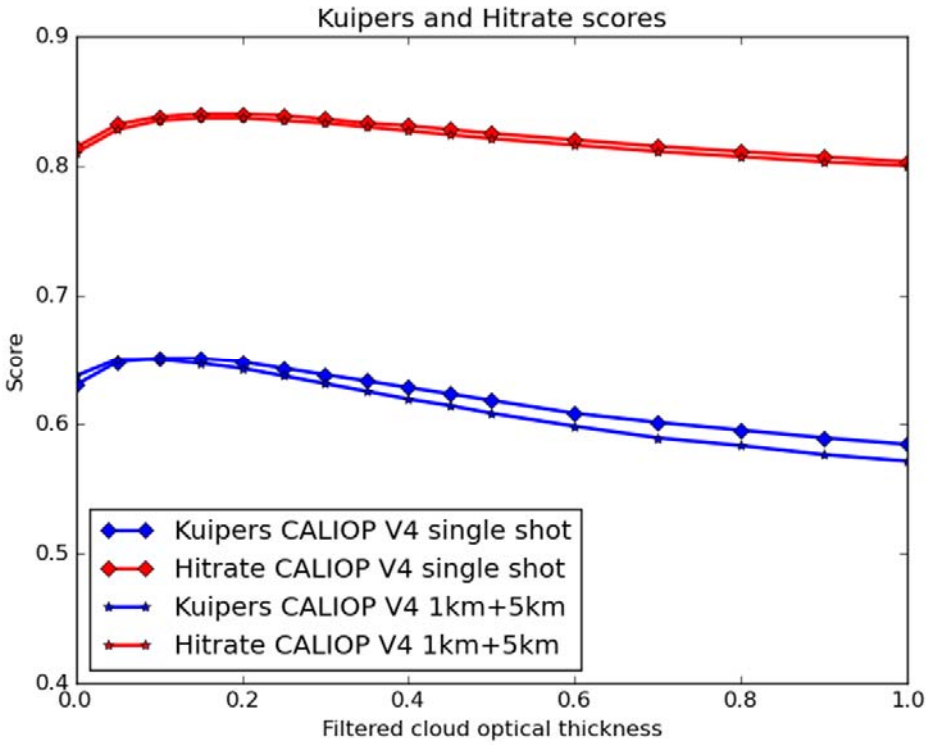


Figure 2: Total number of CALIPSO-CALIOP matchups with NOAA-18 and NOAA-19 AVHRR observations in the time period October 2006 to December 2015. Results are presented in a Fibonacci grid with 75 km resolution.

Formaterat: Centrerad



970

Figure ~~12~~3: CALIOP-based validation scores (Hitrate and Kuipers) as a function of filtered cloud optical thickness (see text for explanation) for 80 matched NOAA-18 orbits between October and December 2006. Validation is based on CALIOP version 4.10 CLAY products and show results from two alternative validation methods (single shot or combined 1 km + 5 km, see text for explanation).

975

980

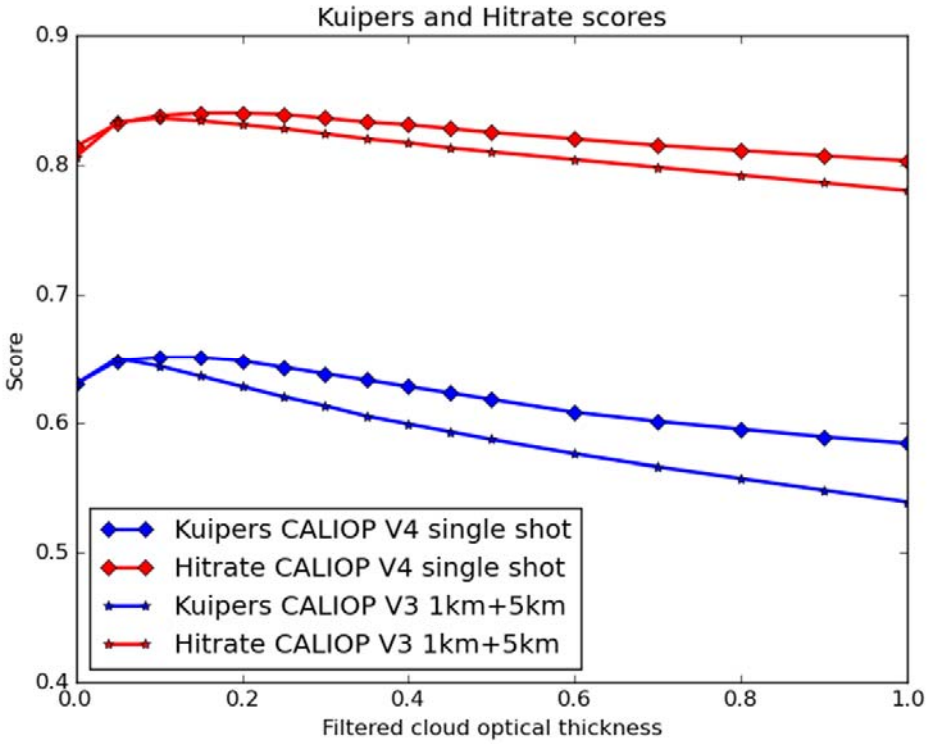


Figure 24: CALIOP-based validation scores (Hitrate and Kuipers) as a function of filtered cloud optical thickness (see text for explanation) for 80 matched NOAA-18 orbits between October and December 2006. The curves compare results based on CALIOP version 4.10 CLAY products computed with the new method based on single shot information (denoted "CALIOP V4 single shot") with results based on CALIOP version 3.01 CLAY products computed with the old method based on combined 1 km + 5 km data (denoted "CALIOP V3 1km+5km").

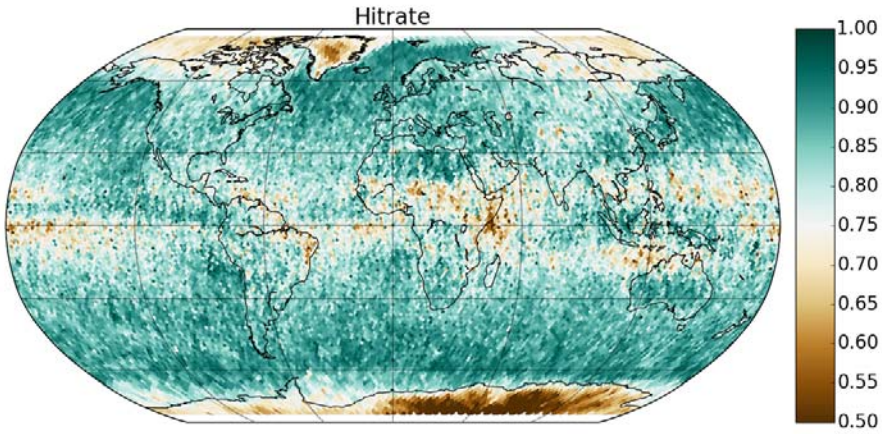


Figure 45: Global presentation of the CLARA-A2 cloud mask Hitrate parameter with a horizontal Fibonacci grid resolution of 75 km. Validation results are based on comparisons with the original CALIPSO-CALIOP cloud mask. Same underlying matchup dataset as in Fig. 23.

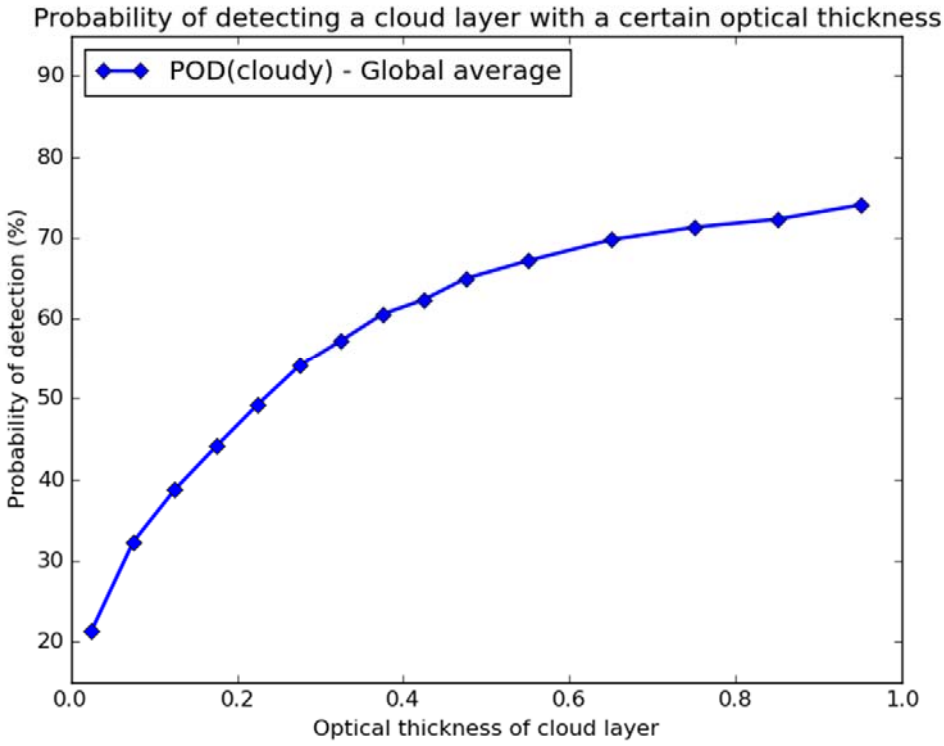


Figure 56: Global estimation of the probability of detecting a cloud with a certain cloud optical thickness. Calculations are based on all available AVHRR-CALIOP matchups over the time period October 2006 to December 2015.

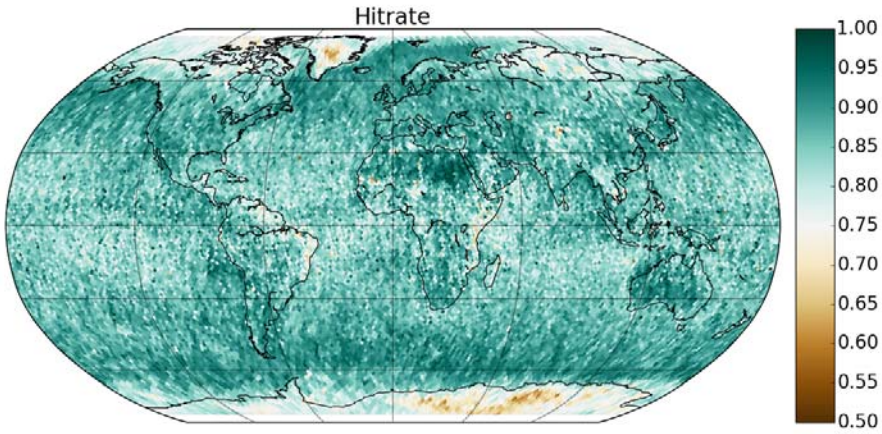


Figure ~~6~~7: Peak Hitrate results for the CLARA-A2 cloud mask achieved after filtering the CALIOP cloud mask with the cloud optical thickness value of 0.225. Same underlying matchup dataset as in Fig. ~~2~~3. Results are presented in a Fibonacci grid with 75 km resolution.

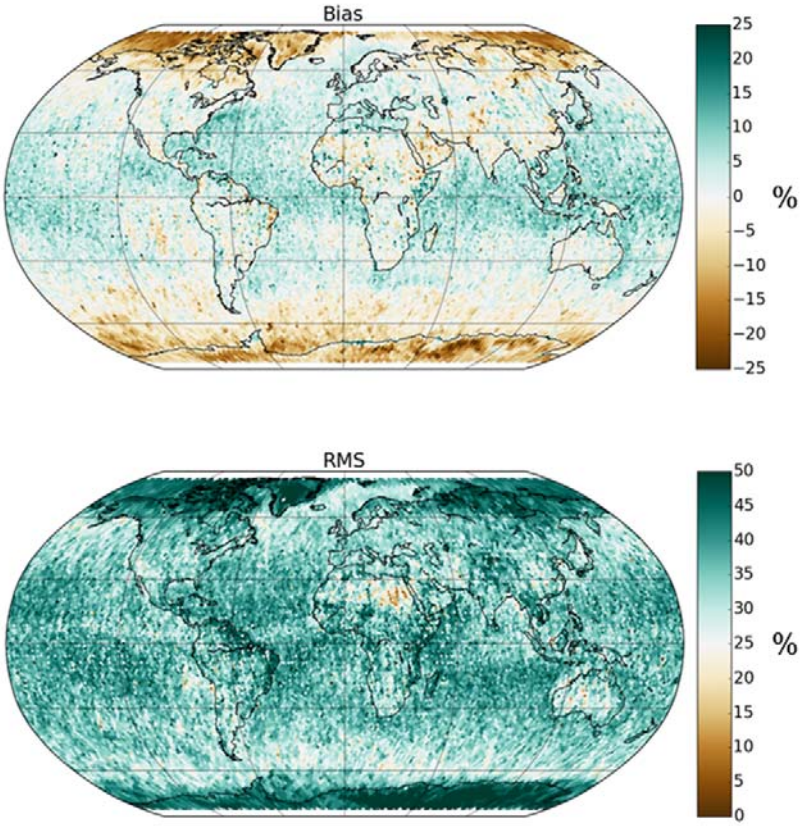


Figure 7~~8~~: Mean Error (Bias) and bias-corrected Root Mean Squared Error (RMS) for the CLARA-A2 cloud amount achieved after filtering the CALIOP cloud mask with the cloud optical thickness value of 0.225. Same underlying matchup dataset as in Fig. 2~~3~~. Results are presented in a Fibonacci grid with 75 km resolution.

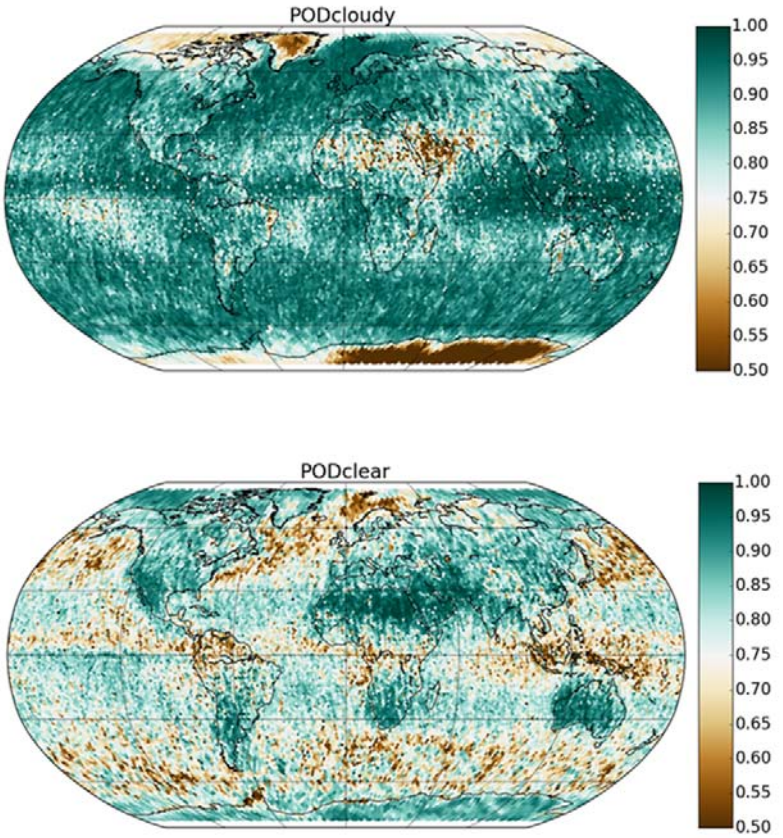


Figure 89: Probability of detection of cloudy (top) and clear (bottom) conditions for the CLARA-A2 cloud mask achieved after filtering the CALIOP cloud mask with the cloud optical thickness value of 0.225. Same underlying matchup dataset as in Fig. 23. Results are presented in a Fibonacci grid with 75 km resolution.

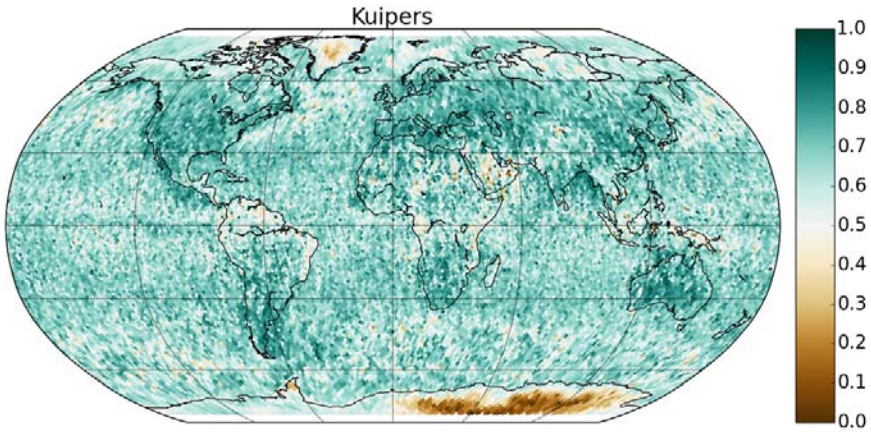


Figure 910: Kuipers score for the CLARA-A2 cloud mask achieved after filtering the CALIOP cloud mask with the cloud optical thickness value of 0.225. Same underlying matchup dataset as in Fig. 23. Results are presented in a Fibonacci grid with 75 km resolution.

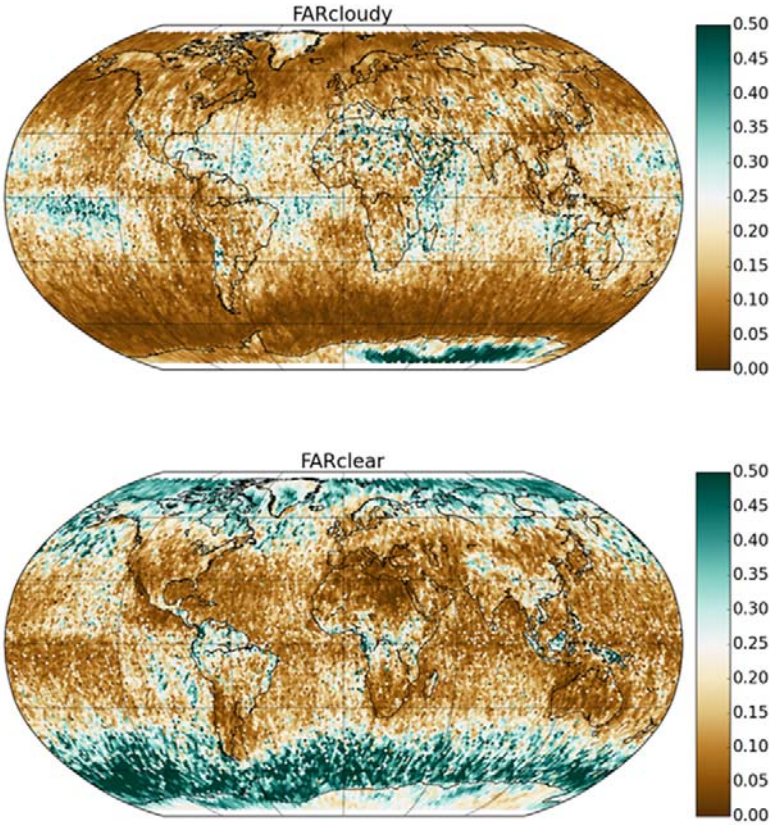
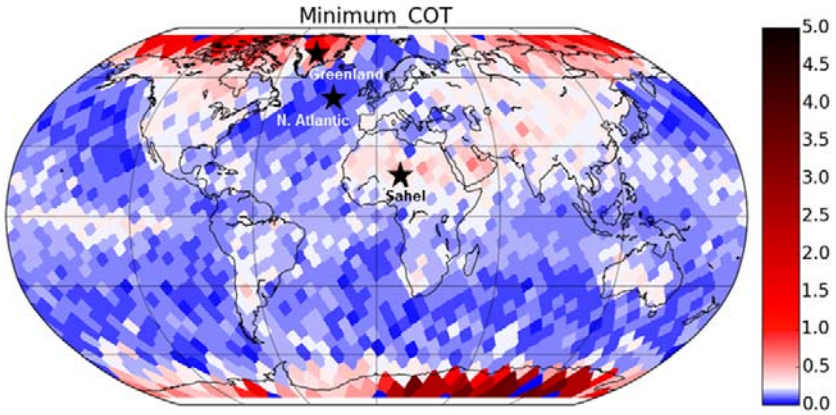


Figure ~~10~~11: False alarm rates for cloudy (top) and clear (bottom) predictions for the CLARA-A2 cloud mask achieved after filtering the CALIOP cloud mask with the cloud optical thickness value of 0.225. Same underlying matchup dataset as in Fig. 2~~3~~. Results are presented in a Fibonacci grid with 75 km resolution.



Formaterat: Centrerad

Figure ~~11~~12: Global map of estimated cloud detection sensitivity of the cloud mask of CLARA-A2 (see text for explanation). Results are calculated from the same dataset as visualized in Fig~~ure~~ 2~~3~~ but in a coarser Fibonacci grid resolution of 300 km. Conditions in the three marked locations (black stars) are analysed further in Fig. 13~~2~~. ~~Grey areas denote areas with values exceeding 1.0.~~Values below the global mean value of 0.225 are coloured in blue shades and values above the global mean value are coloured in red shades.

1085

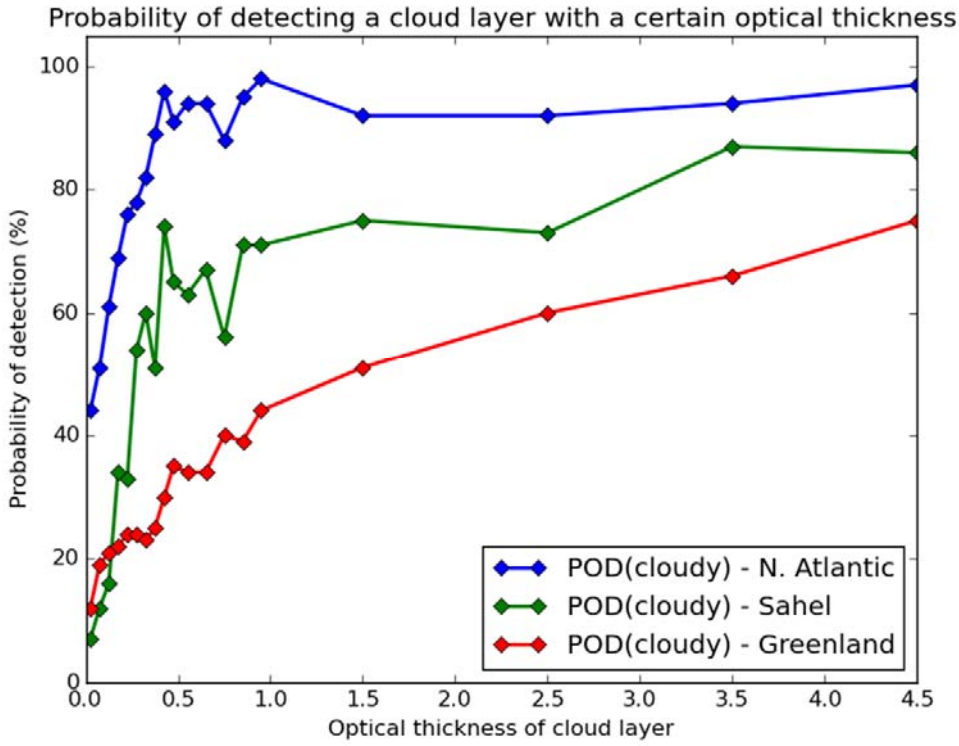


Figure ~~12~~13: Same as Fig. ~~6~~5 but for the individual grid points marked ~~out~~ in Fig. 1~~2~~1.