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).