First 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

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 will do our best to accomplish this. We will reply to the 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.

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 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). 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).

In conclusion, we will add statements to the text based on the above reasoning justifying better why we think the results are still relevant for characterizing the entire CLARA-A2 data record. We can also propose a small change to the title as the following:

"Improved characterization of AVHRR global cloud detection performance based on CALIPSO-CALIOP cloud 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 will do that (with reference to the reply to 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 will follow the recommendation of including both a Discussion section and a Conclusion section.

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 reviewer. We suggest improving the description in three ways:

1. We will introduce a short summary 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 will add a clear illustration (new figure) 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 will expand the discussion of these results in the new Discussion section. 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 (Figure 11) 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 will include some of these clarifications and proposals/suggestions in the new Discussion and Conclusion sections.

FINAL REMARK:

- We will clarify the unclear aspects listed among the short comments

- Finally, thanks for suggested editorial, syntax and language improvements. These are invaluable for non-native English writers like us!