Atmos. Meas. Tech. Discuss., 6, C288–C293, 2013
www.atmos-meas-tech-discuss.net/6/C288/2013/

Atmospheric
Measurement
Techniques
Discussions

Interactive comment on "On the optimal method for evaluating cloud products from passive satellite imagery using CALIPSO-CALIOP data: example investigating the CM SAF CLARA-A1 dataset" *by* K.-G. Karlsson and E. Johansson

K.-G. Karlsson and E. Johansson

karl-goran.karlsson@smhi.se

Received and published: 13 March 2013

We thank the Referee for the nice comments about the manuscript. We will certainly adjust and correct some of the technical and editorial mistakes that are mentioned. Regarding some questions and call for some additional studies we can state the following:

1. On the lengthy discussion of statistical scores

Well, we started this work with the intention to investigate the limits for cloud detection for the applied cloud screening method (PPS) that was used when composing the

CLARA-A1 dataset. But at that time it was very very unclear which statistical measure that was most useful for determining this limit. Little guidance on this existed, probably since no really powerful "ground truth" or reference information existed before active A-train sensors were launched. So, we thought that we'd better use a set of different scores and during the course of investigating/writing make conclusions on what was the best approach. Especially, we wanted a measure that most clearly showed the impact of the optical thickness of clouds for the success of cloud detection. Initially, one kind of assumed that quantities like the Mean Error or the Hitrate would be the most useful ones but it turned out that these are very much dominated by cloud detection problems which have nothing to do with the thickness of clouds being observed but which rather depends on non-separability issues (like separating clouds from very cold ground surfaces or very bright land/desert surfaces and snow). Also, the Kuipers score is even more sensitive to these aspects. So, in the end it turned out that the POD and FAR quantities showed the most obvious sensitivity to the cloud optical thickness property and we decided to use them.

We think that since there is no clear guidance on what to choose as the best measure of cloud detection sensitivity from previous studies, it could be fruitful to keep the description of all those scores as it is in the manuscript. We really think that to keep it there gives a good illustration of the behaviour of these scores and, especially, the differences from each other. It clearly shows how different validation scores emphasize different things and that there is no "perfect" score that "does it all". Also, it illustrates the very important conclusion that the performance of a cloud scheme does not only depend on how thin or thick clouds are. A substantial part of mis-classifications comes from non-separability issues. We think it is important to illustrate both aspects so therefore we wish to keep the description of all validation scores.

2. Arbitrary choice of 1 % threshold

We admit that the 1 % threshold for the (POD+FAR) rate of change as a function of filtered cloud optical thickness is arbitrarily chosen. However, one can also see this as our own definition of the cloud detection limit. Since there is no general guidance on this (see what was written above) we made our own definition and this was the most sensible one, we thought. We noticed that above a certain filtered cloud optical thickness the POD and FAR quantitites reached a kind of "saturation level" where a further increase in the filtered cloud optical thickness would not mean much. So, the achieved minimum cloud optical thickness value here determines the level where the scheme has reached its optimal performance (our interpretation). Below this threshold more and more clouds are being missed as we decrease the filtered optical thickness. So, we wanted to have a measure that well enough described where the rapid increase/decrease in POD/FAR quantities at very low cloud optical thicknesses ceased. For that we needed a threshold that was not too small since the "saturation level" is not really a constant value (some small changes occur also at higher optical thicknesses). In that respect the suggested threshold 0.2 % will be too low since it gives a value that is "too far" away from where the real action is (i.e., where rate of changes are rapidly changing). We will discuss the use of different thresholds further in the manuscript but we would once again state that the choice of threshold is more a matter of definition than a real critical issue.

3. Another CFC threshold than 50 % for CALIPSO 1 km data

The Referee asked us to test what it means if the threshold is set to 30 % or 70 %, respectively. We have made such a test and concluded that changes are rather small. A lower threshold means that CALIPSO cloudiness increases and vice versa. If we look at the case when we use the threshold 30 % (meaning more clouds for CALIPSO) we get a change of the Mean error (in the unfiltered case) from - 14.4 % to - 17.9 %. Also for POD(cloudy) the changes are in the order +/- 2-3 % for the threshold alternatives 30 and 70 %. But, what is more important is that the shape of all curves are not affected in any significant way (i.e, just slightly shifted up and down) which means that we get more or less the same values for the cloud detection limit as before. So, results are in this respect rather robust. We will report these additional results in the final manuscript.

4. Regarding added or removed clouds (page 1103, line 26)

We discuss two different things here. We add clouds to the 5 km CALIPSO dataset if more than 50 % of the 1 km columns are cloudy in the case when 5 km data reports no cloud layers. We see this as a problematic flaw of the 5 km dataset that should be corrected.

Regarding the cloudy FOVs that are removed from 5 km data we are just looking at all the cases when less than 50 % of the 1 km columns are cloudy in the case when 5 km data has a cloud layer. This is a more questionable case (quite separate from the first being mentioned). It could be so that the 5 km CALIOP analysis has found a very thin cloud layer as a result of just a fractionally covered 5 km FOV (as the 1 km data is suggesting) instead of a completely filled view. But we cannot be 100 % sure (since the other tre 1 km columns could theoretically also be cloud filled but here the cloud could be too thin to be detected). In this case we chosed to rely more on the 1 km dataset than the 5 km one, and this is the thing that could be questionned. But we have to make a choice here so we do it and mention the problem in the text. This illustrates quite well that it is not easy to get a consistent picture of results in 1 km and 5 km datasets. We hope that furter reprocessing efforts in the future might improve the consistency between the various CALIOP resolutions. (By the way, the sensitivity studies reported in 3 above shows that this issue does not seem to be very important for the end results - the latter seem to be rather robust whatever CFC threshold in CALIPSO data we use).

5. The definition of FAR (cloudy/clear)

We maintain the view that FAR quantities are correctly defined. The false alarm rate should given an estimation of the frequency at which a method (forecast or scheme), which produces two different answers (yes or no), gives the wrong answer. So, it should deal with the fraction of (PPS) cases when the Yes answer (in our case PPS saying Cloudy) is wrong (CALIPSO says clear). The measure as such is looking exclusively at the PPS cases (of the answers yes or no) and NOT at the CALIPSO cases since we want to describe the performance of the method (PPS) at these specific conditions (yes or no). Therefore we think the Referee is wrong when saying that the reference should be the CALIPSO observations. The situation is completely the opposite for the POD case. Here we want to know how efficient the scheme (PPS) is in detecting all the cloudy or clear cases (with reference to CALIPSO). So, this distinction between the FAR and POD quantities is important. But, the referee might be correct in that the wording could be confusing here (it is, indeed, easy to mix up the two cases). We will reconsider the formulations here to be as clear as possible in the final paper.

The Referee questions the case in Figure 8 where the property FAR_cloudy decreases with decreasing optical depth. However, we think it is quite clear what happens: The false alarm rate for PPS saying it is cloudy should increase with increasing filtered optical depth since all of the cases below this optical depth will now be considered as cloud-free. But PPS does detect some of the filtered clouds and thus, these cases will now be reported/treated as falsely detected clouds and the FAR(cloudy) values would consequently increase. So, we maintain that this description is correct and consistent. But again, we will again reconsider formulations to see if we can make things even more clear.

6. FAR and POD versus d(FAR)/d(optical depth) and d(POD)/d(optical depth) (page 1112 line 24).

We do not fully understand the comment here. Even if we understand that figures could be plotted alternatively (although leading to the same conclusions) we still think there is a value in showing the absolute values of the POD and FAR quantities as a function of the filtered cloud optical thickness in the figures. To add figures plotting instead the gradient as a function of cloud optical thickness for each POD/FAR quantity could of course be done but would not add much in our opinion. We also would like to point out that we are not plotting a continous function, i.e., results as a function of any filtered cloud optical depth. We selected only some discrete values of optical thickness (16

values) since a full continuous analysis would be too resource demanding. Therefore, the gradients are still rather roughly described and the alternative plotting would then not add much in the details. We hope it is acceptable to maintain the current figures.

7. Cloud detection during twilight conditions

The recommendation to "ignore the solar channels" during twilight is understandable and seemingly logical but it is not realistic in practice. The reason is that one of the most important (for PPS) channels is the 3.7 micron channel (channel 3b) of AVHRR. This channel cannot be considered a pure infrared channel since it also recieves radiation from reflected sunlight during daylight hours. It means that this channel changes its appearance when the sun is setting. Thus, we cannot use the nighttime approach if there are still contributions from some reflected sunlight remaining in the signal. This fact rules out the possibility to just switch on the nighttime scheme during twilight. This is most critical for low-level water clouds which are the clouds that reflect most efficiently the sunlight. In practice it meanst that these clouds would "appear too warm" to be a cloud if there is any small contribution from reflected sunlight. Such a cloud would risk to be interpreted as a cloud-free pixel. Thus, new errors are introduced. The alternative to just run with AVHRR channels 4 and 5 (at 11 and 12 micron) during twilight is not an option. This will indeed give PPS results three systematically very different behaviours and would risk to mess up any cloud climatology (due to the extra inhomogeneity in results). Therefore, our approach will be to try to keep to the current situation and to find ways of mitigating the problems during twilight. To repeat: There is no seamless scheme that easily is capable of handling the transition from day and night. The typical nighttime scheme will not produce the typical nighttime results in twilight situations and vice versa for the typical daytime scheme.

Interactive comment on Atmos. Meas. Tech. Discuss., 6, 1093, 2013.