Review of AMTD paper by A. Werkmeister et al. entitled

”Validation of CM SAF cloud fractions: Can cloud cover be reliably derived by satellite data at Hannover, Germany and Lauder, New Zealand? – a comment"

General impression

This paper addresses the prospect of replacing manual (human) observations of cloudiness (SYNOP) with automatic measurements or observations from Hemispheric Sky Imagers (HSI) to be used as references for the evaluation of satellite-observed cloudiness. The subject is certainly a very important one in the light of the current accelerated loss of the traditional SYNOP observations, also considering the historic length of SYNOP observation records. Thus, it is not only a matter of getting a proper replacement but also to find ways of continuing the observation record forward in time without introducing artificial discontinuities. The study is using HSI observations from two geographical positions, one at Hannover, Germany and one at Lauder, New Zealand. Observations are compared to traditional SYNOP observations as well as to two satellite-based datasets generated by the EUMETSAT Climate Monitoring Satellite Application Facility (CM SAF).
Unfortunately, the study is found to have several questionable components and is therefore proposed to undergo a major revision before a manuscript can be accepted. Some of the main problems of the study and the manuscript are the following:

- Too short study periods to be able to make reliable conclusions
- The manuscript contains inadequate and frequently incorrect references to other publications
- Some statistical scores are used in an incorrect way.
- The data processing and the methodology are not described with enough details.
- The used satellite datasets are not correctly described.
- The interpretation of the deviation between the different observation sources lacks a complete view (e.g., no discussion of effects of different viewing geometry/conditions)
- The text needs some significant editorial efforts (including language checks and an attempt making it more focused and consistent).

In the following, these aspects will be described in detail.

Comments, questions and critical remarks:

Title:

- The title is over-ambitious and partly questionable! Considering that also the ground-based observations at the mentioned locations have their weaknesses or, at least, individually unique characteristics, it is not realistic to believe that the question will find a very clear answer. More critical – the current study uses a very short period in time for the analysis and it is therefore impossible to make any general very confident conclusions. Also, the title is language-wise rather complicated (i.e., three parts that are not evidently linked). A simpler and more direct title would be preferred (perhaps "Validation of CM SAF cloud fractions using manual and automated surface observations in Hannover, Germany and Lauder, New Zeeland").

Abstract:

- The abstract text needs most probably a revision after taking into account several of the aspects mentioned below in subsequent comments. For example, the meaning of the sentence on line 8 starting "The standard deviation…" is very unclear and needs to be formulated in a more concise way.

1. Introduction:

1. Page 11147, line 2: The conclusions from Clement et al. (2009) are poorly formulated. To state that low-level clouds act as a positive feedback is incorrect. I suppose the authors mean that changes in low-level cloudiness act as a positive feedback? Furthermore, the next sentence mentioning the relation between changes in cloudiness and changes in temperatures seems rather trivial (it is a well-known feature).

2. Page 11147, line 21: The sentence here is rather remarkable ("The study of clouds has been conducted already in the last decades"). The fact that manual observations of clouds have been going on for more than 100 years on many surface stations indicate that this parameter has been of interest much longer than just the last decades. This sentence (and the following paragraph) should be reformulated.

3. Page 11148, line 6: The findings about the character of human observations during day and night, as expressed by Boers et al. (2010), are not correctly interpreted here. A too simplified view is given.
   Boers et al. (2010) just state that the human observer reports less occurrences in the 1-2 octa category during day compared to during night . This differs from other remote sensing instruments giving the opposite picture. Also, the occurrences in the 7 octa category appear to be higher during day for the human observer. But it is also higher for the other instruments, although not as much as for the human observer. However, from this you cannot with certainty conclude that "the observer is mostly underestimating cloud coverage during the day and overestimating at night". I mean, what is the truth? We really do not know!

   In addition, you need to accumulate the results for all octa categories to get a complete picture. When doing that the pattern is less obvious (Figure 7 in Boers et al, 2010). For example, octa categories 3-7 are always higher during day for the human observer but octa categories 0-2 and 8 are always higher during night. Then, what is the net effect? From Figure 7 the net effect seems to be close to zero, thus it is not possible to make the conclusion stated here. Reformulate!

4. Page 11149, lines 9-11: The description of the investigated algorithms is incorrect. You have mixed up algorithms for SEVIRI and AVHRR – they are different which should be made clear. The AVHRR algorithm (denoted NWC SAF PPS 2010) is described by Dybbroe et al. (2005) while the SEVIRI algorithm (denoted NWC SAF MSG cloud software version 2010) is described by the two papers by Derrien and LeGleau (2005 and 2013). Notice also that the CM SAF adoption of the SAF NWC PPS algorithm in the climate dataset CLARA-A1 is described by reference Karlsson and Stengel, 2012. Thus, this reference should rather be used later in section 2.1.5 than here. Please correct!

2. Data retrieval and processing:

1. Page11152, lines 9-14: I am quite surprised that a geographic distance of 10 km between the SYNOP and HSI observations is considered to have no impact on cloud observations. Sharp gradients in cloud climatologies exist in many regions (e.g., near cost-lines). Especially, since there should also be directions where the HSI instrument is partly blocked close to the horizon (e.g. as seen in Fig. 2, lower right portions of HSI images) you should see some deviations. You have to illustrate this conclusion more convincingly (i.e., add evidence in plots).

2. Page 11152, lines 15-27 + Page 11153 lines 1-6: This paragraph is again full of incorrect references to the various algorithms for AVHRR and SEVIRI. Check item 4 above in the Introduction part and correct!

3. Page 11154, line 7: The reference Karlsson et al., 2009, is describing the validation of AVHRR products. Thus, it has nothing to do with SEVIRI data (which is the subject of this paragraph)!

4. Page 11154, line 9: AVHRR is a sensor, not a satellite!

5. Page 11154, line 14: Why do you not use data from NOAA-17? It should be included in the CM SAF dataset.

6. Page 11155, line 8: The GAC resolution is approximately 4 km! The GAC pixel results from averaging 4 out of 5 original HRPT pixels (1.1 km at best) and skipping 3 lines.

7. Page 11155, last lines: Very unclear if the time matching of AVHRR data and HSI observations is correctly done. A full GAC orbit takes about 102 minutes so what overflight time did you actually use? Please confirm that you have extrapolated the time using the line numbers since basic Level 2 products give only the time for the first scanline in the filename.

8. Page 11156, lines 1-5: Your description here is very confusing. This paragraph is said to describe AVHRR data processing but you are here definitely describing how you treat SEVIRI data. In addition, it is very unclear for the reader whether you have used original SEVIRI data from CM SAF defined in the 1185x1185 grid (with resolution approx. 15 km) or if you have used original SEVIRI Level 2 data (in original resolution 4.5 km at this location) and calculated the 13.5 km average yourself. This paragraph indicates the latter but in other parts of the text you can get a different interpretation. Please be clearer in your descriptions.

3. Methodology:

1. Page 11156, lines 15-25: I don't understand why you use both the mean absolute deviation and the standard deviation. They give at least partly the same information. Why are you not using the Mean Error or Bias? For me a Bias estimate without a sign is useless. The sign should tell you whether you underestimate or overestimate a parameter (the accuracy) and the standard deviation tells you if the estimate is uncertain (the precision). Notice that I here talk about the standard deviation of differences which is equal to the unbiased (bias-corrected) Root-Mean Squared Difference. Please reconsider the choice of parameters.

2. Page 11157, lines 10-15: I think you have made a serious error here in the definition of the POD and FAC parameters.

   - Firstly, why do you use notation FAC when you use the most common term False Alarm Rate (FAR) in the text? You should use FAR which is the most common notation for this quantity.

   - Secondly, the observation that in your results POD+ FAC=100 % shows that you have not understood the meaning of these quantities (alternatively, you mean something else).
   The Probability of Detection is a measure trying to estimate the fraction of all <u>observed</u> cases of a certain category (here, either Cloudy or Clear) where the <u>prediction</u> (of Cloudy or Clear) is correct. Thus, the denominator of Equation 5 must be the total number of observations of <u>either</u> Cloudy or Clear (notice, not the sum of all observations).
   The False Alarm Rate is a measure trying to estimate the rate of failure for the <u>predictions</u> of Cloudy or Clear. Thus, the denominator of Equation 6 must be the total number of predictions of either Cloudy or Clear. Thus, since you have different denominators in the two equations you do not get automatically 100 % if you sum them up.

4. Results:

1. Page 11158, lines 1-5: I am pretty convinced that the difference you see between SEVIRI and HIS may be explained by natural reasons which have nothing to do with any potential 'failure' of the satellite estimation. More clearly:

   Differences in the projected area of observation for the two data sources may explain a large part of the observed differences. For example, the satellite observation is constructed from several pixels each having approximately the same resolution and viewing angle while the HIS observation is not observing the area in the same 'linear' sense (the near-zenith portion of the area gets a higher weight than the portions with higher viewing angles closer to the horizon).

Thus, my conclusion is that you should not expect a very good agreement if you are not attempting to try to correct for differences in viewing conditions. So, that SEVIRI is "unable to detect the same changes in CFC" is not surprising and cannot really be concluded as a clear limitation of the SEVIRI retrieval. It can rather be a sign of that the HSI observation is made with higher weight to the local vicinity of the observation point while the SEVIRI observation is representative of a wider or larger area.

2. Page 11158, lines 6-10 (+ many occurrences in the entire Section 4): The incorrect definition of POD and FAC quantities (see item 2 above) makes this discussion difficult to follow and the conclusions made here may therefore be questionable. *(For this reason I cannot really give any comment to the text on pages 11159-11161 – thus, the entire Section 4 needs to be revised).*

3. Page 11162, lines 1-9: No reader has a chance to follow the discussion here unless you describe what you mean with the Cloud Contamination Factor (CCF)! This is not done. You need to describe that the original CMa results says that a cloudy pixel can be labelled as either cloud-filled or cloud-contaminated. Thus, the CCF discussion only concerns the latter fraction of all the cloudy pixels. Furthermore, CM SAF is currently NOT applying a CCF value (i.e., assuming 100 % cloud coverage also for the cloud contamined CMa category).

4. Page 11162, lines 9-25 + Page 11163, lines 1-27: The lack of details on how the CCF has been applied (according to the previous point) leads to the suspicion that you have applied a CCF factor to ALL cloudy pixels instead of only to pixels labelled with the category Cloud Contaminated. Such a correction would be scientifically doubtful. I mean, as a first approach, one should trust the information about the pixels labelled as completely cloud-filled before applying other corrections. The topic as such is quite interesting but the description of the methodology must be clearer here to really understand what you have done.

5. Conclusions and discussions:

1. Page 11164, section 5.1: Even if I agree with the authors that a stable and consistent machine-operated measurement is to prefer compared to the SYNOP observation, this study is not convincing enough to show what error characteristics that we can expect when comparing to HSI observations. The study period is far too short to give a full picture. If staying with this in a revised version of the manuscript, it should rather be described as a preliminary study or a case study.

Editorial remarks:

- Page 11149, lines 16-20: Just make clear that you are discussing the validation of cloud fraction here (the CM SAF validation reports cover a wide range of parameters).

- Page 11149, line 18: Abbreviations INST, DM and MM are used without being first explained.

- Page 11150, line 14: Word “resilience”!! Not correct here. I guess you mean “resemblance”.

- Page 11150, line 16: “…the authors will describe…”. Bad language in this context. Sounds like you are not responsible for this text! Just write “In the next section, we will describe….”.

- Page 11164, lines 24-26: Unfinished sentences. Please reformulate (merge).

- Page 11176, Fig 2 caption: The term ACS is never explained (assumed to be HSI?).