Review of AMTD paper by Grossi et al. entitled

„Total column water vapour measurements from GOME-2 MetOp-A and MetOp-B''

General

This paper introduces an improved version (version 4.7) of the GOME Data Processor (GDP) algorithm for deriving total column water vapor (TCWV) from observations of the GOME-2 instruments aboard EUMETSAT's MetOp-A and MetOp-B satellites. The GDP 4.7 water vapor algorithm is used by the EUMETSAT's Satellite Application Facility on Ozone and Atmospheric Chemistry Monitoring (O3M-SAF) for the operational generation of GOME-2 TCWV products and improves upon the previous version by the introduction of an empirical correction of scan-angle dependency. The TCWV estimates derived from GOME-2 on Metop A and B are intercompared for the overlap period of 8 months. Additionally, TCWV estimates from both GOME-2 instruments are compared to other satellite-based TCWV estimates and a reanalysis product (ECMWF Interim Reanalysis, ERA-interim).

The topic and the results presented are in principle interesting and well suited for AMT. There are several issues, however, due to which the manuscript is proposed to undergo a major revision before it can be accepted. Major points of criticism:

- Imprecise and misleading usage of terms,

- "real" validation (involving collocated – in space and time - data pairs) using ground-based measurements such as radiosonde observations is missing

- Comparison methods applied differs between comparison data sets used which makes it difficult to compare and interpret the results:

    More details are given below.

Major comments

1. Imprecise and misleading usage of terms

The imprecise and misleading usage of terms makes the text hard to follow. For example, the terms "validation", "comparison", "cross-comparison" are used without any clear distinction. "Comparison" is surely the most appropriate term for the results presented so far. The term "validation" should be reserved for comparisons against a reference of known accuracy, here for example ground-based observations such as radiosonde observations, based on collocated (in time and space) data pairs.

The term "collocated" is also often used in an imprecise way. Both for data pairs matched in time and space (SSMIS) and when just referring to a grid box by grid box comparison of averages (ERA-interim).

Within the text several terms are also used to describe the same thing. This leads to confusion as it implies that there are some distinctions where there may really be none. Therefore better only use one.term consistently e.g. H2O columns, H2O VCD, TCWV, WV are used to refer to total column water vapor estimates and ECMWF data, ERA-Interim data to refer to the ERA-interim data set. For the latter, ERA-interim would be more suitable as that's the official name of the data set.

Please be more precise and use clear definitions. Additional work concerning the language/phrasing is also needed. (see Editorial Remarks) in order improve the readability.

2. **"real" validation (involving collocated – in space and time - data pairs) using ground-based measurements such as radiosonde observations is missing**

The authors aim at performing a " sound assessment of the quality of the (presented) satellite products" but the necessary validation using ground-based observations such as radiosonde data as a reference is missing. I assume this kind of validation was done but it is not clear why it is not added to the manuscript. Please clarify! Such a validation would ideally include both previous and improved algorithm version. The presented comparisons against other satellite-based and reanalysis products are still very useful and valuable to check the consistency of the spatial patterns and can serve as an additional means to check the performance of the product.

3. **Comparison methods applied differs between comparison data sets used which makes it difficult to compare and interpret the results**. So far for the three comparison data sets three different comparison methods are applied:

   - Era-interim : daily mean (Era-interim) vs. single day-time measurement, on 1.5° grid

   - SSMIS: collocated data pairs (5 a.m. vs. 9:30 a.m.) , no information about regridding

   - SSMI-MERIS: Level 3 daytime only, but for SSM/I no cloud screening is applied, 0.5 ° grid

The comparison should be made more consistent by using same gridding and same filtering, e.g. using not only spatially but also temporally matched observations

(like for the SSMIS comparison) which would allow to apply a common (GOME-2 like) cloud screening to all comparison datasets.

In any case, the statistics should be calculated separately for ocean, (maybe also coasts) and land parts (for Fig. 7 and Table 2). So far the bias time series (Fig. 7) and statistics (table 2) are not comparable as the data sets have different spatial coverages (ocean-only, land+ocean).

Minor comments, questions, critical remarks:

P 3022,L6: "long-term stability": the term seems inappropriate as the data period studied covers 6.5 years only which not really be considered to be long long-term. Reformulate.

P 3022,L 28: "independent": As 2 of the 3 comparison datasets are also satellite-based products using similar types of instruments (SSM/I and SSMIS) and SSM/I and SSMIS radiances are also assimilated in ERA-interim Therefore, the comparison data sets are not completely independent.

P3024, L13: "only satellite observations": what about reanalysis data sets, weather models..?

P3025, L 26: "input quantity": input to what?

P3025, L27f: "useful for numerical weather predictions": how?   For assimilation? Give more details.

P3025, L 28: "in contrast to other satellite data sets": What about HIRS, IASI, AIRS, ATOVS based data sets? These are also available over land and so far completely missing in your list of existing satellite-based water vapor products.

P3026, L 9: "pointed at large deficiencies": Also for GOME-2 retrievals? Give more details about the results! Which retrieval was used for GOME-2, a previous GDP version?

P 3028, L 7: "monthly narrow-swath day": does this refer to the change to narrow swath of GOME-2A ? Sounds like it happens only once per month? Reformulate.

P3036L17: "monthly mean distribution": you mean you look at the spatial distribution of the bias, right? Reformulate, in order to be more clear.

P3038,L7: "low statistics": why don't you use more data for the GOME-2 intercomparison? You have in total 8 months of overlap.

P3038,L19-22: Here you have to give more details otherwise it is not clear what you want to say. What is the PRD document? How is the optimal accuracy threshold defined and which value does it have?

P3039,L17: what do you mean with fractional difference?

P3039,L22 "smaller number of data points due to the cloud selections": why don't you use more months then (see also comment above)? You have in total 8 months of overlap.

P3039,L24: Change title to "Comparison result" (see Major comments)

P3039L25f: "A sound assessment of the quality" and "validation" would imply for me the comparison against ground-based measurements like radiosonde data (see Major comments). Reformulate!

P3040,L5: "at all available time steps" what do you mean?

P3040, L23: "all validations" => all comparisons

P3010,L28ff: The results in Fig. 7 are presented before the data sets and the methodology are properly introduced. I suggest to create a new "comparison data sets" subsection (5.1) within section 5 where all comparison data sets are described.

P3040,L28ff : The bias timeseries are not comparable as the data sets have different spatial coverages (ocean-only, land+ocean). See major comments above.

P3043, L12: "forecast 12h values"? does this refer to the 12hour forecast based on 00 and 12 UTC analysis? Why do you use the forecasts and not the analysis data (6-hourly resolution)?

P3044,L2: "good accuracy": I would call it good agreement.

P3044, L13: "very low number of collocations"? you compare monthly means, so probably you mean that the GOME-2 monthly mean is based on 8 measurements only? Anyways, striving for a more consistent approach to compare the different datasets (spatial and temporal matched, as done for the SSMIS comparison) would increase the comparability.

P3044,L13ff: "also problems of the ECMWF..": give more details on input data, preferably also already in the "comparison data set" section"

P3044,L18: "regions without..", you mean grid boxes or pixel without severe cloud cover, as the cloud screening is applied on the Level 2 estimates?

P3044,L21-P3045,L3: as far as I can see from the maps the region you refer to is actually not the Sahara desert but rather covers the Sahel zone and areas even more south, where the month of August forms part of the rainy season. That's why the high surface albedo discussion does not apply. Reformulate!

P3045,L13: move SSMIS related text general "comparison data sets" section as stated earlier.

P3046,L3ff: "if we evaluate..": does not make sense to me. Maybe you mean "mean SSMIS" data instead of "mean ECMWF" data? Reformulate!

P3046,L5ff: "as for ECMWF, we select only daily co-locations": the term co-location in the context of the ECMWF comparison is confusing. You compare monthly means at each grid box. Anyways what you do for SSMIS is completely different compared to ECMWF (or better ERAinterim), here you use the descending F16 orbits only (i.e. Level 2 data), for ECMWF you use daily averages. In general it would be best to make the comparison methodology among the different data sets more consistent (to reduce sampling related issues)!

P3046,L10: "5.08p.m." you probably mean a.m.

P3046,L10: "drifting": why don't you use F17 then instead which is drifting less and has a more close overpass time?

P3047,L2: But in case of intense rain, there will be no valid TCWV estimate and therefore no SSMIS-GOME-2 match-up pair, so these cases do not influence your results, they only decrease the number of available match-ups.

P3047,L10: "1% per 1h time difference": how do you get this number?

P3047,L12:"intradaily" => diurnal

P3047,L20f: provide also the overpass time for F13, and F14

P3047,L26: Why don't you use the available daily product? This would make the comparison closer to what was done for SSMIS?

P3048,L6: "values are larger..": therefore better use daily data and make comparison as done for SSMIS see comment above.

P3048,L25: "potentially important differences": which region are you exactly talking about?

P3049,15ff: also here the area of large differences is situated south of the Sahara, so the attribution to (problems with) high albedo values is questionable. Please provide details and results of the correlation analysis you did

P3049,L20ff: it would be more interesting to see these scatter plots for collocated data pairs (so far only possible for SSMIS comparison) and then also for the other comparisons (GOME-2 vs. SSMIS, GOME-2 vs. ERA-interim).

P3050,L27: "because of the reduced overlap between the GOME-2A- and GOME-2B orbits is lower is the tropics". Reformulate and explain why.

P3051,L1: "Assimilated data" => reanalysis data

P3051,L9: "surprisingly good agreement": Really? What did you expect? Did you do similar comparisons based on the previous retrieval version? Then it would be interesting to see the improvements.

P3051,L10: why four data sets? You mean three (ERA-interim, SSMIS, SSM/I-MERIS).

Editorial Remarks

1. Additional work concerning the language/phrasing is needed. The authors are strongly urged to carefully check their manuscript on spelling and grammatical errors once again. Some examples are given below, but list is by no means exhaustive :

1) P3024,L13: offers => offer

   P3024,L22: constrained on => constrained to

   P3028,L3: at the 15 of July => on 15 July; check other dates!

   P3035,L29: The net effect…is to reduce => is a reduced bias

   P3044,L22**:** "expected", rather estimated or similar..

   P3047,L8**:** North Ocean Atlantic => "North Atlantic Ocean"

   P3049,L5**:** higher data => higher data values

   …

2. There are many abbreviations which are not introduced at all, introduced several times or not introduced when used first but later, for example GOME-2A GOME-2B, HOAPS, WV, UVN, O3M-SAF, ERA-interim, TCWV.. This list not exhaustive. Please check carefully.

3. Figures and Tables

   a. Several figures are too small and therefore hard to read, Fig. 2, 3, 7, 8. Their size needs to be increased.

   b. Table 2: As the spatial coverage differs between the data sets statistics should be provided separately for ocean (maybe also coast) and land areas. It is not mentioned that SSM/I+MERIS statistics are based on a shorter time period only. How do you calculate the margins you provide for bias and RMSE?

   c. Fig7**:** How did you calculate the global means? Did you apply area-weighted average or just averaged over all grid boxes?