Authors' response to Interactive comments of Anonymous Referee #1 on "Comparison of GOME-2/Metop total column water vapour with ground-based and in situ measurements" by N. Kalakoski et al.

We thank the referee for careful reading of our manuscript and for the detailed comments. We will incorporate these comments to the revised manuscript. Below, we list the referees' comments followed by our answers (in blue).

The manuscript describes the validation of the improved GDP v4.7 operational retrieval of total column water vapour from GOME-2 using radiosonde observations and ground-based GPS retrievals. The validation concentrates on the calculation of the relative differences of the total column water vapour retrievals and the interpretation of the dependencies of these differences is done by means of the median and some percentiles (5th, 25th, 75th, and 95th) of the distribution.

The topic is of course well suited for publication in AMT, especially in the special issue. Moreover, in the same special issue, the GOME-2 total column water vapour product has been validated/compared with measurements of other instruments in 2 other manuscripts in AMTD:

· _Antón, M., Loyola, D., Román, R., and Vömel, H.: Validation of GOME-2/MetOp-A total water vapour column using reference radiosonde data from GRUAN network, Atmos. Meas. Tech. Discuss., 7, 9573-9601, doi:10.5194/amtd-7-9573-2014, 2014

· _Grossi, M., Valks, P., Loyola, D., Aberle, B., Slijkhuis, S., Wagner, T., Beirle, S., and Lang, R.: Total column water vapour measurements from GOME-2 MetOp-A and MetOp-B, Atmos. Meas. Tech. Discuss., 7, 3021-3073, doi:10.5194/amtd-7-3021-2014, 2014

Furthermore, as already mentioned in a short comment, a recently published paper in AMT also compares GOME-2, radiosonde and GPS retrievals of total column water vapour:

· _Van Malderen, R., Brenot, H., Pottiaux, E., Beirle, S., Hermans, C., De Mazière, M., Wagner, T., De Backer, H., and Bruyninx, C.: A multi-site intercomparison of integrated water vapour observations for climate change analysis, Atmos. Meas. Tech., 7, 2487-2512, doi:10.5194/amt-7-2487-2014, 2014

My major concern with the manuscript in its present form is that the findings obtained here have not been opposed thoroughly (enough) to similar findings in the mentioned references. I will give some examples:

· _Although, as I understand from the author's response to the referee comments to the Grossi et al. manuscript, the present manuscript and the Grossi et al. manuscript are coupled, e.g. the result of the wet bias of GOME-2B against GOME-2A is not explicitly confirmed here. The major difference of version 4.7 with the previous version is the introduction of an empirical correction of scan-angle dependency, but it is not explicitly stated from the validation in this work if this correction works well or not (see also below). Also, Grossi et al. stated that the quality of the GOME-2 water vapour data might depend on solar zenith angle, surface albedo and cloud fraction due to approximation in the retrieval algorithm (p. 12526, L16-18). This statement has been investigated by constructing Fig. 6. However, from the text, it is not clear if these dependencies are really present or not in the

case of the SZA and the cloud fraction.

· Both the present manuscript as the Antón et al. manuscript use radiosonde data to validate GOME-2 total column water data. This manuscript uses only GDP v4.7 retrievals, while 86% of the data used in Antón et al. is retrieved with GDP v4.6. This manuscript uses the radiosonde from the IGRA database (a rather inhomogenous database of different radiosonde types, processed differently), Antón et al. uses the homogeneous GRUAN database (all radiosondes types are RS92, all data processed consistently). These differences are vital within the framework of validation studies. Then, as a reader, you expect some comparisons of the findings of both studies, particularly with respect to the dependencies of the GOME-2 radiosonde differences on solar zenith angle, cloud fraction, and total column water vapour.

· As has already been mentioned in the short comment by Roeland Van Malderen: some of the findings in the manuscript should be confronted with theirs as well (some examples are given in the short comment).

We agree with the reviewer that these recent results need to be directly addressed in the manuscript. Manuscript will be revised accordingly. Please see answers to detailed comments for details.

A second major comment is the imprecise usage of terms like “comparison “and “validation” that are really linked to the main purpose of the paper. For instance, in the title and abstract, “comparison”, “compared” and “comparisons” (2*) are used, while the conclusions begin with “we have performed the global validation”. So, is this a validation or comparison paper? To my opinion, and knowing the background of the authors, this is a validation study, since the purpose is really to compare the GOME-2 retrievals against references of known accuracy, such as radiosonde and GPS ground-based observations. But, then, a description of the accuracies of radiosonde and GPS retrievals of total column water vapour is missing in Section 3: “Ground-based data sources”. In this context, please also note that radiosonde observations are in-situ observations and not ground-based.

I also think that the important section 5, with the title “Results and discussion” should be reorganized. First, you might explain and argue on the used statistical method (relative and not absolute differences, median and not mean, percentiles (5th, 25th, 75th, and 95th) of the distribution), then show the overall agreement and define subsections for each of the investigated dependencies of the differences.

We agree with the reviewer re. reorganization. Manuscript will be revised accordingly.

Finally, the English could also be improved. Very often, the article word (the, a) is missing or is used improperly. I advise you to ask a native speaker to read the manuscript. Different spellings of the same word are used throughout the manuscript (e.g. vapor and vapour, co-located and collocated). Please choose one and use it consistently throughout the manuscript afterwards. Also, identical formulations are repeated closely after each other (e.g. “The comparisons are performed…” in two sentences in a row in the abstract).

We agree, revised manuscript will be read by external proofreader.

Taking these considerations into account, I believe the paper can be accepted for publication after a major revision by the authors. I think that the paper can greatly be improved with the suggested supplementary comparison of the findings with other studies and proposed corrections and I am willing to review it anew afterwards. The core of the research is very

interesting and well established, but there is a need to dig a little deeper and add some perspective to the work by comparison with the literature.

Thank you for your comments. We agree that the suggested revisions will greatly improve the paper.

Detailed comments

· Title: comparison → validation

Agreed, will change the title.

· P12518, l6: GPS does not "observe" water vapour, but water vapour is retrieved by GPS. A GPS system registers time delays. Therefore, change to "co-located radiosonde observations and Global Positioning System (GPS) retrievals"

Changed accordingly.

· P12518, l6-8: both sentences start with "The comparisons are performed". Please use "validation" and another formulation for one of them.

Changed accordingly.

· P12518, l11: please remove "and screened for soundings with incomplete tropospheric column"

Changed accordingly.

· P12518, l14-20: radiosondes are in-situ measurements rather than ground-based. You should use the term GPS total column water vapour "retrievals", rather than "observations".

Changed accordingly.

· The abstract might mention something about the investigated dependencies of the differences on SZA, cloud fraction, surface albedo, geography.

Mention added to the abstract.

· _P12518, l23: in the Kiehl and Trenberth 1997 paper, it is written that water vapour accounts for about 60% of the greenhouse effect **for clear skies.** So, please add this.

Sentence clarified.

· _P12519, l26-27: the microwave measurements are typically limited to ocean areas. Infrared observations on the other hand have the …

Changed accordingly.

· _P12520, l20: Grossi et al. (2014) compared "it" to SSMIS measurements etc. Now "it" refers to the algorithm and so you are stating that an algorithm is compared with measurements.

Changed accordingly.

· _P12520, l26: have been extensively compared with SSM/I observations instead of "validated against"

Changed accordingly.

· _P12520, l26-27: "Data have been found to generally slightly …" What an awkward phrasing. Please reformulate.

Changed accordingly.

· P12522, l12: Replace title in "In situ and ground-based data sources"

Changed accordingly.

· _In Section 3, you should write some lines about the homogeneity of the datasets and give

accuracy estimates (see e.g. Van Malderen et al. (2014) and references therein).

Discussion of the homogeneity and accuracy added to manuscript.

· _Also, specify which auxiliary meteorological data have been used to convert the GPS Zenith Total Delay measurements to TCWV.

Suominet uses data from sites that have co-located surface meteorological sensors. Will update the paragraph accordingly.

· _P12523, l3-5: The discussion about the biases between radiosondes and GPS is so incomplete and is wrong. Igondova (2009) (is this peer-reviewed? I think they are better references than this one) mentioned a wet bias of 0.135 kg m^{-2} of GPS against radiosondes (the differences is actually $PWV_{GPS} - PWV_{radiosonde}$), and I did not find the sentence you quoted in the original Wang and Zhang (**2008**) publication. On the contrary, to my knowledge and also in this publication, most common radiosonde types have a dry bias against GPS PWV retrievals. The paragraph should be rewritten considerably.

You are correct, major mistakes here. Discussion removed, some sentences are added to paragraphs describing the IGRA and SuomiNet datasets.

· _P12523: In section 4, you are using co-location for both spatial co-location and time coincidence. The text will be clearer if you make the distinction between both and speak about coincidence for time.

Changed accordingly.

· _P12523, l24-26: Please rewrite your radiosonde –GOME 2 co-location criteria. "GOME-2 measurements that are co-located with the radio soundings within GOME-2 pixel" is such an awkward phrasing, and also not very clear. What do you mean? You calculated the trajectories of the radio soundings and subsequently looked for co-location with the GOME-2 pixels? Or you selected radiosonde measurements with sites located in the GOME-2 ground pixel? Your next sentence also does not help in clarifying this criterion.

Changed accordingly.

· _P12524, l1: by integrating "the specific humidity measurements" from "the" surface up to …

Changed accordingly.

· _P12524, l7: Please reformulate "GOME-2 and GPS measurements co-located within GOME-2 pixels". Again awkward phrasing.

Changed accordingly.

· _P12524, l9-11: It is not clear what the maximal allowed time difference between the GOME-2 and GPS measurements actually is. If you have GPS measurements every 30 minutes, it will be 15 minutes, as you stated. But it is not clear from the text what you do if you have gaps in the GPS time series. Then you allow a larger time difference? Please reformulate.

Changed accordingly.

· _P12524, l12: We have used only GPS measurements that have "a" formal error (as specified in "the" data files) not exceeding 0.3. → 0.3 kg m^{-2} or 0.3%? Please specify.

Specified. Unit is mm (kg m^{-2}).

· _P12524, l21-24: Argue why you are using this statistical methodology to compare GOME-2 measurements with radiosonde and GPS retrievals of TCWV.

I considered several methods for representing the distribution of the co-located data points, in order to take to account the large number of data points and the range of observed PWV. I considered the current visualization to be best suited for illustrating the behavior of the full data set, both for very small PWV amounts and very large.

· _P12525, l1-2: “This suggests that GOME-2 water vapour estimates are less reliable above 50 kg m^{-2}.” On which grounds do you make this statement? Are there indications that the GOME-2 retrieval does not work well at high TCWV values? High TCWV values can be associated with clouds, so can these have an impact on the retrieval? Have you so much confidence in the RS and GPS retrievals above 50 kg m^{-2}?

Removed the mention of reliability, added mention of high cloud top heights and albedos. Cited Anton et al., who investigated these parameters.

· _P12525, l2-5: “The range and number of outliers (i.e., large differences, which are seen in Fig. 2), is however smaller in comparison with GPS than in the comparison with sondes. This might be due to a smaller time difference between GOME-2 and GPS measurements, or due to a more robust water vapour estimates in GPS data.” → Perhaps you should look up in the literature how GPS TCWV retrievals behave with respect to radiosonde observations for large TCWV values!

Removed the sentences, unnecessary speculation on relative small difference**.**

· _P12525, l10-11: “This agrees also with the differences between radiosonde and GPS data reported in Wang and Zhang (2008).” As already mentioned, Wang and Zhang (2008) mentioned a dry bias of radiosondes w.r.t. GPS. And if you want to compare radiosonde data with GPS TCWV retrievals, do it directly and not via the comparisons with GOME-2.

Agreed, sentence omitted.

· _P12525, l14-15: “The shape of the scatter plots (Fig. 2) suggests that the overall biases depend on water vapour abundances.” → What about “The overall biases are independent of the water vapour abundances between 8-50 kg m-2 . At the edges of the TCWV range, a depencency might be observed, but might be related with specific instrumental shortcomings for measuring TCWV.” Could this statement be valid instead? Why (not)? Please argue.

You are correct, suggested formulation is more informative in this case. Updated accordingly.

· _P12525, l19-20: “At low H_2O values, below 8 kg m^{-2}, a large positive bias is clearly visible, especially in comparisons with the GPS data.” Why is this to your opinion?

Low water vapour abundancies occur mostly at high solar zenith angles. Anton et al. show positive biases at the case of low cloud fractions and cloud top albedos for high SZAs. We will investigate this, discuss findings in the revised manuscript.

· _P12525, l26-27: “No significant difference can be seen in behavior between GOME-2A and B.” How this relate to the findings in Grossi et al. (2014)?

Grossi et al. reported a small wet bias for GOME-2B. Updated accordingly.

· _P12526-12528: Especially the section 5.2 could be restructured and organized better. A division in subsections could improve the readability of the different investigated dependencies. Now all results are presented directly one after the other.

Updated accordingly.

· _P12526, l2-9: Here, you should explain more about the scan angle dependency and its consequences for the TCWV retrieval. In this part, it is not mentioned anywhere that this results in a bias between the H_2O product for the west and east part of the swath and the central ground pixels (Grossi et al., 2014). This information is necessary to interpret Fig. 5.

Updated accordingly.

· _P12526, l12-15: “As observed in Fig. 5, the scan-angle dependence of GOME-2 H2O data is small. However, the western edge of the GOME-2 swath shows about 5% higher water vapour column than the eastern one in comparisons with the radiosonde observations. In comparisons with the GPS observations, both edges of the swath show wet bias about 10%

compared to the center of the swath." → So, what can you conclude about the empirical correction applied in GDP v4.7? Based on these findings, does it work well? Why (not)?

We will expand the discussion on this topic in the revised manuscript. Some dependency on the scanning angle is still visible, although it is not clear how much of this is due to other issues such as diurnal variation.

· _P12526, l16-27: How do these findings on the dependency of the relative differences on SZA and cloud fraction relate with the results obtained in Grossi et al.(2014) and especially in Antón et al (2014)?

Anton et al. reported similar behavior for both dependencies, updated accordingly.

· _P12527, l7-8: You should include in Table 2 the number of co-locations for the different surface types. There might be a sampling size issue. Discuss.

Number for co-locations added to table. While much smaller than in the case of land co-locations (318000, 103000 and 34000 for land, sea and ice, respectively), number of co-locations for sea and ice co-locations should be large enough for statistical analysis.

· _P12527, l15-29, P12528, l1-16: This paragraph can be shortened considerably in case of only considering the seasonal and latitudinal dependence of the bias with respect to radiosondes. At the end of the paragraph, you give the argument yourself: the GPS network is too sparse, especially in the Southern Hemisphere, so that the zonal monthly medians are based on data of just a few stations (even with questionable GPS or meteorological data quality). So, concentrate in this paragraph only on the GOME-2 radiosonde biases! Fig 9. can also be omitted, just refer to Fig. 1 to point to the GPS network being sparse, especially in the Southern Hemisphere.

While I agree that the paragraph would be clarified by omitting the GPS comparisons, I would prefer to keep the GPS discussion in the manuscript for the sake of completeness. We added reference to figure 1 to the caption of figure 9.

· _P12527, l18-19: "When compared with sondes, GOME-2A generally has a wet bias in the Southern Hemisphere and a dry bias in the Northern Hemisphere."→ What is the reason for this?

Bias in southern hemisphere is most likely related to bias observed over oceans high latitudes by Grossi et al.. Likewise, they observed dry bias over land in northern hemisphere. Distribution of the stations might also affect this issue. We will add discussion of this issue to revised manuscript.

· _P12527, l21-22: "Seasonal variations in the differences can be seen at mid-latitudes, especially in the Southern Hemisphere. These seasonal variations at mid-latitudes are in a broad agreement with the general dependence of GOME-2 biases shown in Fig. 3: a negative/smaller bias in wet seasons (summer) and a positive/larger bias in dry seasons (winter)." → OK, but would you then not expect a negative/smaller bias at low latitudes (high TCVW) and a positive/larger bias at high latitudes (low TCVW)?

This issue is also likely complicated by the distribution of the stations. Coastal stations are more likely to show wet bias and the inland stations dry bias. We will investigate this further for revised manuscript.

Suggested technical corrections corrected, revised manuscript will be read by external proofreader.