Answers to Referee n°1

We thank the referee for the taking the time to review this paper and providing constructive comments and suggestions. The referee's comments are repeated below in black italics and our answers are given in blue.

General Comments

The work presented focus on comparing atmospheric integrated water vapour (IWV) estimated from ground-based GPS observations with the corresponding IWV values in four nearby grid points of the ERA Interim reanalysis. The structure of the manuscript is straightforward and reasonably easy to follow, although I needed quite some time until I was familiar with the nomenclature and the symbols.

Thank you for the comment. You are right that we used a number of symbols to quantify our results in the text (rather than paraphrasing) but they were all defined in the Appendix. This may actually require some back and forth reading but the advantage of an Appendix is that all the definitions are grouped in one place.

The stated motivation for the work was to identify GPS stations where ERA Interim is not recommended to be used when searching for inhomogeneities in the GPS time series of IWV.

Actually the goal of the study was stated as "to better understand to which extent model errors, GPS errors, and representativeness errors can be distinguished, what is the limit set by representativeness differences on the best achievable agreement between global reanalyses and station observations, and explain their contribution to the geographical and seasonal dependencies reported in previous publications." (page 2 line 32 to page 3 line 1).

Later we mention the application of these results to homogenization: "The results from this study are important to homogenization work where IWV data from reanalyses and GPS observations are used jointly…" (page 3 line 16-17).

A question that is not answered after reading the manuscript is an approximate quantitative relation between representativeness statistics and the size of the break in the GPS IWV time series. I think like this: if representativeness errors (at a specific GPS site) are stable with time, it should still be possible to detect a break in the GPS time series if it is above a certain size? It would be interesting to have the authors ideas about how large, or small, breaks that could be detected, given some example values of the representativeness statistics.

This is an interesting question but it would require to establish a rule between our representativeness statistic and the results from a homogenization method. So this is clearly beyond the scope of this paper. Our recommendation here is simply to discard the stations that were detected as outliers based on the proposed thresholds for this dataset (page 11 line 31-33).

Actually, it is difficult to give a more general answer. The size of breaks that can be detected depends strongly on the statistical test or homogenization method used. One of the problems is namely that representativeness errors show in general a strong seasonal variability (see Fig. 8 and 9). In this respect, we think it is primordial that the homogenization method takes the non-stationarity of the variance into account.

Figures 2–6 are presented and discussed in Section 3. Some of them have red dotted lines defining limits in order to identify outlying results/stations. However, it is only in Section 4 that these limits are explained. I think it would help the reader if they were introduced already in Section 3. Related to this

it is clearly stated that the method is subjective. Nevertheless, if the method is to be applied by others, it would be informative to also document the reasoning behind the choices. For example, why did you choose non-symmetric limits for the mean differences in Figures 2 and 3?

Regarding the outlying stations, we actually refer the reader to Section 4 when we present the first figure in Section 3 (page 5 line 18). To make it clear that the red dotted lines are related to the outlying sites we clarified the sentence in brackets:

"(the outlying stations, defined beyond the red dotted lines, will be discussed in Section 4)."

The reason why we added the red dotted lines and named the outlying results on the figures presented in Section 3 is that this avoids us to duplicate the figures later in Section 4.

The choice for the limits is subjective because we think that the results are very unpredictable when analysing a global network (due to the variety of climates, equipment, and reanalysis performance). So it is necessary to inspect visually the results and determine the limits beyond which the results do not look "normal". We believe this approach is quite robust thanks to the combination of several representations of the results such as shown in Figure 2-6 (i.e. function of latitude, altitude, mean vs. std. scatter plots, etc.). This methodology can be safely applied to other datasets.

The reason why we chose non-symmetric limits wrt to zero for the mean differences is because the distribution is not centred on zero. This is quite clear in Fig. 2, 3, and 5. Choosing symmetric limits here would remove more stations on one side, which is not wanted.

Specific comments

P1,L19: It is not surprising that the comparison results are significantly improved when the worst 15 sites (of 120 sites) are removed. It would be informative to quantify the improvement.

It is quantified in the next sentence (20 to 30%).

P2,L25-26: Is that not obvious? I mean it is stronger than "a tendency".

Well, many studies reported the same ("Absolute differences have a tendency to be larger in moister and warmer regions/periods while relative differences tend to be larger in colder and drier regions/period, globally.") and some hypotheses have been made about the reasons but no clear explanation has been found. So, though the conclusions are not new, it doesn't make them obvious as long as they are not fully explained.

Actually, the dependence of absolute and relative errors on IWV and other factors depends on the observation/processing method and the nature of the noise/error sources. E.g. lidar water vapour measurements errors follow different statistics.

P2,L32-34: Are representativeness errors never to be referred to model errors? I interpret the definition of a representativeness error as that the only cause is the limited model resolution? If this is correct it can be stated explicitly, because I can also argue that the limited resolution of a model can be the cause of "model errors".

We distinguish model errors (i.e. error due to the model physics) and representativeness errors (more directly linked to the model resolution and the fact that it cannot represent small scales that are sensed by the observing system). We think the difference is clearly stated page 2 line 15-17. But you are right, the limited model resolution can also be a cause for model errors (e.g. when convection is parameterized vs. explicit). This is one of the reasons why better results are found with the AROME model (page 6 line 25 – page 7 line 5).

It is actually difficult to discuss further the model errors in this study because we have no diagnostic to evaluate them contrary to the representativeness error (according to our definition) for which we proposed a statistic. This statistic can be computed easily from the IWV at the four surrounding grid points and can be used to detect outlying sites, e.g. for the purpose of GPS IWV homogenization.

P3,L15: Explain/give examples, what is meant by "atmospheric environment" already here? Although it is clarified later when presenting Figure 9, my reading stopped here wondering what atmospheric environment could be out of many different things?

We added "(mean IWV and variability)"

P4,L9-11: I think you should mention that the GPS time series used have passed some kind of quality check, because a very large break should have an impact on the overall standard deviation of the differences GPS – ERA Interim

We added "and the data have been screening beforehand".

P6,L5: It cannot be taken for granted that the discrepancy is not due to GPS errors just because the formal errors do not increase. For example, a nearby installation of say a metallic structure may introduce significant multipath errors without affecting the formal errors.

In general, when the noise in the measurements is increasing the formal errors are increasing too because they are rescaled based on the "a posteriori variance factor". So we would expect that a sudden increase in multipath error would be detected by the screening procedure, though such a case has not yet been clearly identified. On the other hand, since we compute statistics over 16 years it is likely that an undetected temporary increase in multipath errors would not impact strongly our statistics.

P6,L30-33: An additional explanation could be that you only required 15 days of data for a specific month in order to be included. That would also affect the reduction of the standard deviation, unless it is very rare that so much data are missing from a month?

In this study "Monthly averages are computed directly from the 6-hourly values within the given month to the condition that at least 60 values are available (similar to Parracho et al., 2018)." (page 4 line 7-8). We don't think this can impact the statistics of differences because the GPS and ERAI data are time-matched before the monthly values are computed for each dataset. So there is no sampling difference.

P7,L2-5: Can you compare this standard deviation of 0.81 kg/m2 to what is obtained for stations located in the same area of the present study, in order to quantify the improvement obtained for the higher resolution model?

There are 12 IGS stations in the AROME-WMED domain. The median standard deviation of IWV differences GPS-ERAI over these 12 stations amounts to 0.98 kg m^{-2}. This number is indeed larger than the 0.81 kg m^{-2} found with AROME-WMED (over 661 stations), but the difference is not only due to resolution but also to different and more modern model physics in AROME-WMED (in addition to different spatial sampling 661 vs. 12 stations).

P7,L13-15: Perhaps the GPS sites that do not show an improvement using bi-linear interpolation are located close to one of the four grid points that is more representative compared to the others?

This hypothesis can be tested from Figure 7 where the grey bars shows the results for each of the four grid points ordered by increasing horizontal distance from the GPS station. At 8 out of 15 sites, the closest grid point gives the smallest std. dev. of difference, so about 50% of the cases.

P11,L11: delete "strong" and just give the value? It should be up to the reader to decide what is a strong and a weak correlation

Given the different nature of the two variables compared we think that the 0.73 value can be considered as strong (this is a different situation from the comparison of a similar variable from two difference data sources where we would require at least a value of 0.90 to be strong). And in general, we think authors should give their interpretation of results and not leave the task to the reader.

P12,L23: delete "good", or state your definition for "good". Which parameter values do you typically see for Antarctica that is not seen globally?

In Antarctica, the comparison statistics exceed the global thresholds at 4 out of 5 sites (see Fig. 2 for the values of the thresholds). We completed the sentence to clarify this point:

“where the comparison failed at 4 sites out of 5.”

P12,L25-26: This last sentence is not clear. It is the word “also” that raise questions. Because isn’t that what you have done in the study? And what is meant by “other observation types”

You are right the sentence was not clear. We reformulated it:

“The methodology described in this paper can also be applied to assess the consistency and representativeness of other data sources (e.g. climate models, satellite IWV data) and other observation types (e.g. surface humidity, temperature, etc.).”

Fig. 3: The figure caption refers to Figure 2. Are really the black dashed lines in Figure 3 of order 5 to 9?

You are right, they are linear fits. We added the information in the captions.

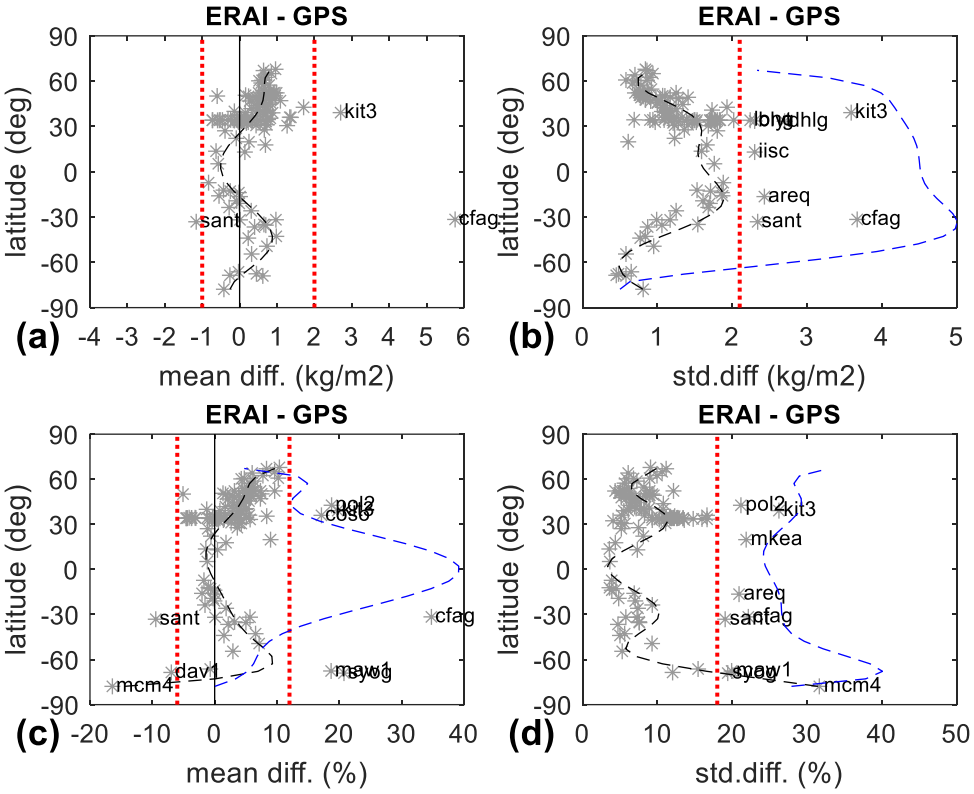
Technical Corrections

Thank you for the careful reading. All the suggested corrections have been implemented in the text.

Figures: the red dotted lines are kept thin because we don’t want to highlight them when we discuss the overall results in Section 3. Later in Section 4, we focus on the named station results, so again the thresholds do not need to be emphasized too much.

An example with linewidth=1 is given below and we think it emphasizes too much the red dotted lines.

We made the blue lines dash-dotted as suggested in Fig. 2 and moved slightly the station labels in Fig. 5 (for dav1 and a few other stations).


ERAI - GPS
latitude (deg)
mean diff. (kg/m2)
(a)
kit3
sant
cfag
ERAI - GPS
std.diff (kg/m2)
(b)
kit3
areq
sant
cfag
ERAI - GPS
mean diff. (%)
(c)
sant
cfag
mcm4
dav1
ERAI - GPS
std.diff. (%)
(d)
pol2
kit3
mkea
areq
sant
cfag
mcm4