The referee's comments are presented followed by our responses in blue.

Anonymous Referee #1

General Comments

The manuscript presents results for IWV mean values and linear trends over time periods of 15-20 years in a global perspective. As far as I know this paper is unique and it offers additional knowledge to the paper by Wang et al. (2016) which in some aspects is similar, such as the use of the GPS dataset from the IGS. At places I think the manuscript reminds of a text book. For example, I think there is no need to explain the general distribution of the IWV, as in the beginning of Section 3, with mean values of the IWV for different areas. In general, I find the text to use too many words. I will mention a couple of additional examples below. If the text is more focused on what is new it will be easier for the reader to grasp the important results, i.e. these that are new or are contradicting previously published results. It is a bit disturbing that the time period with GPS data ends in 2010, covering a 15 year period, when we are now in 2018. I wonder if the additional 5-7 years of data would influence the overall results. Trends are small and a longer time period would be beneficial. Perhaps also the GPS stations have been more stable in terms of less changes of equipment during the recent years? Perhaps there is a another manuscript in preparation?

I note that the last section is called "Summary" rather than "Conclusions". Perhaps this is intentional? I miss conclusions in terms of what have you found that affects other studies? Do you have recommendations concerning the use of the different data sets? What are the future problems to address, et cetera? Given that the summary section to large extent repeats results already presented, it can be significantly shorter.

>> We thank the referee for the constructive comments that helped improving the manuscript. The referee is right in pointing that this work is in some aspects similar to Wang et al., 2016, but also that it is unique and offers a great deal of additional knowledge, such as separating the analysis in seasons (DJF and JJA), which is important to help understanding the underlying climate processes, and also evaluating modern reanalysis, which are widely used by the climate community.

We took the general comments into account and tried to make the manuscript shorter overall. Specifically, the general description in Section 3 was almost completely removed and the Introduction and Conclusion sections were largely rewritten (according the specific comments given below). The introduction focuses more directly on the topics addressed later in the paper, and the conclusions state more clearly what is new and what are future problems to address.

Regarding the comment about the period used in this study, the reason why it stops in 2010 is that it uses the IGS repro1 dataset (the same Wang et al. 2016 used, with slight differences in the post-processing as explained in Section 2). This is the only official reprocessed IGS tropospheric dataset available to date. Tropospheric data from a 2nd more recent reprocessing which goes until end of 2013/2014 are only available from individual analysis centers and are not qualified so far. They might be evaluated in a future study and namely require to check the homogeneity of the more recent years. However, in order to solve the new homogeneity issues highlighted in our study, a complete reprocessing of all stations would be preferable.

If the extra 5-7 years were included the results might change slightly for two reasons. First, additional inhomogeneities are expected both in the GPS data (e.g. due to recent equipment changes) and in the reanalyses (due to recent observing system changes). Second, we know that the trends are not exactly linear, hence the period of study has an impact on the trends computed. That is why we are careful to state the trends are for 1995-2010.

The GPS stations have undergone systematic equipment changes in recent years to adapt to the new GNSS signals from Galileo, Beidou, etc. These changes are not expected to have as dramatic impact as

observed with some older equipment as more accurate processing models are available but the exact impact on IWV trends from these data has not yet been evaluated.

To be noticed also is the new figure 3 presenting differences in the means and interannual variability between ERA-Interim and MERRA-2. In the first submission only trend results were shown for MERRA-2. We added this figure for the sake of completeness and higher consistency in the discussion of both reanalyses. It allows us to draw more general conclusions on the performance of modern reanalyses.

Specific comments

pages 2-4: The introduction could be much more focused. Many of the issues here are not addressed by the results presented later.

>> The introduction has been in large part re-written.

page 5, line 20: I do not see a problem of using all the 20 stations in the western USA, because you never present any global averages or trends. Please explain why overrepresentation would be an issue.

>> We do present global statistics (i.e. including all stations) for the means and standard deviations of IWV differences in Table 1.

In addition, it would be impossible to distinguish between points when all 20 stations are included in the maps, so it is not useful to retain all stations.

page 6, lines 8-11: I would like to see stricter requirements on the data availability in order to have representative values for a month and for a season. It is stated that GPS offers continuous data coverage back to 1995 so from that point of view it does not make sense to be so "generous" when accepting data.

>> The two reasons for adopting these criteria were the following:

1) GPS is not really continuous (it contains gaps in the time series), so we need to apply a selection.

2) However we do not want to eliminate too many months, because when we compare with ERA-Interim, in the maps, they are not time-matched, so the elimination of too many months would impede their comparison.

To avoid overstatement, we removed the word “continuous” from the description of the GPS data.

page 6, subsection 2.3: This method used for the calculation of the trends has become popular recently. Here I miss one or more quantitative example(s). All IWV results are presented as absolute (kg/mˆ2/decade) or relative (%/decade) trends for the IWV. Temperature trends are presented e.g. in Subsection 5.3. What are the corresponding resulting uncertainties obtained with this method compared to the classical least squares fit? Furthermore, it would indeed be interesting (for at least a few sites) to present the differences between trends obtained by these two methods.

>> Before deciding to use the Theil-Sen method to compute the IWV trends, we compared it to the Least Square method. Both methods were applied to the monthly mean ERA-Interim and GPS IWV data, and the differences between the trends were plotted and are presented in an Appendix that was added to the paper.

page 7, line 23: The words "good agreement" needs to be defined. Actually I think there is no information in this sentence. The quality of the agreement is discussed in detail thereafter and "good" means different things to different readers.

>> The entire sentence here was removed.

page 10, lines 3 and 6: again unclear, what is good agreement?

>> Same sign, the text has been corrected to make it clearer: *"In general, the monthly trends computed at the GPS stations are consistent in sign and magnitude with ERA-Interim (Fig. 6)."*

page 10, lines 14-16: This statement is not needed. It is just statistics (that are expected) and not used later.

>> The statement was removed.

page 11, lines 3-4: It is a bit unclear if all coastal and mountainous sites have a problem with representativeness? Otherwise is not this too much hand waving?

>> Not all, the sentence was corrected: *"Representativeness differences are suspected at some mountainous and coastal sites (e.g. AREQ, CFAG, KIT3, MAW1, SANT, SYOG and the other sites discussed in the previous section)."*

page 11, line 9: again the term "good agreement" is used. Try instead to describe quantitatively how they agree (and perhaps not agree? For example if a difference between two trends is 0.2 kg/mˆ2/decade, is that a good agreement or is it a disagreement?

>> The sentence was changed: *"Over the oceans, results that are significant in ERA-Interim are consistent (i.e. same sign) with those obtained by Wang et al. (2016), despite the fact that they are not always significant over land in the latter study."*

page 11 line 11: "... none of the values computed by Wang et al (2016) are significant but the drying over western Australia is also observed." How is it observed when the values are not significant?

>> This sentence was removed.

page 11, line 17: Here is an example where you refer to "low values". Does that mean that they are comparable to the uncertainties (in kg/mˆ2/decade, this connects to the comment on page 6)?

>> This part of the sentence was removed. In fact, only the absolute trends are low (because IWV has low values in this region), the relative trends are not small.

page 12, line 20: Here you define that "good agreement" means that some features are confirmed. However, you can just state that the features are confirmed and ignore to add the subjective wording "good agreement"

>> Sentence has been re-written to reflect this: *"Overall, the seasonal trends estimated from the GPS data confirm the features discussed above for ERA-Interim."*

page 19, line 3: define "high IWV gradients"?

>> High IWV gradients are regions where IWV varies strongly (e.g. around the ITCZ, mountain regions). This definition was added to the sentence.

Technical Corrections

>> All the corrections suggested by the referee have been implemented.

page 1, line 18: "Monthly IWV trends" is unclear, trends over a month or trends for all the January months et cetera, or (total) trends based on all monthly means?

>> The abstract has been rewritten, but later in Section 4, it is specified that trends computed from the time series of monthly means are referred to as monthly trends.

page 1, line 25: "found to not" –> "found not to"?

page 1, line 30; and page 2, lines 3 and 4 (as well as many additional places in the manuscript): Leave a space between a value and its unit, according to SI also for the units percent (%), degree north and east (◦N and◦E), and degree centigrade (◦C).

page 2, line 6: IPCC report is not in the reference list, specify which year and give an URL address? => no longer referenced

page 2, line 11: Tropics –> tropics (not a name)

page 3, line 5: annual and –> annual, and

page 3, line 9: GPS data has –> GPS data have

page 3, line 11: data is –> data are

page 3, line 14: This data is –> These data are

page 3, line 15: ERA-Interim , –> ERA-Interim,

page 4, line 1: e.g. dynamical –> e.g. the dynamical

page 4, line 8: In section 5 – > In Section 5

page 5, line 23: rate –> temporal resolution (rate is not measured in minutes)

page 6, line 7: such a statement requires a reference, otherwise it should not be stated.

page 6, line 23: no need to repeat the requirement from above

page 8, line 6: section 2 –> Section 2

page 10, lines 5-6: well documented –> well sampled ?

page 11, line 1: section 5 –> Section 5

page 12, lines 18-19: "More details will be given in the discussion section." There is no section with the title "Discussion"?

>> Summary section has been replaced with a “Summary and conclusions” section and the text has been corrected.

page 12, line 28: In this section MERRA-2 is –> MERRA-2 is now?

page 13, line 31: (Fig. 9c and d) –> (Figs. 9c and d)

page 14, line 1: (Fig. 5c and d) –> (Figs. 5c and d)

page 14, line 2: (Fig. 9d and d) –> (Figs. 9c and d) ?

page 15, line 23: (Fig. 11d, h) –> (Figs. 11d and h)

page 15, line 23: (Fig. 11b, f) –> (Figs. 11b and f) and again two times on line 27, same page, and many times on pages 17 and 18 ...

page 18, line 22: the dry season flow –> the flow in the dry season

page 19, line 2: In this paper we –> We

page 19, lines 10-11: again, define "monthly trends"?

page 19, line 20: seasonal trends –> long term trends for winters and summers? (this is the same language issue as "monthly trends")

page 25: font size of station names is too small to be readable

pages 38 and 41: font size of the text within the figure frames is too small to be readable

page 41, fig. 16 –> Fig. 16

page 41, caption: siries –> series

Anonymous Referee #2

Received and published: 20 April 2018

The manuscript includes a lot of information. It (1) compares PW monthly and seasonal means, interannual variability, and linear trends between reanalyses and GPS data for 1995-2010, (2) studies PW trends for 1980-2016 using two RA products, (3) looks at the relationship between PW and surface temperature trends, and then (4) tries to link the dynamics with PW trends and variability. The authors have done a lot of work, but it is hard to figure out what the focus of this study is and what original results are achieved. In the major comments below, I raised several major concerns. Based on that, I think that the manuscript needs major revision or is resubmitted later.

>> We thank the reviewer for the comments which we tried to implement in the revised manuscript. The manuscript has been revised throughout to highlight what is new and original in our results and the conclusion section has been rewritten. The introduction has also been re-written and refocused on the main questions addressed in the manuscript, including additional references.

Major comments:

1. Scientific originality: The scientific originality first starts from the review of prior studies and the motivation of this study. As I mentioned below, some important references are missing in the introduction. Then the authors have to provide rationale on why they want to study those four things (listed above). Have they never done before? Are your data better than that previous studies used? Are you going into more depth on those topics? I didn't see the strong motivation explained in the introduction. All those topics have been studies extensively before. What new and significant results does this study provide? The authors touched so many things, but didn't emphasize their originality. The authors try to describe all things they have done in a tedious way, so the manuscript looks more like a work summary, rather than synthesized scientific paper. I think that previous studies have done a lot for #1, #2, #3. Maybe the focus should be on briefly summarizing your results to establish the bases on using reanalysis data, and then on linking the dynamics with PW variability. For the first three, your results should be compared with previous studies, and then emphasize new results you found.

>> The introduction has been re-written and additional references have been included. We have emphasized our motivation and explained why we use these IWV datasets, why we focus on means, variability, and trends, and how our approach and data differ or complement previous studies. We have also tried to shorten our descriptions of the results (e.g. the description of the mean IWV has been squeezed, as also required by the first reviewer), and compared our results to previous studies when possible. Finally, we have rewritten the last section to include not only a summary of results but also a discussion of the conclusions, in terms of what is new and what future work should be done on this topic.

To answer more specifically the questions posed by the reviewer, only few studies investigated the means, variability, and trends in IWV and their link with surface air temperature. Major deficiencies in older reanalyses and observations-based datasets have been evidenced in past studies. Compared to those, we used modern reanalyses (ERA-Interim and, especially the very recent MERRA-2) and a new independent reprocessed 16-yr long ground-based GPS dataset to validate them. To our knowledge it is the first time reprocessed GPS IWV data are used to validate global reanalyses over such extended periods. We showed that the modern reanalyses still suffer from uncertainties in data sparse regions such

as Africa and Antarctica, and that IWV trends and variability at regional scale are dominated by changes in atmospheric circulation rather than surface temperature variations predicted by Clausius-Clapeyron.

2. Technical quality: Again too much information is provided including too many topics, tedious descriptions of all results and too many plots. After you decide the focus, the manuscript should be reorganized and be shortened.

>> The rationale has been sharpened and text shortened. More specific comments would help.

3. References:

The manuscript didn't include some of relevant references. I mentioned this in several places in Specific comments. You can find a lot of references from the review paper by Guerrova et al (2016, Review of the state of the art and future prospects of the ground-based GNSS meteorology in Europe, AMT).

>> New references have been added to the paper:

Dee, D. P. and Uppala, S.: Variational bias correction of satellite radiance data in the ERA-Interim reanalysis. Q.J.R. Meteorol. Soc., 135: 1830-1841. doi:10.1002/qj.493, 2009.

Mears, C. A., Wang, J., Smith, D., and Wentz, F. J.: Intercomparison of total precipitable water measurements made by satellite-borne microwave radiometers and ground-based GPS instruments. J. Geophys. Res. Atmos., 120, 2492–2504. doi: 10.1002/2014JD022694, 2015.

Meynadier, R., Bock, O., Gervois, S., Guichard, F., Redelsperger, J.-L., Agustí-Panareda, A., and Beljaars, A.: West African Monsoon water cycle: 2. Assessment of numerical weather prediction water budgets, J. Geophys. Res., 115, D19107, doi:10.1029/2010JD013919, 2010.

O'Gorman, P. A. and Muller, C. J.: How closely do changes in surface and column water vapor follow Clausius–Clapeyron scaling in climate-change simulations? Environ. Res. Lett. 5, 025207, 2010.

Robertson, F.R., Bosilovich, M.G., Roberts, J.B., Reichle, R.H., Adler, R., Ricciardulli, L., Berg, W., and Huffman, G.J.: Consistency of Estimated Global Water Cycle Variations over the Satellite Era. J. Climate, 27, 6135–6154, https://doi.org/10.1175/JCLI-D-13-00384.1, 2014.

Schneider, T., O'Gorman, P. A., and Levine, X. J.: Water vapor and the dynamics of climate changes, Rev. Geophys., 48, RG3001, doi:10.1029/2009RG000302, 2010.

Sherwood, S. C., Roca, R., Weckwerth, T. M., and Andronova, N. G.: Tropospheric water vapor, convection, and climate, Rev. Geophys., 48, RG2001, doi:10.1029/2009RG000301, 2010.

Thorne, P.W. and Vose, R.S.: Reanalyses Suitable for Characterizing Long-Term Trends. Bull. Amer. Meteor. Soc., 91, 353–362, https://doi.org/10.1175/2009BAMS2858.1, 2010.

Trenberth, K.E., J.T. Fasullo, and J. Mackaro, 2011: Atmospheric Moisture Transports from Ocean to Land and Global Energy Flows in Reanalyses. J. Climate, 24, 4907–4924, https://doi.org/10.1175/2011JCLI4171.1

Wang, J., and Zhang, L.: Systematic errors in global radiosonde precipitable water data from comparisons with ground-based GPS measurements. J. Climate, 21(10), 2218-2238, https://doi.org/10.1175/2007JCLI1944.1, 2008.

Zhao, T., Dai, A., and Wang, J.: Trends in tropospheric humidity from 1970 to 2008 over China from a homogenized radiosonde dataset J. Climate, 25(13), 4549-4567, https://doi.org/10.1175/JCLI-D-11-00557.1, 2012.

Specific comments:

1. GPS vs. GNSS: I would suggest that GNSS is used instead of GPS.

>> In this study we use a GPS-only solution, so it is correct to refer to it as GPS.

2. P1, L13: Do gaps affect monthly mean if they last longer?

>> If the reviewer means if gaps in the 6-hourly GPS IWV time series affect the monthly means: yes, and the impact is larger if the gaps last longer, that is why we reject months which have less than 60 values (half of the expected values). This ensures that we obtain representative monthly means, while not eliminating too many data (necessary for the comparison with ERA-Interim monthly mean fields like in Fig. 2, which do not have gaps).

3. P1, L24: What temperature? Surface or upper air?

>> The surface air temperature (2 m) was used. This information has been added to the abstract.

4. Abstract: I didn't learn a lot anything new from this.

>> Thanks for the notice. We hope that the revised abstract will be more instructive.

5. P2, L5-8: Lots of papers have discussed this. Please list some references.

>> Several references have been added.

6. P3, L10, Fig. 1: This map only shows 104 stations, much less than thousands of available stations. It is not convincing about "growing". I understand that those are the stations used in this study. But it is not convincing to use this to make your point here.

>> The term "growing" has been removed from the text. Note that even if thousands of stations are operated routinely, not all are publicly available and/or of scientific quality. Here we consider the IGS network which is a freely available global scientific network counting nominally about 400 stations.

7. P3, L28: There have been quite a few studies using GNSS PW to evaluate reanalysis products. First of all, you should summarize the prior studies on this, and then describe what is different (unique) about this study.

>> We don't think a summary of all prior studies is useful here because there is quite poor agreement between them due to the study of different periods and regions, and use of different datasets with various issues (e.g. older reanalyses had time-varying biases due to observing system changes and incorrect assimilation of recent satellite data, and most observational datasets did not use reprocessed and homogenized data). Nevertheless, we cite a few studies dealing with the same scope as ours and discussing these issues. This leads us to the main objectives and novelties of this study which are the assessment of modern reanalyses and the use reprocessed GPS data. These points have been made more clear in the revised introduction.

8. P5, L1: Should you just simply average the surrounding grid points or have more complicated regression? It depends on the location and topography. See Fig. 1 in Mears, C., J. Wang, D. Smith, and F. J. Wentz, 2015: Intercomparison of total precipitable water measurements made by satelliteborne microwave radiometers and ground-based GPS instruments. J. Geo-phys. Res. Atmos., 120, 2492–2504, doi:10.1002/2014JD022694.

>> The method we used is not simply an average of the surrounding grid points. As stated:

*"A bilinear spatial interpolation is computed from the model IWV estimates at the 4 grid points surrounding each GPS station. The IWV model estimates are then recomputed from the pressure level data by vertically integrating the specific humidity between the height of the GPS station and the top of the atmosphere. Most GPS station heights fall between two pressure levels and the specific humidity data can be interpolated. However, for stations located below 1000 hPa (the lowest pressure level) the reanalysis data are extrapolated. Interpolation and extrapolation are done linearly for specific humidity and temperature, and exponentially for pressure."*

9. P7, L10, the standard deviation is calculated from seasonal mean values. Is it right? In other words, only 16 data points are used to calculate standard deviation? Is the standard deviation statistically significant given only 16 data points?

>> Yes. The two-sample F-test was used to determine whether the difference in variances between GPS and ERA-Interim were statistically significant.

10. Fig. 5: The trends are calculated using monthly PW anomaly, not monthly mean, correct? Fig. 5 is a similar plot as Fig. 4 in Wang et al. (2016). You need to discuss how your results compare with Wang et al. (2016) here.

>> Yes, the trends are calculated using monthly PW anomaly, as described in Section 2.3: *"The Theil-Sen method was applied to the anomalies obtained by removing the monthly climatology from the monthly data."*

The results of Fig. 5 were compared with Wang et al. (2016) a little further ahead in the text, so as not to break the discussion of the agreement between GPS and ERA-Interim. We decided to keep the organization this way in the revised paper.

11. P10, L33-34: For WUHN, the big change in Oct 2006 is due to the radiosonde type change from Shang-M to Shang-E. Radiosonde data over land are the main source of upper air humidity for reanalyses, so any inhomogeneity in radiosonde data would be reflected in the reanalysis data.

>> Thank you for this information. It has been added to the manuscript: *"In fact, the break at the end of 2006 is associated with the change in radiosonde from the Shang-M to Shang-E, which is assimilated by ERA-Interim (Wang and Zhang, 2008). Zhao et al (2012) found that prior to this change there was a 2 kg.m-2 wet bias in the radiosonde data at the Wuhan station, in comparison with GPS. This moist bias is also observed in ERA-Interim prior to the end of 2006."*