Replies to the Comments:

The authors thank the reviewers for their insightful comments. In the following, the comments are included in black while our replies are given in blue.

General comments:

This work describes the annual, semi-annual, and quasi-biennial (QBO) variability of water vapour, as calculated from a large number of satellite-retrieved data sets from 1984 to 2014, with some differences arising naturally from the different time periods considered. Other differences in the behaviours are probably the result of a combination of effects, including sampling issues, systematic but non-constant error sources, vertical resolution differences, clouds, aerosols, and non-local thermodynamic effects on some retrievals, among other possible issues. Overall, the analyses are performed in a way that is sound and the paper is well written overall, but a few language adjustments and corrections are needed.

Some of my concerns and criticism relate to the legibility of some of the Figures, or some of the points shown in these Figures, as a result of the large number of data sets that were used in this comprehensive analysis and summary. Also, having the representation of so many MIPAS retrievals is somewhat distracting in terms of the end results which should state whether some instruments are clearly showing differences versus others, rather than too much of retrieval A versus retrievals B, C, D,... for the same instrument. An alternative approach would be to try to clarify the MIPAS results and summarise them by using some average result, although there may be too many ways to decide how to do this, which is why showing everything in a sort of politically correct way may still be the "best way". One could envision simply throwing out some of the outliers among the MIPAS results to show maybe only one or two (for the two main MIPAS periods) "cleaner" (or average/median) MIPAS results, and then discussing or showing more clearly how this compares to other satellite instruments/results.

General response #1:

We definitely do not want to combine different results from a given instrument. The focus is not really on instruments, but rather on the observational database to which all data sets contribute. The different MIPAS data sets all have their right to exist (and likewise multiple data sets from other instruments). The observations during the high resolution period (2002 - 2004) are different to those in the reduced resolution period (2005 - 2012). The observations in the nominal mode (NOM) differ from those in the middle atmosphere mode (MA). In addition all these measurements were obtained at different times. The different MIPAS processors use a wide range of choices to analyse the water vapour abundance in the atmosphere. This concerns the type of retrieval, retrieval constraints, a priori usage, altitude resolution, spectral signatures and spectral database considered, consideration of 2D variations, treatment of cloudiness, non-LTE, …, just to mention a few. The characteristics of the different MIPAS data sets will hopefully become much clearer once the WAVAS data set

overview paper is available. We are quite sure that the different choices are the main reason for the discrepancies among the MIPAS data sets from the different processors. Arguably we have considered some of the reasons listed above in Sect. 5.2, expanding this seems beyond the scope of this analysis. We also agree upon that a similar situation could easily occur for other instruments if a similar range of retrieval choices would be considered.

A certain motivation behind our decision to include all available MIPAS data sets was to give the data user an objective criterion which data set to use. Maybe just for a certain purpose or a larger range of case studies. We do not want either to judge which data set is best nor to do anything with the data that does not help the data user, as averaging all the data sets. Such averages will not be available from any server for the data user; they will have to work with the individual data sets.

It is somewhat disconcerting to see that sometimes the range of results is quite large, but as this is the reality, we also need to better understand the limitations/errors involved, which also includes errors from different retrievals. Such retrieval errors would also exist if one had several groups working on other instrument data, although the actual spread of results is unknown. In the end, it is clear that the same atmosphere is being observed by all these instruments, but this spread in results does make it somewhat challenging for a potential comparison to model results (not for this paper itself); how would a modeler go about choosing what to compare to (brief suggestions are certainly welcome)?

General response #2: We agree with sentiment that the spread of results is disconcerting occasionally. But this was also part of the motivation for this study.

As pointed out above we do not want judge which data set is the best for a specific comparison with model simulations. In a comparison to a model simulation it is easier to consider (or eliminate) effects of temporal and spatial sampling, the time period considered and the vertical resolution. The differences due to other reasons are clearly more difficult. Also the simulations will have some systematic errors. In the end we think that the standard deviations provided for the amplitudes and phases will be a good starting point.

In certain sensitivity studies, it would also help to try to clarify what may be able to specifically explain some of the differences between the results (in particular, regarding the sampling). Also, certain differences are probably not significant; by that I mean that since the results are obtained through regression fits, there should also be a way to obtain uncertainties in these results. For example, in places where the amplitudes (or phases) are small, the results may well not be significant, and/or the differences between the various results are not significant. In theory at least, this could be quantified better, so that not so much time is spent worrying about certain differences in the end. As this could represent a fairly large amount of work, this is just a suggestion to look into.

General response #3:

The sensitivity study for the sampling problem is arguably rather specific as it uses the temporal coverage of the ACE-FTS observations in the tropics as characteristics. As such it cannot explain why ACE-FTS v2.2 shows such a large amplitude for the annual variation around 3 hPa to 2 hPa in the tropics (see Fig. 2). This is far beyond the sensitivity. The main point we would argue is the large sensitivity in the Antarctic. A data set with non complete temporal coverage can easily exhibit differences of ±0.2 ppmv in terms of the annual variation amplitudes in this region. In relation to our Fig. 3 this concerns besides ACE-FTS (even though the temporal coverage at high latitudes is different than the tropics) also the MAESTRO, HALOE, POAM III, SCIAMACHY limb (at least partly), SCIAMACHY lunar occultation and SOFIE data sets. This has been noted in the text now.

The errors are a standard output of our regression analysis. We did not use these errors so far for several reasons. First of all it was never the intention to show if the variability characteristics of two data sets are significantly different or not. Our focus was more on the spread among the data sets in general (that was also a reason to not omit or combine multiple data sets from one instrument). In addition the errors of our regression analysis are probably not suited best to decide upon significant differences. As outlined in Sect. 3.2 the regression has been performed without any consideration of autocorrelation effects and empirical errors. As a consequence the errors are too small, not on par with the atmospheric variability. The autocorrelation effects and empirical errors can be accounted in different ways, which would it make more difficult to make our results easily reproducible. This was certainly a main rationale behind our analysis. The consideration of autocorrelation effects and empirical errors leads also to a reduction of available results. Tests for the annual variation have shown a reduction of more than 6%. This concerns primarily data sets with less observations and gaps in their temporal coverage.

For sensitivity studies we have now considered the errors (without any consideration of autocorrelation effects and empirical errors) and implemented the error propagation from the start to finish. The features discussed in Sect. 5.2 have been checked in this context. There have been some smaller changes, however mainly due to the retraction of the MLS v3.3/3.4 data set or the revised screening of the MIPAS-IMKIAA data at the uppermost altitudes.

Also, it would be better if this paper could refer back to some of the references listed (for example on page 3, lines 26/27) in terms of how the current results might differ, or agree, with previous findings.

General response #4:

In our analysis we have considered almost all satellite observations that were employed in the listed references. The only exceptions are Nimbus-7/LIMS (Remsberg et al., 1984), UARS/MLS (Carr et al., 1995; Mote et al., 1996) and UARS/CLAES (Mote et al., 1996). For the work of Remsberg et al. (1984) it is a bit difficult to say as they "just" present the latitudinal cross section in December and May. The work by Carr et al. (1995) focused on the semi-annual variation in the upper tropical stratosphere (our key feature #1). The amplitude that can be derived from the UARS/MLS observations (around 0.2 ppmv) falls well within the spread of our analysis. A similar conclusion can be drawn for the variability results in the lower tropical stratosphere as presented by Mote et al. (1996).

From the ground-based microwave observations presented by Seele and Hartogh (1999) amplitudes between 0.75 ppmv and 1.25 ppmv can be derived for the annual variation in the lower mesosphere at 69°N. This result is comparable to our work. Observations at Table Mountain (34°N) and Lauder (45°S) indicate smaller amplitudes in the same altitude region (Nedoluha et al., 1996), in good quantitative agreement with the results of our study.

It has not been a key motivation to compare our results to previous studies. Therefore, we have not added this little comparison to the manuscript.

Finally, using the present tense more abundantly would be useful in my opinion, and trying to make some data sets more visible could also help, although this is clearly a challenge, if the number of points is not reduced (I also imagine that a large number of colored lines is not really a viable solution).

General response #5:

In terms of colour optimisation we think we have reached our limit. As pointed out later in response #16 we have tweaked a bit the colour for the SAGE III data set to enhance visibility. Coloured lines we have tried before. This may work for the amplitudes, but for the phases it seemed not a good choice due their cyclic nature. For consistency we went along without any lines for the amplitude and phase. We also had discussions in the past in which sequence the data sets should be plotted. Currently it is done in a alphabetic order. A random approach would always change the figures (unless we keep this constant for a single figure). Some figures will be both in the manuscript here and the report later and in this context a random approach seemed not a good idea either.

After the above items are addressed as well as possible (without needing to embark on a large amount of extra work), this would definitely be a useful paper to see published. The paper's usefulness does depend on its clarity regarding the final results and potential explanations for some of the larger differences.

Specific comments:

Comment #1: I recommend that you use the present tense in most of the paper (just a suggestion); I find that there mixture of past and present could be improved upon.

Response #1: Present tense was used generally in Sect. 4, while there is a mixture in Sect. 5. There past tense is used in the result summary and also in the description of the sensitivity studies we performed and there out. This felt most natural to us. We attempted to employ this as consistent as possible and have looked into this again in the revised version.

Comment #2: The Abstract values for amplitudes would be useful as percentages, also probably in the Conclusion section (not just 0.2 ppmv, 0.1 ppmv).

Response #2: We had a number of discussions among the authors if the results should be shown on an absolute scale, a relative one or even both. Showing both would have blown up the size of the manuscript considerably and this idea was thus rejected. We decided upon the absolute scale since it does not require a reference explicitly. For the examples given in the comment we would probably need a range of percentages due to different altitudes considered. That is why we decided against percentage values.

Comment #3: P1, L3, probably worth adding "vapour" after "water", and also on L5. - L9, "In these regions, the standard deviation over all data sets..." - L10, "For the phase, the larger differences between data sets are [or were] found in the lower mesosphere."

Response #3:

L3 - Word added.

L9 - Start of sentence changed.

L10 - Left as is.

Comment #4: "The standard deviations of the phases for all data sets are [were] typically ..." - L2, The amplitude and phase differences among the data sets are probably caused by a combination of factors, including differences in temporal and spatial sampling, and temporal variations in systematic [retrieval?] errors. I would note that it would be helpful if you were able to point to the largest likely sources of such differences, given the work that has been done already, although it seems that this may be challenging to converge upon. - L10, add a comma after "greenhouse gas". - L20, change "effect" to "affect".

Response #4:

L1 - Changed as suggest using past tense.

L2 - The two sentences have been rewritten as: “The amplitude and phase differences among the data sets are caused by a combination of factors. In general differences in the temporal variation of systematic errors and in the observational sampling play a central role.”

We thought we had already given a qualitative assessment of the relative the importance, by mentioning the sampling biases and the time dependence of systematic errors first. To make this more obvious we included the word “central” now. The term “in general” should indicate that this is the typical picture for the entire latitude-altitude domain considered in our work. At specific latitudes and altitudes, however, some of the factors named later can be of larger importance. Beyond that it is very difficult to provide a quantitative assessment of the relative importance of the different factors.

L10 - Comma added.

L20 - Corrected.

Comment #5: P3, L23-24, Reword this, e.g. "A complete understanding of water vapour changes requires a good description of annual, semi-annual, and quasi-biennial oscillation variations (denoted here as QBO variation)." - L25, "shorter-term". - L33, "QBO variations which are subsequently summarised".

Response #5:

L23-24 - The sentence has been rewritten as follows: “A complete understanding of water vapour changes requires also a good knowledge of short term variability, such as the annual and semi-annual variation or the variation caused by the quasi-biennial oscillation (which we denote here as QBO variation).”

L25 - Fixed.

L33 - Changed.

Comment #6: P4, L16, "referred to the WAVAS-II...".

Response #6: Fixed.

Comment #7: P5, L3, "which we collectively..." - L20, maybe putting the "1 + " at the front of this equation would make it clearer.

Response #7:

L3 - Fixed.

L20 - Good idea. The “1” comes first now.

Comment #8: P6, L4, "with the phase denoting the time of maximum in the semi-annual fit between...". - L23, add a comma after "expected".

Response #8:

L4 - We started a new sentence, which goes as follows: “The phase denotes the time of the maximum in the semi-annual fit that is found between January and June.”

L23 - Done.

Comment #9: P7, L5, change "Additionally" to "In addition". - L22, "screened amplitudes"

Response #9:

L5 - Changed.

L22 - Changed.

Comment #10: P8, L1, The largest uncertainties were found in the lower mesosphere. For the reference data sets, we considered those that have a more or less... - L10, such plots are provided for all data sets considered in this work. - L11, allowing for a direct comparison. - L12, Finally, a summary is provided in the form of...in Sect. 3.3. - L20, vapour as a function of...

Response #10:

L1 - We changed this section to: “Quantitatively, the largest differences were found in the lower mesosphere. As reference data sets we considered those that have a more or less complete coverage of the latitude-altitude domain considered here, i.e. the MIPAS~V5H~NOM or V5R~NOM data sets from the different processors and the MLS data set.”

L10 - Changed.

L11 - Changed.

L12 - Changed.

L20 - Fixed.

Comment #11: P9, L2, I think you mean upper stratosphere (not sure why middle really). Also, late summer and early autumn would apply for Aug-Sep, which is really what you should state, since you are referring to the SH, where summer and autumn do not occur in Aug-Sep, I would say. - L5, allows for more effective downwelling of moister air from above, these moister values arising from methane photochemistry. - L12/13, occurs at a rather constant level... but more of a seasonal characteristic. - L20, except that feature #5 occurs at altitudes above the water vapour maximum. During winter,… - L26, add a comma after "coverage", and also after "data sets)", and say "or sometimes only a subset of those." - L31, as a function of altitude.

Response #11:

L2 - Yes, it should be the upper stratosphere. It appears that everybody has slightly different definitions what upper and middle stratosphere mean. Maybe referring to the upper half

would make everything easier? The reference to late summer and early summer is incorrect and has been fixed.

L5 - This part of the text has been rewritten. As indicated we are drafting a manuscript to characterise this feature more in detail. Our latest (and hopefully final) results indicate that the vertical transport is not as dominant for the inter-hemispheric differences as previously thought.

L12-13 - Changed.

L20 - Changed.

L26 - Fixed.

L31 - Changed.

Comment #12: P10, L16, "on the order of" [please change this everywhere, e.g. on L26, L35 too] - L19, data sets did not have sufficient temporal coverage... - L25, 20 hPa, good agreement ... - L29, deviate in obvious ways from...

Response #12:

L16 - All occurrences replaced.

L19 - Changed.

L25 - Changed.

L29 - Changed.

Comment #13: P11, L9, while the majority of data sets - L14, in the form of - L20, is revealed [not reveals] - L20, The former lie typically in regions where

Response #13:

L9 - Fixed.

L14 - Fixed.

L20 - Changed to “the latter” as the text was referring to the standard deviation in relative terms.

Comment #14: P12, L4, Note that the amplitude - L32, a breakup of the vortex breakup sounds a bit strange. Do you mean an interruption of the vortex breakup?

Response #14:

L4 - Fixed.

L32 - Simply the vortex breakup was meant. The doubling has been removed.

Comment #15: P13, L9, SSAO allows waves to propagate further up only if they have horizontal propagation directions opposite to the zonal wind...

Response #15:

L9 - Changed.

Comment #16: P14 - Fig. 9 is an example where SAGE III data, in yellow, are very hard to read/find even on a screen view of the plot(s), and on a printed version too. - L30, do you mean "despite the small absolute standard deviations" or "as a result of the small absolute standard deviations"?

Response #16:

Fig. 9 - As a little tweak, we changed the colour from “yellow” (FFFF00) to “yellow2” (EEEE00) according to the RGB standard colour table. This is a little bit darker but still clearly distinguishable from MLS (“gold2”, EEC900). A further change of the colour scheme should be avoided since any alteration means that everyone in the WAVAS consortium has to change the plots again.

L30 - “Despite” is the word we had in mind.

Comment #17: P15, L2, delete "above". - L16, adding a bit more clarification regarding the "profound effect" [maybe from the Jackson ref.] would be useful. - L22, add a comma after "feature #1". - L31, This is also a characteristic of key feature #2...

Response #17:

L2 - Done.

L16 - We have slightly rewritten the text. In particular we changed terminology from “profound effect” to “interaction”. Essentially the QBO modulates the vertical eddy transport of westerly momentum originating from Kelvin and gravity waves. As this momentum is responsible for the westerly forcing of the SAO the two variability patterns are interlinked in the upper tropical stratosphere.

L22- Fixed.

L31 - Changed.

Comment #18: P16, L26, there are a few data sets that...

Response #18:

L26 - Fixed.

Comment #19: P17, L2, add a comma after "patterns". - L14, change "lowest" to "poorest". - Summary, here (again), it would sound nicer if the present tense could be used. - L26, add

a comma after "patterns" - L29, add a comma after "the key features" - L30, add "and at" before "high latitudes"

Response #29:

L2 - Fixed.

Tense: As pointed out in response #1 we used the past tense to describe the results that we had obtained. As a summary comes first after that it seemed natural to used also in the past tense there.

L14 - Changed.

L26 - Fixed.

L29 - Fixed.

L30 - Added just “at”.

Comment #20: P18, L2, 86%, and considering a standard deviation... - L4, change "could" to "can" or just "are often observed".

Response #20:

L2 - We actually started a new sentence here.

L4 - We left it as is.

Comment #21: P19, L5, add a comma after data sets. - L14, sampling bias is due not only to the actual... - L15, but also to the atmospheric variability... - L17, We investigated the influence of incomplete coverage throughout the year. - L22, Contrary to this, in the middle stratosphere... and change "could be" to "was" or "is". - L26, changes exceed[ed] 50% on occasion. - Also, did the sub-sampled data sets help to *improve* the comparisons in these sensitivity test, it is not clear what the result really is [it would make sense if it did, otherwise, there are other factors that cause the differences]?

Response #21:

L5 - Fixed.

L14-15 - We think there is something wrong with the suggestion. We rewrote the sentence as follows: “This is due to the fact that the variability within each time and latitude bin can also cause a sampling bias, besides the sampling pattern itself.”

L17 - Actually it was important to us to mention that we investigated upon one aspect of the sampling bias. In accordance we wrote now: “We investigated the influence of incomplete coverage throughout the year as one aspect of the sampling problem.” We also started a new paragraph to indicate a logical separation between the discussion above and the sensitivity study that follows.

L22 - Changed.

L26 - Text changed. The primary objective of this sensitivity study was to get a feeling how the variability characteristics change if the sampling is not complete temporally. Looking at the standard deviations among the data sets we found that it is generally smaller when the temporal sampling is complete compared to the case with measurements only during the four months. We have have added that result to the text.

Comment #22: P20, L3, The intermediate frequency holds the information... - L11, a time variation of some systematic errors (in particular above 30 hPa) that affects the results - L22, in the vicinity of the hygropause

Response #22:

L3 - Changed.

L11 - Changed.

L22 - Fixed.

Comment #23: P21, L4. These results are fairly useful, and it would be good to expand a bit on the specifics, meaning what sort of change in resolution leads to a 50% change in amplitude? It is somewhat unfortunate also that you do not give a specific example or two whereby this aspect could really help to make the comparisons of amplitude (or phase) better between specific retrievals. - L21, delete "the data sets used here are based on" - L24, In a similar way, the observations are affected by aerosols.

Response #23:

L4 - We have expanded the text to provide a better grasp for the influence af the altitude resolution on the variability characteristics in these particular regions.

L21 - Changed.

L24 - Changed.

Comment #24: P22, L10, only on the latitude range - L14, polewards of 45 - L15, add a comma after variation - L17, accompanied by a substantial decrease of the agreement - L23, MIPAS data sets increased significantly polewards of 35... - L29, There are many ways to derive the characteristics of different variability patterns. - L32, add a comma after "Sect. 5.2".

Response #24:

L10 - Fixed.

L14 - Fixed

L15 - Fixed.

L17 - Fixed.

L23 - Changed.

L29 - Changed.

L32 - Fixed.

Comment #25: P23, L18, shows a sample comparison - L33, which show both negative ... The differences between ...

Response #25:

L18 - We removed the word “example”.

L33 - Changed.

Comment #26: P24, L25, add a comma after altitude. Also, add "and" before "in extreme cases". - L30/31, I would delete "as the variability in those is typically large" [or clarify]. - L33, change "could be" to "are".

Response #26:

L25 - Comma added and text changed.

L30-31 - Text was rewritten as follows: “Many regions with large spreads coincided often with large variability, resulting in relatively small relative standard deviations (<50\%). Contrary, regions with small spreads in amplitude often exhibited low variability, leading to larger relative standard deviations (>50\%).”

L33 - Left as is for consistency tense-wise.

Comment #27: P25, L4, Other reasons include the different time periods used... [and delete "also contribute to the differences" at end of sentence]. - L12, tape recorder, that still exhibits fundamental differences versus observations,...

Response #27:

L4 - Changed.

L12 - Comma added and text changed.

Comment #28: Fig. 5 caption, Line 4, change "with respect of the MLS" to "with respect to the MLS".

Response #28: Fixed.

Comment #29: Fig. 6 caption, sentence 3, change "were" to "where".

Response #29: Fixed.

Comment #30: Fig. 9 caption, change Fig to Figs

Response #30: Fixed.

Comment #31: Supplement text, page 1, 3 lines from bottom, also exhibit some outliers (?).

Response #31: The section was rewritten as: “Figure 1 provides a good first overview of the spatial coverage of the individual data sets. It also shows those outliers in the variability amplitudes which we screened to provide reasonable estimates of the uncertainties of the variability characteristics (see Sect. 3.3).”

Comment #32: Also, Fig. 2 caption there, change "were" to "where", and in Fig. 3, change "phased" to "phase".

Response #32: Both fixed.