Author response to reviewer's comments on

"Comparison of ground-based and satellite measurements of water vapour vertical profiles over Ellesmere Island, Nunavut"

by Weaver et al.

Reply to Reviewer #2

The authors would like to thank reviewer #2 for their attention to detail and helpful comments.

The reviewer's comments are included in italics. Replies are in blue.

Major comments:

Figures 6 and 9 appear to be identical. It is impossible that they can look exactly the same given what they are meant to show and the obvious differences between the 125HR and RS92 profiles in Figure 5. Also, values stated in the text for specific satellite-RS92 differences don't match up with what's shown in Figure 9. See specific examples below for pages 15 and 16. Finally, Figure 9 shows difference profiles for MIPAS and SCIAMACHY vs RS92 while the text in Section 3.2.4 explicitly says that no MIPAS or SCIAMACHY measurements were coincident with radiosondes. As Figures 6 and 9 are the most important Figures in this paper, it became impossible to continue my review past page 15. My hope is that the authors not only include the correct Figure 9 in the next version, but also take to heart the remainder of my comments and those of the other reviewer(s) that will improve the paper.

(1) The correct version of Figure 9 was included in the initial submission of the manuscript during submission; however, minor modifications to improve the readability were suggested during the technical review. When updating the file for re-submission, the lead author mistakenly included a second copy of Figure 6 where Figure 9 should have been. This has been corrected, and should satisfy the other concerns raised about consistency between the text and figure. We apologize for this unfortunate mistake.

I think there are also problems with some of the mean bias values in Tables 2 and 3. For example, for the MIPAS IMK retrievals (v5 and v7) at 12 km in Table 2. The mean difference from the 125HR is given as -0.3 ppmv and -1.4%. If the biases that produce these values are normally distributed, they imply that the mean MIPAS retrieval at 12 km is between 18 and 25 ppmv (-0.25/0.014 and -0.35/0.014). This is way too wet for stratospheric air, and is 3 to 4 times the mean MIPAS IMK retrieval at 12 km (approx. 6 ppmv) shown in Figures 6a and 9a. Another example of this problem is found in Table 3.

Water vapour abundances near 20 or 50 ppmv would indeed be well outside expected values in the stratosphere and were not observed in the measurements presented. This can be seen in the panel (a) of the profile comparison figures (i.e., Figures 5, 6, and 9), which show the mean abundances of profiles used for comparisons in this study.

We have calculated the mean absolute difference at each altitude level using:

$$\Delta_{abs}(z) = \frac{1}{N(z)}\sum_{i=1}^{N(z)}[X_i(z) - Y_i(z)], \quad (1)$$

and the mean relative difference using the mean of the percent differences as:

$$\Delta_{rel}(z) = 100\% \times \frac{1}{N(z)}\sum_{i=1}^{N(z)}\frac{[X_i(z)-Y_i(z)]}{Y_i(z)}, \quad (2)$$

rather than calculating the relative difference between the mean profiles using, i.e.:

$$\Delta_{mean}(z) = 100\% \times \frac{\frac{1}{N(z)}\sum_{i=1}^{N(z)}X_i(z) - \frac{1}{N(z)}\sum_{i=1}^{N(z)}Y_i(z)]}{\frac{1}{N(z)}\sum_{i=1}^{N(z)}Y_i(z)} = 100\% \times \frac{\sum_{i=1}^{N(z)}[X_i(z) - Y_i(z)]}{\sum_{i=1}^{N(z)}Y_i(z)}. \quad (3)$$

The absolute difference and percent difference can be combined to calculate the typical abundances only if the percent difference has been derived using the mean profiles of two datasets, e.g. using Equation 3. This cannot be done if the percent difference is derived using the mean of the individual differences and percent differences, e.g., using Equation 2. To ensure the method we used is clear, Equations 1 and 2 have been added to the text of the methods section.

To illustrate the importance of this distinction, let's consider the comparison between MIPAS (IMK v7) and the 125HR at 12 km.

The mean MIPAS abundance was 6.5 ppmv and the mean 125HR abundance was 6.8 ppmv.

Calculating the individual differences between coincident measurements and taking the mean, i.e. applying Equations 1 and 2, results in the following:

Δ_{abs} (12 km): −0.3 ppmv
Δ_{rel} (12 km): −1.4%

If these values were combined to calculate 'typical' abundances, the result would be inaccurate and misleading, as pointed out by both reviewers.

If we were instead to apply Equation 3, ie., to calculate the percent differences using the difference between the mean profiles, we get:

Δ_{abs} (12 km): −0.3 ppmv
Δ_{mean} (12 km): −4.4%

If we calculate a typical abundance from these values, we get:

$H_2O = \Delta_{abs} / \Delta_{mean}$ = 0.3 ppmv / 4.4% = 6.8 ppmv

This is the original reference value for water vapour abundances, and how the both reviewers expected the numbers to be related.

However, if we examine the mean of the differences, rather than the difference of the means, this calculation of typical abundances is no longer possible.

We could also consider a simple example of two datasets, X and Y, so that the full calculation and numbers can be readily written out:

X = (1, 3, 5)
Y = (2, 2, 8)

The mean of X is: 3
The mean of Y is: 4

The difference between the two means is: −1
The percent difference between the two means (Δ_{mean}) is: −25% (using Y as the reference).

However, we get a different percent difference by taking the mean of the individual percent differences:

$$\frac{X-Y}{Y} * 100\% = \left(-\frac{1}{2}, \frac{1}{2}, -\frac{3}{8}\right) * 100\% = (-50\%, 50\%, -37.5\%)$$

Mean percent difference (Δ_{rel}) = $-$ 12.5%

Only in the first case, i.e., the percent difference between the means, can the original value be recovered, i.e.:

−1 / −25% = 4,

i.e., the original mean of Y.

…. Mean bias values for AIRS vs RS92 at 12 km are -2.0 ppmv and +5.2%. How can the mean absolute bias (ppmv) and mean relative bias (%) be of opposite signs if the biases are normally distributed? Either there are errors in the mean values presented in these Tables or the distributions of the differences that produce the mean biases are very skewed. If the former, please double check the Table values and make corrections. If the latter, quantifying the biases using Gaussian statistics (i.e., mean and standard error of the mean) is not warranted.

The distributions of the differences are generally Gaussian. For example, Figure 1 shows a histogram of the differences between AIRS and 125HR measurements at 6.4 km.

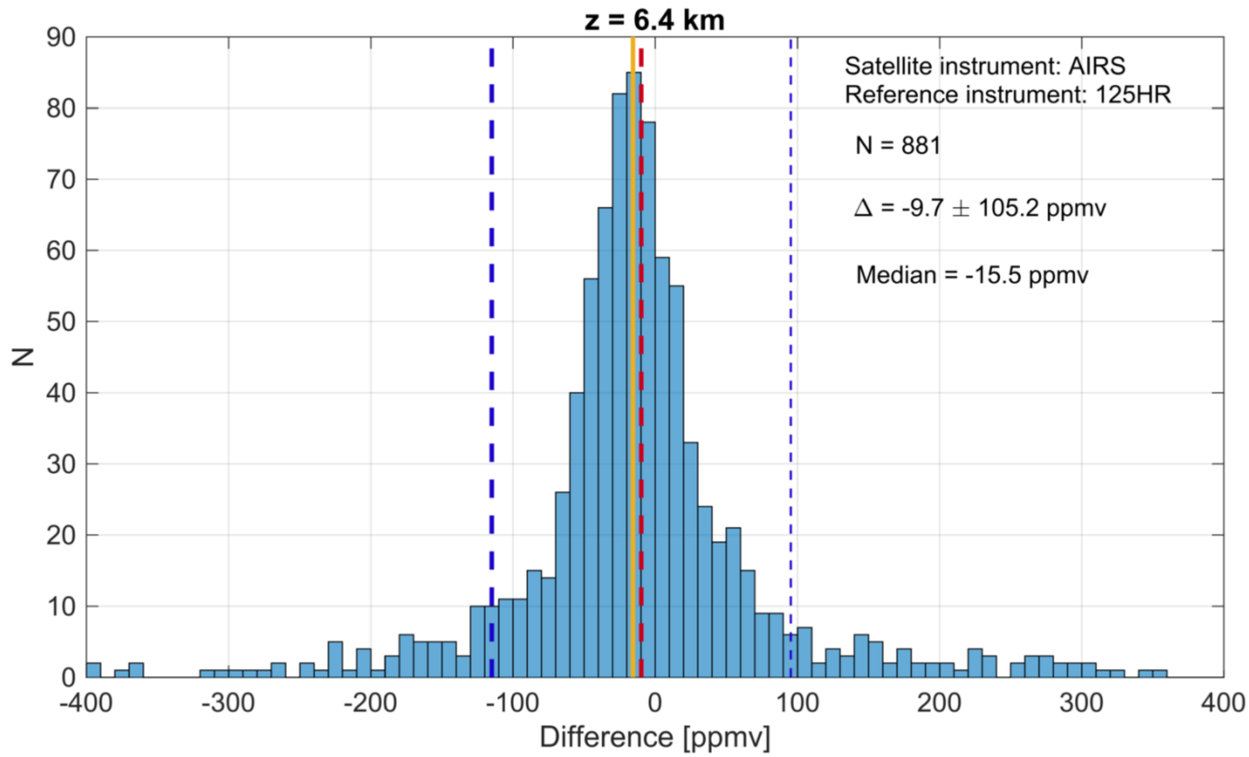

Figure 1: Histogram of differences between AIRS and 125HR water vapour measurements at 6.4 km. The dashed red line is the mean of the differences; the blue dashed lines show one standard deviation above and below the mean. The solid tan line shows the median of the differences.

In a few cases, the sign of the absolute and percent differences are not the same. There are a few reasons for this.

In some cases, e.g., AIRS vs. radiosondes comparison at 12 km, the number of coincidences is relatively small (N = 50). As the number of coincidences becomes small, we expect the approximation of a Gaussian distribution to be less justified. Indeed, in the case of ACE-FTS vs. 125HR at 6.4 km, the standard error in the mean indicates the mean absolute and percent differences, which are in this case of opposite sign, are not significantly distinct from zero.

In other cases where there is large number of coincidences and a roughly Gaussian distribution, there is a small skewness in the distribution that differs enough between the absolute differences and percent differences that the means land on opposite sides of zero. The skewness is not large, but has this effect because the mean of the overall distribution is close to zero relative to the range of values involved. Histograms of absolute differences and percent differences between AIRS and GRUAN at 6.4 km are shown in Figure 2 and 3. These illustrate how the small differences in the skewness of the absolute difference and percent difference distributions, nearly centered at zero, can have means with opposite signs. Also note that the medians of the absolute differences (Figure 2) and percent differences (Figure 3) have the same sign (negative).

If the median differences and percent differences are examined, all comparisons have the same sign at all altitudes. Median differences have been added to Table 2 and 3, which summarize the results.

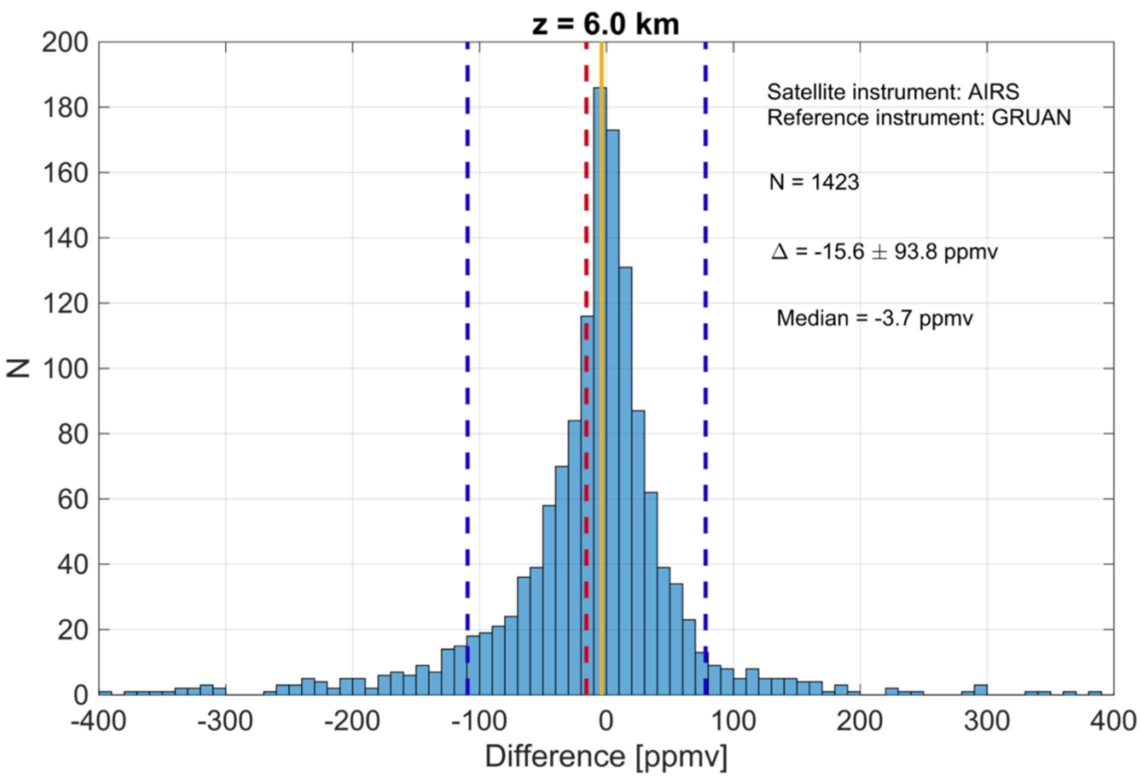

Figure 2: Histogram of absolute differences between AIRS and GRUAN-processed radiosonde water vapour measurements at 6 km. Lines defined as in Figure 1.

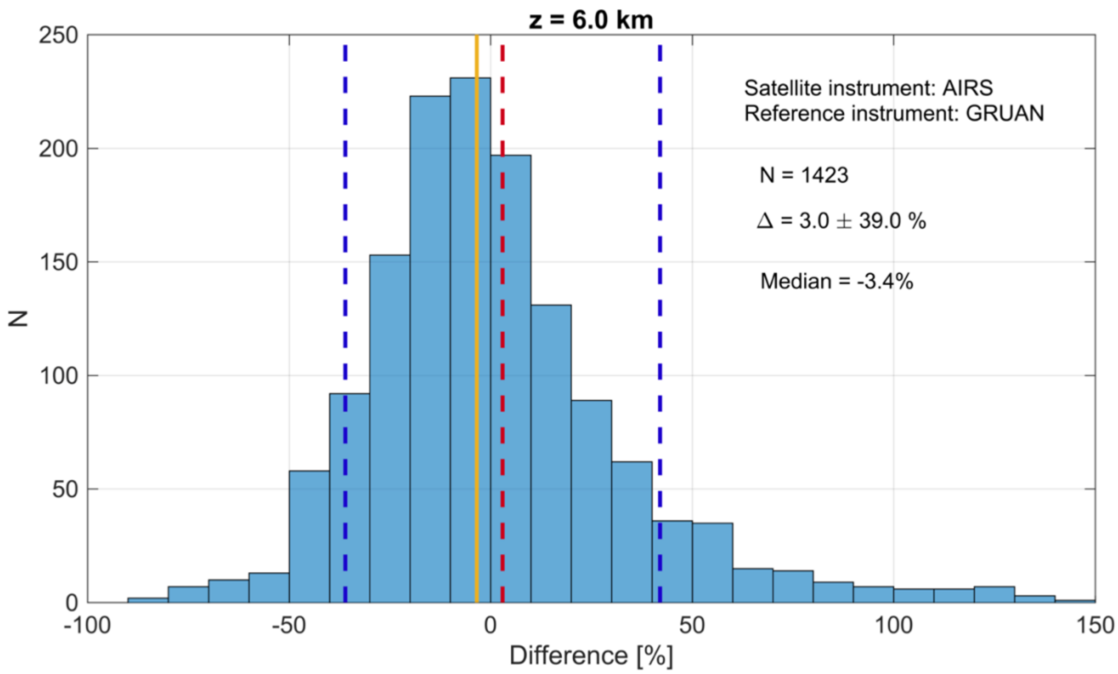

*Figure 3: Histogram of percent differences between AIRS and GRUAN-processed radiosonde water vapour measurements at 6 km. Percent differences calculated using: (AIRS – GRUAN)/GRUAN * 100%. Lines defined as in Figure 1.*

Correlation coefficients and correlation plots are of limited quantitative value in a paper focused on measurement biases between pairs of instruments. Two sets of measure- ments can be well correlated even though there are huge biases between them! Cor relation plots can show biases, but only qualitatively, so consider if the three Figures with correlation plots reveal any quantitative information not already revealed by the profile differences and/or time series of differences. If the correlation plots are deemed unnecessary (my opinion), some (if not all) of the Supplemental Figures could become part of the main manuscript. Please see my specific comments below for Page 15 Line 1 (P15 L1).

It is true that correlation coefficients need to be carefully interpreted. In this study, they are used in combination with the differences to show how well the measurements agree. In particular, the correlation plots illustrate how closely the measurements agree and how much variation in the differences exist, i.e., the spread in the values. This information is also shown in the difference timeseries; however, it is useful to examine the datasets as a whole – e.g., not as a timeseries. This can reveal, for example, if there are measurement biases or differences that affect measurements at larger vs. smaller abundances. The use of these plots is common in the validation literature, e.g., with the FTIR MUSICA product (Schneider et al., 2010), other water vapour measurement techniques (Buehler et al., 2012), ACE and OSIRIS satellite products (Adams et al., 2012), and other satellite missions such as GOSAT (Frankenberg et al., 2013; Ohyama et al., 2017) and MLS (Vömel et al., 2007). The use of the correlation coefficient as a part of an overall assessment of agreement between datasets has been even more widely used, e.g., for comparisons between ACE-FTS profiles and other satellite datasets (Sheese et al., 2016).

The Introduction describes the importance of water vapor in the UTLS and how accurate measurements of WV in the global UTLS are needed. The focus of the paper therefore seems drawn towards WV measurement biases in the UTLS. But this focus becomes lost when you start to compare WV measurements at altitudes as low as 1 km. Why do you apply the same spatial and temporal coincidence criteria to the stratospheric and lower tropospheric data even though the spatiotemporal variability of WV in these regions is very, very different? My advice is to focus this paper on the crucial UTLS region and leave out or downplay the lower tropospheric comparisons.

One of the key questions to be answered in the work is "how low, in altitude, can ACE profile measurements of water vapour be trusted?" This question necessarily involves measurements as low as 4.5 km (in the case of ACE-MAESTRO). While the paper could exclude profiles with values below this altitude, it is useful to see the comparisons at all available altitudes for the retrieved profiles of the datasets used so that the context of the observed agreement at altitudes of particular interest are interpreted in their full context. For example, if the AIRS measurements were to suddenly diverge from the radiosondes at 4 km, and show a large bias in the lower troposphere, their agreement in the upper troposphere would be placed in a different context than the consistent agreement observed throughout the troposphere in this study. Also, including these available results gives a more complete assessment of what vertical profiles are available at Eureka. Moreover, the results at tropospheric altitudes motivate a study that focuses on the use of AIRS water vapour data in the high Arctic.

That said, it is certainly true that the spatio-temporal variability increases greatly at lower altitudes with important consequences for the selection of coincidence criteria. For that reason (and others), tighter coincidence criteria were examined. In Section 3.2.4, which discusses the AIRS comparison results, the paper notes that a much tighter coincidence criteria of 25 km and 2 hours shows similar comparison results.

General Comments:

P2 L20 what exactly does "modest vertical resolution" mean? Please be more quantitative here. The vertical resolution of FTIR measurements is very important information for this paper that compares satellite retrievals to the FTIR measurements.

The vertical resolution of the FTIR measurements varies; the mean DOFS are 2.9 for the PEARL 125HR MUSICA product. This has been added to the text in Section 2.2's description of the dataset used in this study, as suggested in the comment for P6 L20.

P2 L20-22 Radiosonde humidity sensor measurements also require substantial corrections for solar radiation effects, calibration biases and slow response times in the cold UTLS. It surprises me that frost point hygrometers and lidars are not mentioned here even though the current global coverage of frost point hygrometer sounding sites is starting to surpass the coverage of FTIRs.

The paper's introduction focused on the approaches used in this study and those that are most widely used. While frostpoint hygrometers (FPHs) offer definite advantages over radiosondes and FTIR spectrometers, their geographic deployment is much less widespread than radiosondes. In addition, they typically acquire measurements less often than radiosondes and FTIRs, i.e., some sites launch them only monthly, and their data timeseries are usually shorter. Moreover, there have been no FPH measurements taken from Eureka. This is regrettable. The nearest sites where FPH measurements are taken are Ny Ålesund, Svalbard, and Barrow, Alaska, which are both roughly 2000 km away. For these reasons, FPH measurements are noted in the conclusions as a promising area of future work, as it would be valuable to add them to the suite of instruments at PEARL/Eureka.

P2 L28 "assessing the accuracy and quality" - what does quality mean here if not accuracy?

"Quality" has been removed as redundant, as suggested.

P3 L1 I believe UT WV measurements will also be compared, not just those in the stratosphere and lower mesosphere.

This sentence has been reworded to include the upper troposphere.

P3 L26 move lat/lon to L20 (description of Eureka location)

This has been done.

P4 L23 why is the humidity sensor "no longer able to report a meaningful value"? Is it the cold ambient temperature? Is it the low number density of WV? The solar heating effects on the sensor? Please be more specific.

Original text in that paragraph and the following paragraph notes that Miloshevich (2009) shows that the RS92 radiosonde capacitance sensor responds accurately at low temperatures (–70°C) and at low abundances (5 ppmv) but that low pressures are a limiting factor.

P5 L2 why describe the Miloshevich et al. (2009) limits when Dirksen et al. (2014) improves the correction algorithms and expands the upper altitude limits of "meaningful" RH measurements by the RS92?

Dirksen et al. (2014) improves the correction algorithms, but the resulting GRUAN data product does not set out upper altitude limits. This motivates the use of a filtering approach for this study based on the calculated uncertainties resulting from the Dirksen analysis technique. In a few cases, the uncertainty of the GRUAN-processed humidity profile remains below the filtering threshold well above 15 km, e.g., to 25 km. Out of an abundance of caution, the altitude limit suggested by Miloshevich et al., 100 hPa, on the radiosonde measurements was cited and applied as an additional quality control filter, resulting in any profile at Eureka that passes the uncertainty filtering being limited to a maximum height of 15 km, which is approximately the altitude of 100 hPa. This is also roughly the boundary for the upper troposphere and lowermost stratosphere, the area of specific interest of this study.

P6 L20 this would be a good place to mention the vertical resolution of the MUSICA FTIR WV profiles

This has been added:

“The mean degrees of freedom for signal (DOFS) of the Eureka MUSICA retrievals are 2.9.”

P7 L20 "Correlations between ... were observed to be greater than ..." Why are correlations important in this inter-comparison? Two data sets can be extremely well correlated, even when there is a very large bias between them. Correlation is not a good measure of the agreement between two data sets.

Correlations have been discussed above in the reply to the major comments.

P8 L2 what is the vertical resolution of ACE-MAESTRO WV retrievals in the UT and LS?
P9 L11 what is the vertical resolution of Aura MLS WV retrievals in the UT and LS?
P9 L25 what is the vertical resolution of Aura TES WV retrievals in the UT and LS?

Approximate values are given in the instrument descriptions as available.

The ACE-MAESTRO vertical resolution is approximately 1 km.

The MLS 4.2.x product document states that the vertical resolution of the water vapour profiles is 1.3 – 3.6 km between 316 and 0.22 hPa. The altitudes used in this study are 316 hPa and the levels immediately above it, putting the resolution closest to the 1.3 km end of the range.

TES vertical resolution varies by altitude, latitude and species. The DOFS have been improved in the most recent version (6) used here, with DOFS between 3 and 5. However, at polar latitudes, in the UTLS, the vertical resolution is 11.6 km, while in the troposphere it is 6.0 km (Worden et al., 2004).

The text has been revised as follows:

> "The ACE-MAESTRO water vapour retrieval algorithm produces profiles with an approximate vertical resolution of 1 km, and is described by Sioris et al. (2010) with updates described in Sioris et al. (2016)."
>
> ….
>
> "MLS water vapour profiles are vertically resolved at pressures less than 383 hPa, with a vertical resolution ranging between 1.3 and 3.6 km from 316 to 0.22 hPa (Livesey et al., 2016)."
>
> ….
>
> "The vertical information content of TES profiles varies; retrievals with less than 3 DOFS are filtered out. In the subset of measurements examined in this study, TES DOFS range between 3.0 and 5.2. At polar latitudes, the vertical resolution is approximately 11.6 km between 400 and 100 hPa and 6.0 km between 1000 and 400 hPa (Worden et al., 2004)."

P10 L22 Stiller et al. (2012) compared MIPAS with many types of WV instruments including frost point hygrometers, lidars, microwave radiometers and an FTIR, not just the CFH.

While the Stiller et al. (2012) study included comparisons to other instruments, the comparison to the CFH was most relevant to the discussion here, as it was the best reference measurement.

P10 L25 "suggest" and "might be" are very waffly terms. Are there 20-40% biases or not?

Conclusive statements regarding the bias of an instrument cannot be derived from comparisons at a single site. The term 'suggest' is intended to convey that these specific results are to be interpreted in the context of the wider validation literature. The specific use of these terms in this instance reflect the terms used by Stiller et al. to describe the results of the cited work.

*P11 L5 Weigel et al. (2016) also compared *SCIAMACHY* v3.01 (not MIPAS v3.01 as written) to in situ instruments made from balloons (FPH) and aircraft (FISH), not just other satellite retrievals.*

Thanks - correction made.

P12 L2 Closest in time or space? How did you determine the time stamp for FTIR spectra, which are often co-added for minutes or hours? Also, radiosondes reach 10 km about 30 minutes after they are launched, so how did you set the timestamps for the RS92 profiles?

The closest pair in time were kept. The timestamp for the FTIR spectra were the scan start time. Scans took about 5 minutes, following standard NDACC procedures and settings. The timestamp for radiosondes was the launch time. These clarifications have been added to the respective descriptions of the datasets.

P12 L9 if the results of comparisons using the closest satellite profile are similar to the results using all coincident profiles, why do you need to show the latter in Supplemental Tables?

The comparisons using all coincident profiles was offered in the supplemental materials for reader's interest, to demonstrate the accuracy of the statement that the results are similar (they are not identical), and to provide a complete record that might be useful for future studies that might want to compare results that use this approach rather than the paired approach used in the main manuscript.

P13 L15-18 "... effectively synthesizing a narrow weighting function, then is possible from any one channels. We use of the width ... to estimate a Gaussian smoother generally overestimates ..." These sentences are very poorly constructed. Please fix them.

The first sentence has been removed while the second has been revised to be:

> "We use of the width of the AIRS weighting functions to estimate a Gaussian smoothing width that generally overestimates the amount of smoothing."

P14 L4 Above, you stated that the FWHM approximates the vertical resolution of the measurement. So why then do the weighting functions for MLS have a FWHM or 1.0 km when the vertical resolution of MLS retrievals is more like 2-3 km?

The MLS data quality document specifies (page 66) that the vertical resolution of the water vapour profile ranges from 1.3 – 3.6 km from 316 – 0.22 hPa. The altitudes of interest are at the highest pressure (lowest altitude) of that range, thus closer to 1 km than the 3-4 km typical of stratospheric altitudes.

P14 L22 Are the 8% and 6% mean differences significantly different from zero? In other words, what are the standard errors of these mean values? It they are not statistically different from zero I would hesitate to call them "biases" because you have no evidence that they are real biases, just mean differences that may equal zero.

The standard errors on the approximate 8% difference between the 125HR and the radiosondes under 8 km altitude ranged between 1 and 3%, suggesting a real difference. SEM values are provided both in the text when specific altitude results are given, and also in the summary of results in Tables 2 and 3. In addition, inspection of individual coincident profiles frequently show a negative RS – 125HR difference. However, caution in this result is justified, given that the expected accuracy of this FTIR water vapour profile retrieval is approximately 10%. Additional text has been added to clarify the standard errors and remind the reader of the expected precision of the FTIR profiles.

P14 L28 I can't see any ACE-FTS differences between 6 and 9 km in Figure 6b that exceed 9 ppmv, so why do you say "was within 11 ppmv"? Also, why report differences for this altitude range when they change from negative to positive at 7 km then become much smaller (in ppmv) and consistent (in ppmv and %) at 8 km and above?

–11.0 ppmv is the difference between ACE-FTS and 125HR at 5.6 km altitude. The text has been revised to state they agree within the suggested 9 ppmv in the 6 – 9 km altitude range. This range had been reported for comparison with other instruments. The text states that the differences are smaller above 8 km, i.e., "between 8 and 14 km, agreement is within 1.4 ppmv and 10%".

Figure 6 I suggest using fewer red and purple curves, as they are difficult to tell apart. Replace some of them with green, orange and gray. Also, I am guessing that you discuss satellite-125HR mean differences at 6.4, 8.0 and 9.8 km because these are the altitudes of 125HR retrievals?

In this study, each instrument is given a colour, which is used consistently across all figures. The suggested colours are used for other instruments, some of which are not in this figure, but are in others. For consistency across figures, the colours have been kept as they are.

Yes, 6.4, 8.0, and 9.8 km are altitudes from the FTIR retrieval grid. This has been noted in Section 3.1.2, in the description of the method:

> "Comparisons between satellite measurements and the FTIR are thus presented on the MUSICA retrieval altitude grid, e.g., 6.4 km, 8.0 km, and 9.8 km."

P15 L1 and Figure 8 I don't see the value of the correlation coefficients or the correlation plots. The focus of this paper is biases. Correlation coefficients can be near unity when biases between instruments are huge! The correlation plots reveal only qualitative information about biases. For example, the linear fits to ACE-MAESTRO vs 125HR show really awful correlations and essentially no quantitative information about biases. The AIRS panels show good correlations and (qualitatively) that AIRS is biased low at 6.4 and 8.0 km because most of the differences lie below the 1:1 line. What does this Figure (and Figures 11 & 12) show that the vertical profiles of mean differences and time series of differences don't show?

Correlations have been discussed above in the reply to the major comments.

Figure 9 I cannot find a single difference between this Figure and Figure 6, even though they are meant to be showing differences from the RS92 sondes and 125HR, respec- tively. The two Figures appear to be identical, even when printed, stacked, and held up to backlighting. Are you sure Figures 6 and 9 are actually showing what they are intended to show? The only way they can be exactly the same is if the RS92 and 125HR mean differences are very close to 0 ppmv and 0%, which they are not (Figure 5). The mean differences presented in the text (P15 L7-8) and in Figure 9 do not agree. I suspect Figure 6 appears a second time as Figure 9 in this manuscript.

This correction has been made and was discussed above in the reply to the major comments.

P15 L19 Your statement here "scatter around the zero line" contradicts what you just concluded, "a dry bias of approx. 10%". The dry bias in ACE-MAESTRO vs 125HR is apparent in Figure 7, so the "scatter" is not "around the zero line" as stated, otherwise there would be no bias.

This sentence has been revised.

P16 L10 "Differences as large as 13% are observed between 8 and 14 km." The suspicious Figure 9 shows no relative differences (AIRS-RS92) exceeding 5% between 8 and 14 km.

This disconnect is due to the aforementioned Figure 9 issue, which has been corrected.

References

Adams, C., Strong, K., Batchelor, R. L., Bernath, P. F., Brohede, S., Boone, C., Degenstein, D., Daffer, W. H., Drummond, J. R., Fogal, P. F., Farahani, E., Fayt, C., Fraser, A., Goutail, F., Hendrick, F., Kolonjari, F., Lindenmaier, R., Manney, G., McElroy, C. T., McLinden, C. A., Mendonca, J., Park, J.-H., Pavlovic, B., Pzamino, A., Roth, C., Savastiouk, V., Walker, K. A., Weaver, D., Zhao, X.: Validation of ACE and OSIRIS ozone and NO_2 measurements using ground-based instruments at 80°N. Atmospheric Measurement Techniques, *5*(5), 927–953. https://doi.org/10.5194/amt-5-927-2012, 2012.

Buehler, S. A., Östman, S., Melsheimer, C., Holl, G., Eliasson, S., John, V. O., Blumenstock, T., Hase, F., Elgered, G., Raffalski, U., Nasuno, T., Satoh, M., Milz, M., & Mendrok, J.: A multi-instrument comparison of integrated water vapour measurements at a high latitude site. Atmospheric Chemistry and Physics, *12*(22), 10925–10943. https://doi.org/10.5194/acp-12-10925-2012, 2012.

Frankenberg, C., Wunch, D., Toon, G., Risi, C., Scheepmaker, R., Lee, J. E., Wennberg, P., & Worden, J.: Water vapor isotopologue retrievals from high-resolution GOSAT shortwave infrared spectra. Atmospheric Measurement Techniques, *6*(2), 263–274. https://doi.org/10.5194/amt-6-263-2013, 2013.

Ohyama, H., Kawakami, S., Shiomi, K., Morino, I., & Uchino, O.: Intercomparison of XH_2O data from the GOSAT TANSO-FTS (TIR and SWIR) and ground-based FTS measurements: Impact of the spatial variability of XH_2O on the intercomparison. *Remote Sensing*, *9*(1). https://doi.org/10.3390/rs9010064, 2017.

Schneider, M., Romero, P. M., Hase, F., Blumenstock, T., Cuevas, E., & Ramos, R.: Continuous quality assessment of atmospheric water vapour measurement techniques: FTIR, Cimel, MFRSR, GPS, and Vaisala RS92. *Atmospheric Measurement Techniques*, *3*(2), 323–338. https://doi.org/10.5194/amt-3-323-2010, 2010.

Sheese, P. E., Walker, K. A., Boone, C. D., McLinden, C. A., Bernath, P. F., Bourassa, A. E., Burrows, J. P., Degenstein, D. A., Funke, B., Fussen, D., Manney, G. L., Thomas McElroy, C., Murtagh, D., Randall, C. E., Raspollini, P., Rozanov, A., Russell, J. M., Suzuki, M., Shiotani, M., Urban, J., von Clarmann, T., and Zawodny, J. M.: Validation of ACE-FTS version 3.5 NO*y* species profiles using correlative satellite measurements. *Atmospheric Measurement Techniques*, *9*(12), 5781–5810. https://doi.org/10.5194/amt-9-5781-2016, 2016.

Shephard, M. W., Herman, R. L., Fisher, B. M., Cady-Pereira, K. E., Clough, S. A., Payne, V. H., Whiteman, D. N., Comer, J. P., Vömel, H., Miloshevich, L. M., Forno, R., Adam, M., Osterman, G. B., Eldering, A., Worden, J. R., Brown, L. R., Worden, H. M., Kulawik, S. S., Rider, D. M., Goldman, A., Beer, R., Bowman, K. W., Rodgers, C. D., Luo, M., Rinsland, C. P., Lampel, M., and Gunson, M. R.: Comparison of Tropospheric emission spectrometer nadir water vapor retrievals with in situ measurements. Journal of Geophysical Research Atmospheres, *113*(15), 1–17, https://doi.org/10.1029/2007JD008822, 2008.

Vömel, H., Barnes, J. E., Forno, R. N., Fujiwara, M., Hasebe, F., Iwasaki, S., Kivi, R., Komal, N., Kyrö, E., Leblanc, T., B. Morel, B., Ogino, S.-Y., Read, W. G., Ryan, S. C., Saraspriya, S., Selkirk, H., Shiotani, M., J. Valverde Canossa, and D. N. Whiteman: Validation of Aura Microwave Limb Sounder water vapor by balloon-borne Cryogenic Frost point Hygrometer measurements, J. Geophys. Res., 112, D24S37, doi:10.1029/2007JD008698, 2007.

Weaver, D., Strong, K., Schneider, M., Rowe, P. M., Sioris, C. E., Walker, K. A., Mariani, Z., Uttal, T., Thomas McElroy, C., Vömel, H., Spassiani, A., & Drummond, J. R.: Intercomparison of atmospheric water vapour measurements at a Canadian High Arctic site. Atmospheric Measurement Techniques, *10*(8), 2851–2880. https://doi.org/10.5194/amt-10- 2851-2017, 2017.