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 #1

The authors would like to thank reviewer #1 for their attention to detail and helpful comments.

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

G1/ Possible erroneous values in tables 2 and 3.

As stated in the rapid access review (initial manuscript evaluation) there seems abnormally high amounts of water vapour in the stratosphere (>10ppmv). Quoting this earlier review:

Table 2: MIPAS: 12km: -0.3 ppmv = -1.4% implies a mean VMR of 21.4 ppmv

Table 3: MLS: 12km: -2.4 ppmv = -4.9% implies a mean VMR of 49.0 ppmv

Could the authors please check analysis and table entries and explain the high amounts of water vapour in the lower stratosphere.

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.

G2/ Defining the UTLS and limiting the scope of analysis to the UTLS.

The UTLS altitude range is not defined. Based upon analysis results presented the UTLS has a range of ~6-12 km. There are comparisons made down to ~1km, and up to 14km (fig 5, 6 & 9). Personally, I found that with the multiple datasets and comparisons spanning many altitude ranges it is hard to put together a coherent picture/story. There does seem consistency in comparisons over the 6-12(14) km range, as reflected in tables 2 and 3.

I suggest the scope of the study be limited to the UTLS only, and define the UTLS. If this approach is taken then the title be changed to reflect the scope. Maybe something like:

"Comparison of ground based-based and satellite measurements of upper troposphere and lower stratosphere water vapour profiles over Ellesmere Island, Nunavut."

A definition for the UTLS altitude range has been added, of between 5 and 22 km, in addition to a definition of the upper troposphere and lowermost stratosphere (UTLMS), i.e., altitudes from 5 km to ~15 km, since the reference instruments have sensitivity only below about this altitude range.

We prefer not to limit the altitude ranges shown as the tropospheric comparisons add to the larger story of what measurements are available in this data-poor region. They also put the UTLMS results in context. As noted in the conclusions, the results usefully motivate further work with the AIRS dataset.

G3/ Context

The introduction states the importance and reasons for accurate water vapour measurements in the UTLS. I think there could be more details on the importance of water vapour effects (and changes in water vapour) in the high arctic, hence the importance of the Eureka measurements.

There is a lack of information on past similar multi-measurement campaigns measuring UTLS-WV, such as MOHAVE-2009 (it is mentioned once in the conclusion). Is this current study the first such measurement comparison activity at high latitudes? I think this would help put this measurement comparison in context.

The first three sentences in the paragraph starting pg 2 line 17 are very weak. They do not add much information. Could such sentences be rewritten with either more information, or a good place to add context as mentioned in the paragraphs above.

Additional context has been added to motivate the study, including:

> "Atmospheric water vapour plays a crucial role in the chemistry, dynamics, and radiative balance of the Earth's atmosphere. Changes to water vapour abundances in the upper troposphere and lower stratosphere (UTLS), which approximately spans altitudes between 5 and 22 km, are particularly consequential for radiative balance (Soden et al., 2008; Riese et al., 2012). Water vapour abundances are expected to increase the most in the lowermost stratosphere (LMS) (Dessler et al, 2013), i.e., altitudes above the tropopause and beneath the tropical tropopause (~17 km), where the radiative impact of additional water vapour is maximum (Solomon et al., 2010). Despite the importance of understanding and monitoring changes to water vapour in this region, accurate long term measurements of water vapour in the upper troposphere and lowermost stratosphere (UTLMS) are limited.
>
> ….
>
> Satellite-based measurements complement ground-based observations by producing frequent global measurements of atmospheric constituents. More than a dozen satellites are currently (or have been recently) making measurements of water vapour. There is interest in assessing the accuracy and quality of these datasets. The Global Energy and Water Cycle Experiment (GEWEX) (Chahine, 1992) conducted a detailed assessment of tropospheric water vapour measurements. It identified many challenges to attaining a global understanding of the water cycle, including large inconsistencies in long-term total column water vapour measurements in deserts, mountainous regions, and the polar regions (Schröder et al., 2017). The conclusions of the GEWEX review of the state of water cycle measurements reiterated the need to improve on satellite profiling capabilities, diligent validation of data products, and to acquire stable, bias-corrected total column and profile datasets.
> In addition, a World Climate Research Programme (WCRP) Stratosphere-troposphere Processes And their Role in Climate (SPARC) activity…"

G4/ Inclusion of measurement uncertainties.

There is passing mention of measurement uncertainties per instrument (e.g. sondes 3-5%, FTIR ~= 10%) in the text, but this does not carry through in the analysis, figures and tables or in comparison commentary.

For instance in table 2: ACE-FTS: 12km: +0.4 ppmv = 9.7% implies a mean VMR of 4.1 pmv

What are the uncertainties at 12km associated with ACE-FTS and the FTIR measurements? If both were 50% then a 9.7% difference lies within the combined uncertainty. Such uncertainty analysis is not undertaken. Without it, it is hard to put the biases in context of instrument performance. I suggest adding some uncertainty analysis and associated commentary.

Minor, but related points:

-Inclusion of uncertainties estimates (over a given range, per instrument) in table 1 would be helpful.

The ACE-FTS dataset does not currently include full uncertainty estimates. The potential for a full uncertainty analysis is limited due to the differences in the information provided by each dataset. For example, ACE-FTS provides an error estimate that represents a statistical fitting error while MLS provides an estimate of the retrieval precision. Other validation work involving ACE datasets, e.g., Sheese et al. (2016), has not used uncertainties to assess the observed biases with other datasets for these reasons. To help inform the bias, the standard error in the mean has been reported, e.g., in Tables 2 and 3.

-In figures 5(c), 6(c) & 9(c) lines are drawn on the +/-10% relative difference. I suspect these have been included as a visual guide. I recommend using lines at 5% (or include lines at 5%) as this is the defined accuracy goal of the study (GCOS goal).

You are correct: the ±10% relative difference lines were added to aid the reader in interpreting the differences. ±5% lines would helpfully note the GCOS goal; however, when attempted, the scale of the figure made this visually too crowded, particularly Figure 9. Also, the 125HR water vapour profile retrieval's expected accuracy is 10%, making this line meaningful for those comparisons.

G5/ Layers, vertical resolution, sensitivity and degrees of freedom.

The GRUAN sonde measurements have high vertical resolution with multiple independent data points. For satellite base measurements there is piece-meal mention of vertical resolution (e.g. MIPAS ~3.3km, pg 10, line 17). There is no mention of the FTIR vertical resolution. Linked to vertical resolution, there is only passing mention of the degrees of freedom (DOFs) of the remotely sensed datasets. In the text it quotes TES DOFs to be 3 to 5 (pg 9, line 27), and FTIR retrieval sensitivity is mentioned in section 2.2. I recommend that table 1 be expanded to include columns stating the approximate/average vertical resolution and DOFs for each instrument over the UTLS region. If recommendation S15 (see below) is also implemented (on author discretion) this would also visually indicate vertical resolution to the reader.

Profile comparisons are analysed and reported on ~1km wide altitude layers (table 2 and 3, fig 6 & 9). Given the relatively coarse resolution of the remotely sensed datasets (along with datasets having less degrees of freedom that the number of levels reported on) there will be considerable inter-layer dependence and layer comparison results will be correlated. In figure 5 there seems to be ~28 levels from ~1km to 11km. Given that the Eureka FTIR DOFs are ~1.7 (Schneider, 2016) there is lack of layer independence.

Could authors please comment on inter-layer correlation and performing comparisons using remotely sensed products on vertical grids finer than their associated vertical resolution? Would it be better to perform partial column comparisons (2 or 3 for the UTLS)? This would reduce interlayer correlation.

The Schneider et al. (2016) paper's 1.7 DOFS refers to a dataset version that is different from the one used in this study. Theirs has a downgraded vertical resolution to align the FTIR H_2O product with the vertical resolution of the retrieved δD. The H_2O product at Eureka has an average DOFS of 2.9. Barthlott et al. (2017) has a useful table comparing the DOFS of these two versions of the MUSICA water vapour products. The reference at that point in the text has been changed to the Barthlott paper for greater clarity on this point.

Comparisons of partial columns would be much more limited due to the variability of altitude ranges available from many of the datasets, particularly those of primary interest here, i.e., ACE-FTS and ACE-MAESTRO. In addition, the altitude range where radiosonde measurements meet the uncertainty filtering applied in this study varies, often significantly from profile-to-profile, again limiting the ability for partial columns to be compared in the UTLS.

The vertical resolution of the sondes is better than 1 km, e.g. between 10 and 100 m. Each of the satellite datasets are retrieved or measured on a different grid. The comparisons with the radiosondes are reported on a 1-km grid so that a mean difference can be calculated (since this requires a regular grid), and also so that results from different satellite datasets can be compared with the others.

The different vertical resolution of the FTIR and comparison instruments is taken into account by smoothing the satellite profiles with the FTIR averaging kernels prior to the comparison.

G6/ Seasonal cycle and seasonality in the TPH

There is no mention of a seasonal dependence in dataset comparisons. All comparisons are made across entire datasets. Looking at Figs 7 and supplementary figures S1 & S3 there seems to be no seasonal bias in comparisons, whilst in Fig 10 (b) there could be a small seasonal bias but nothing mentioned in the manuscript. I think there needs to be a statement or section on seasonal biases (either stating there is a seasonal dependence or not).

There is also no mention on the seasonal variation in the TPH and how this would affect comparisons, especially since the TPH variation could span the current 1km resolution layers. A commentary on TPH height variation in comparisons is required (stating either an impact or lack of impact).

This is an interesting question. No seasonality was clearly seen in the differences. TPH dependence of the comparisons was plotted but no clear dependence was observed. The first paragraph of the discussion section now notes that "no seasonal pattern in the differences were observed, or pattern with respect to the TPH."

The figure below illustrates an example of the TPH vs. differences figures produced to check for impact on the results:

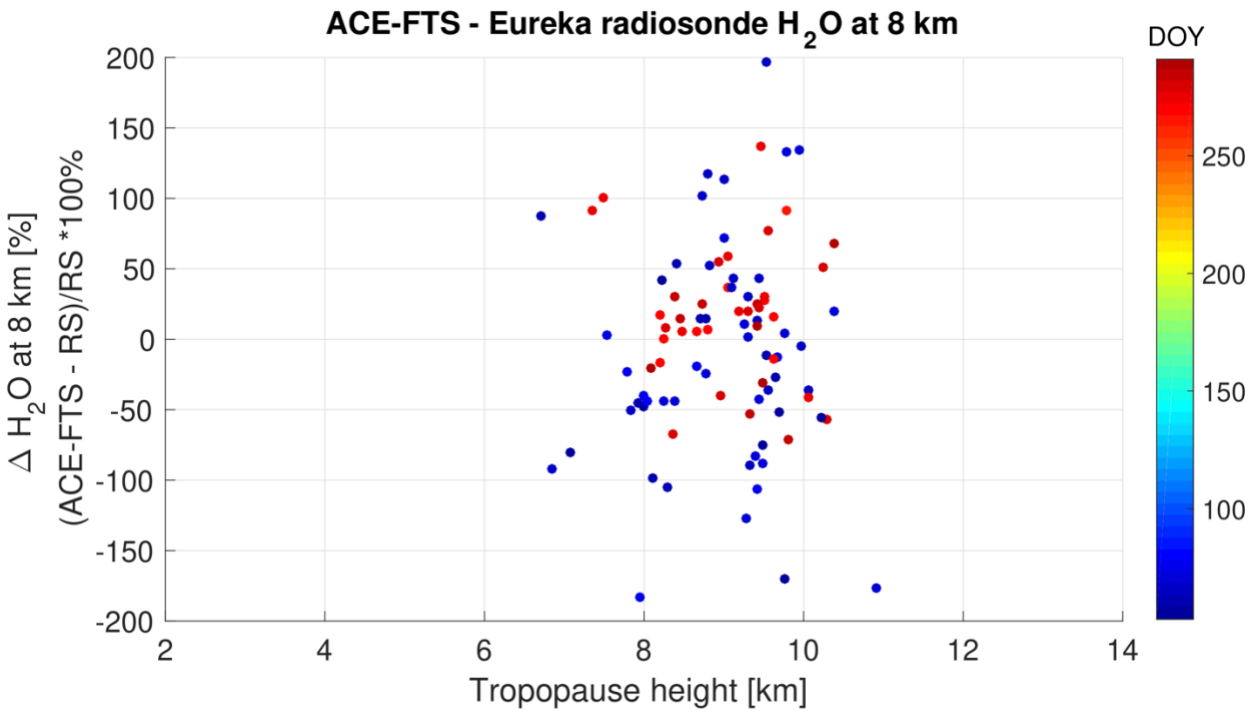

Figure 1: ACE-FTS – radiosonde differences at 8 km vs. tropopause height. Points are colour-coded by day of year (DOY). Tropopause height calculated by GRUAN radiosonde processing.

Specific comments:

S1/ References and referencing:

There is an instance where a paper is referenced in the manuscript, but not in the reference list (Khosrawi, 2018) and conversely there are papers in the reference list, Kurylo, 1991, Sioris, 2016b, & Stevens, 2013 that are not referenced in the manuscript. Can the authors please recheck the manuscript and reference list to make sure all cross-referencing is correct.

Thank you for catching the referencing mistakes. They have been corrected.

S2/ GCOS and WMO are used interchangeably.

GCOS was referenced in the main part of the manuscript, pg2, line 12, but then subsequent reference to the 5% accuracy goal is attributed to the WMO. Maybe for consistency keep GCOS, not WMO? ...or add a WMO reference.

Agreed. References to WMO has been replaced with GCOS.

S3/ Equation 7.
In eqn 7, 'GF' would be better represented as 'GF(z)'.

Thank you; this change has been made.

S4/ Convolving radiosonde VMR profiles with weighting functions: pg 13, line 26 and equation 8.

I think convolving is incorrect terminology, as mathematically it is not a convolution if the weighting function is not static (GF varies with altitude, see fig 4a) and not applicable for instrument averaging kernels. It is also unusual to smooth the high resolution data set (sonde) and report back on the high resolution levels. Usually the smoothed profile is reported on the coarse profile grid. Can the authors please comment on why the smoothed profile is reported back on the high resolution data set levels?

The description of the smoothing procedure in section 3.1 has been modified to state:

"the vertical resolution of radiosonde water vapour VMR profiles were downgraded using the weighting functions"

The radiosonde profiles have variable altitude levels, but measurements are reported roughly every 5 to 10 m in altitude. The satellite datasets all have different, courser, altitude grids. Some of the datasets have different altitude grids from profile-to-profile, e.g., ACE-MAESTRO. A regular grid is needed to put the results on a common basis for comparison. The 1 km grid is a reasonable middle-ground that also allows comparison between the radiosonde and 125HR

results, since many of the 125HR retrieval grid levels of interest are near those values, e.g., 6.4 km, 8.0 km, 9.8 km, and 12.0 km.

S5/ Equations 2 and 3.

Minor point: Usually 'X' is the independent variable (ordinate), and 'Y' is the dependent variable (abscissa). So maybe to hold convention it would be better to have X = reference measurement, Y = satellite measurement (pg 11, line 17). Currently Y = reference measurement.

Satellite – reference is an intuitive way to represent the observed agreement because:

If there is a high bias in the satellite measurement, the difference is positive.
If there is a low bias in the satellite measurement, the difference is negative.

This has been used in other validation literature, such as Vömel et al. (2007)'s MLS water vapour validation using cryogenic frostpoint hygrometer measurements.

S6/ Equation 1.
For completeness, the term $e_S(T)$ should be $e_S(T(z))$.

Thank you; this change has been made.

S7/ Sigma (σ) values in section 3.2.4

There are a series of statistics quoted in section 3.2.4 in which the units are ambiguous, for instance pg 16, line 5: -1.6 +/-1.5% (sigma = 45.9). What are the sigma units? (I gather ppmv?) Also again on line 7 and line 15.

In line 20, there is a statistic: -25.3 +/- 5.9% (sigma = 33.5%) is '%' the correct unit for the sigma value (the issue also reappears in line 23, and other instances)?

Can I recommend that consistency be preserved in the sense of report statistics in absolute units, i.e. ppmv then as relative (%) in brackets, or vice versa, but not to mix the order up at section level (or even keep consistent across the entire manuscript, if possible)

Yes, the units for all standard deviations are the same as the differences preceding them. The text has been updated to ensure that units for the standard deviation are stated explicitly in every instance.

The differences reported in section 3 have been updated to ensure that there is consistency in giving absolute units (ppmv) then relative differences (%) in brackets.

S8/ Quantifying small dry bias.
Could the 'small dry bias' (pg 17, line 21) be quantified in the text.

This has been revised to:

"As shown in Fig. 6, 361 TES measurements showed a dry bias relative to the 125HR of approximately 10% in the lower troposphere, a small dry bias (e.g., −1% at 3.0 km) to a small wet bias in the mid-troposphere (e.g., 3.7% at 3.6 km), and a wet bias (e.g. 20 – 25%) in the UTLS."

S9/ Hexagon symbols in Figs: 5, 8, 11, 12

A pedantic point, sorry, but I'm confused about the use of hexagons as symbols, are these to illustrate a point or an area? I'm assuming data binning, hence its representative of an area. In Cartesian X-Y plotting a hexagon is an interesting choice. Is the data binned within the hexagon region or usual X-Y (rectangular) binning and using the hexagon as a symbol centred in the middle of the rectangular bin?

The figures using the hexagons (Figs. 5, 8, 11, 12) show the density of the points within the area of the hexagonal symbols. This approach was taken because when plotting points for a correlation figure, the overlap between symbols at each point can mask useful information about how many points are in what location. The plots use hexagons rather than squares because this more closely approximates a circle, allowing the furthest points to be more symmetrically situated with respect to the center (e.g., compared to squares or triangles). The efficiency of this approach, including a comparison and discussion of the use of hexagons vs. other shapes, is described by Carr et al. (1987).

S10/ Tables 2 and 3

For completeness could SEM be explained in the table captions?

The definition of SEM has been added to the captions for Tables 2 and 3.

S11/ Figure 3 and accompanying discussion in the text: section 3.1.1

Figure 3 displays a decade of AIRS WV at 400hpa for 2 months: March and July. I'm struggling to find the significance of this figure. On pg 12, line 15 it states that figure 3 shows the spatial variation in water vapour abundance. The data is averaged over 10 years, hence mostly likely averaging out any spatial variation (due to high WV spatial variability). On pg 12, line 18, it states that WV variability is greater in summer (July) than winter (March? or should there be a December or January plot?). Fig 3, 'July', does show larger variability, but stratified in latitudinal bands, is this real? (given the discrete jumps and over 10 years of averaging, I suspect not). There is no commentary on these bands of WV.

At best figure 3 shows a coarse climatology over a large region. Is this what the authors want to convey? If WV seasonal spatial inhomogeneity (i.e. high spatial variance) is to be illustrated then maybe a different visualization should be considered.

Figure 3 was indeed included to illustrate the spatial (in)homogeneity of the water vapour abundances in the area around Eureka. March was used because ACE coincidences with Eureka

measurements occurred most often during March. The results for other winter months were not very different. Averaging over the available decade of measurements was intended to provide a general idea of the abundances in the region.

This figure has been replaced with a plot showing the Eureka-coincident AIRS measurements at 400 hPa in March and July for a specific representative year (2015) without any binning/averaging. (Plot included below.) There is some overlap between points, but this illustration better conveys the spatial variability of H_2O abundances in the area.

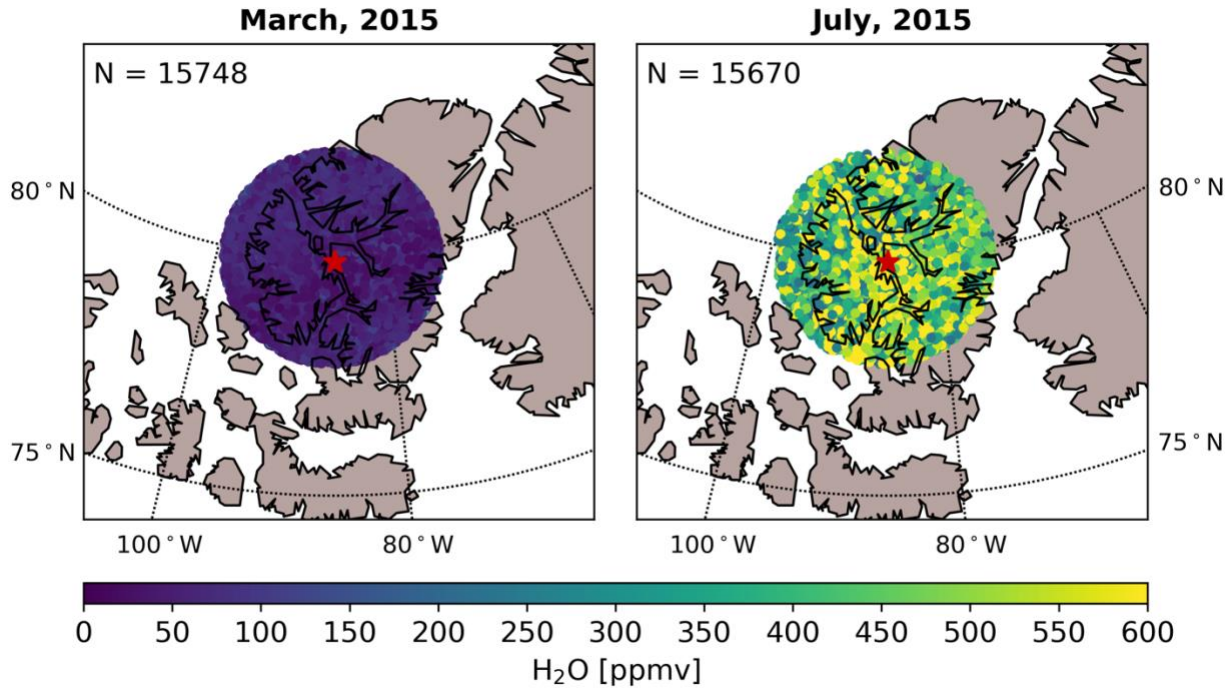

Figure 2: New "Figure 3" showing the spatial variability of H_2O AIRS measurements at 400 hPa near Eureka for two example months, March and July, in 2015.

S12/ Figure 5.

In figure 5 it seems sonde and FTIR data only goes up to 11km. Is this correct? The text states (pg 5, line 4) that sonde data is limited to less than 15km, but pg 5, line 27 states the mean sonde maximum altitude reached is 11.3km (+/- 4.4km). Also in Fig 6, sonde data is up to 14km. Why is sonde data limited to ~11km in Figure 5? I suspect the number of coincidences above 11km (fig 5, d) is too small.

Yes, there are no comparisons reported between the FTIR and sonde above 11 km. This is because only altitudes with $N \geq 15$ were shown throughout the study (noted in Section 3.1, which describes the method). Above 11 km, there were only a few coincidences found between those two instruments. This is largely due to the difference in measurement times; the FTIR takes measurements only during daylight (and operator hours emphasize times between 10 AM and 4 PM local time) while the sondes are launched at 6 AM and 6 PM local time (there are occasional exceptions for additional launches). This is illustrated in the figure below, a histogram

that shows the available MUSICA measurement times by the hour of the day. Daily radiosonde launch times are noted with red dashed lines. Atypical occasional radiosonde launch times are noted with blue dashed lines.

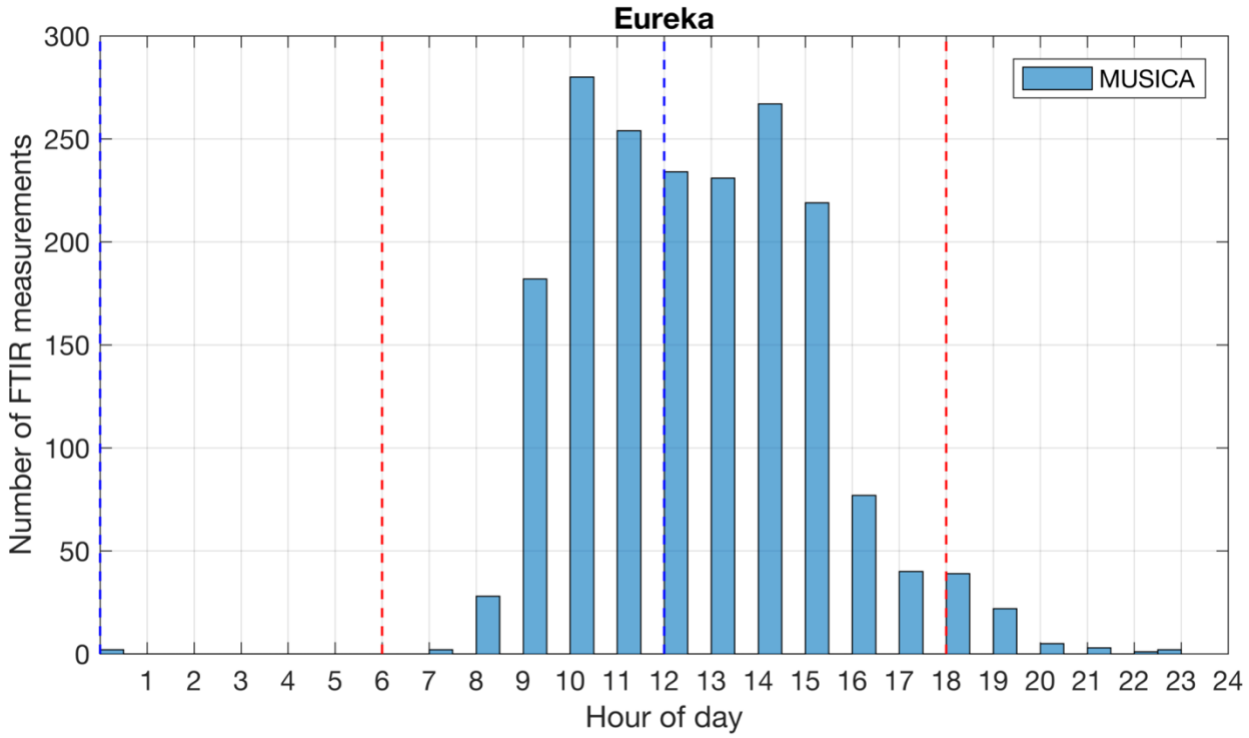

Figure 3: Histogram of Eureka MUSICA measurement times. Red dashed lines indicate typical daily radiosonde launch times (6 AM and 6 PM). Blue dashed lines indicate occasional atypical radiosonde launch times (12 AM and 12 PM).

In addition, the sonde measurements are filtered by uncertainty, which removes many of the measurements above 10 km. Text noting that mean profiles are not plotted for z > 11 km because N < 15 at those altitudes has been added to the Figure 5 caption.

S13/ Figure 5, part 2...

In the legend it states, X = sonde, Y =125HR. If this refers to use in eqns. 2 and 3 then the FTIR dataset is used as the 'reference' dataset. The sonde dataset would have higher accuracy in the UTLS. Should the sonde dataset should be used as the reference?

That is true. However, the difference in results would be only the sign of the statistics. This arrangement was chosen for consistency with other comparisons to the 125HR, as the radiosondes in this case are smoothed using the 125HR averaging kernels.

S14/ Table 1. Valid altitude range for SCISAT

The SCISAT valid altitude range table entry is vague, considering ACEF and ACEM are the primary satellite instrument datasets to be investigated. Could a more definite altitude range be specified?

The altitude range reported in Table 1 is worded in this manner because the valid altitude range of the ACE instruments is varies greatly from measurement-to-measurement (e.g., some ACE-FTS profiles extend only to 15 km at their lowest; in other cases, they extend to 5.5 km). In addition, determining the lowest altitude range where measurements are accurate is one of the objectives of the study.

S15/ Displaying instrument vertical resolution.

Figure 4 illustrates ACEF pseudo-vertical resolution (smoothing) and the 'smoothed' radio sonde profile. Figure 4 could be expanded to include the averaging kernels of other instrumentation. This would be helpful in illustrating the comparative vertical resolutions of the different datasets. Looking at figure 2, there seems a brief period in late 2008 that all datasets overlap (or very close to overlapping). A snapshot day of all datasets measurements and vertical resolutions could be displayed as an example. Such a figure could supplant the current figure 4, or be an additional supplementary figure (an idea the authors may wish to consider).

This is an interesting idea. An examination of all dataset coincidences resulted in one specific day where all datasets had measurements coincident with the 125HR and a few days where all datasets had measurements coincident with the radiosondes. The former and an example of the latter are shown below in Figures 6 and 7. They have been added as supplemental figures.

With regard to the vertical resolutions of the datasets, there will be an examination and presentation of this in a forthcoming WAVAS-II paper by Walker and Stiller. We aimed to avoid overlap with their efforts.

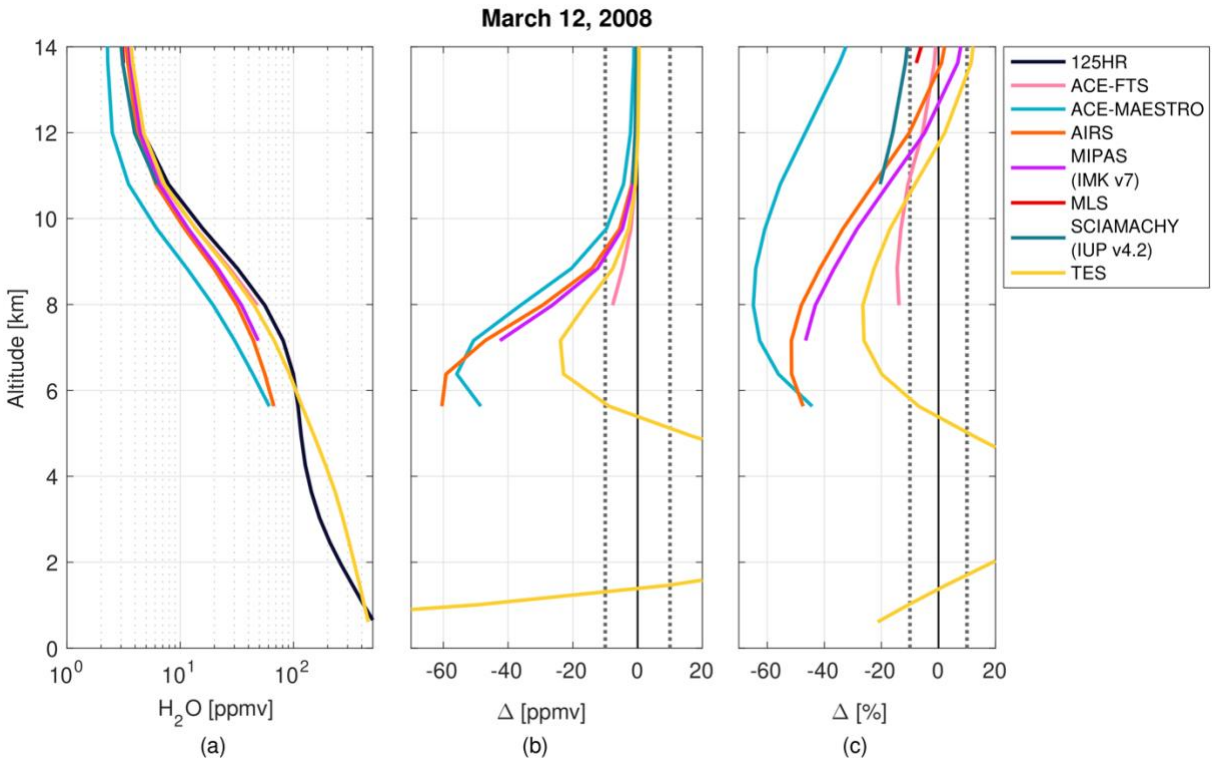

Figure 4: Individual satellite vs. 125HR profile comparisons on March 12, 2008.

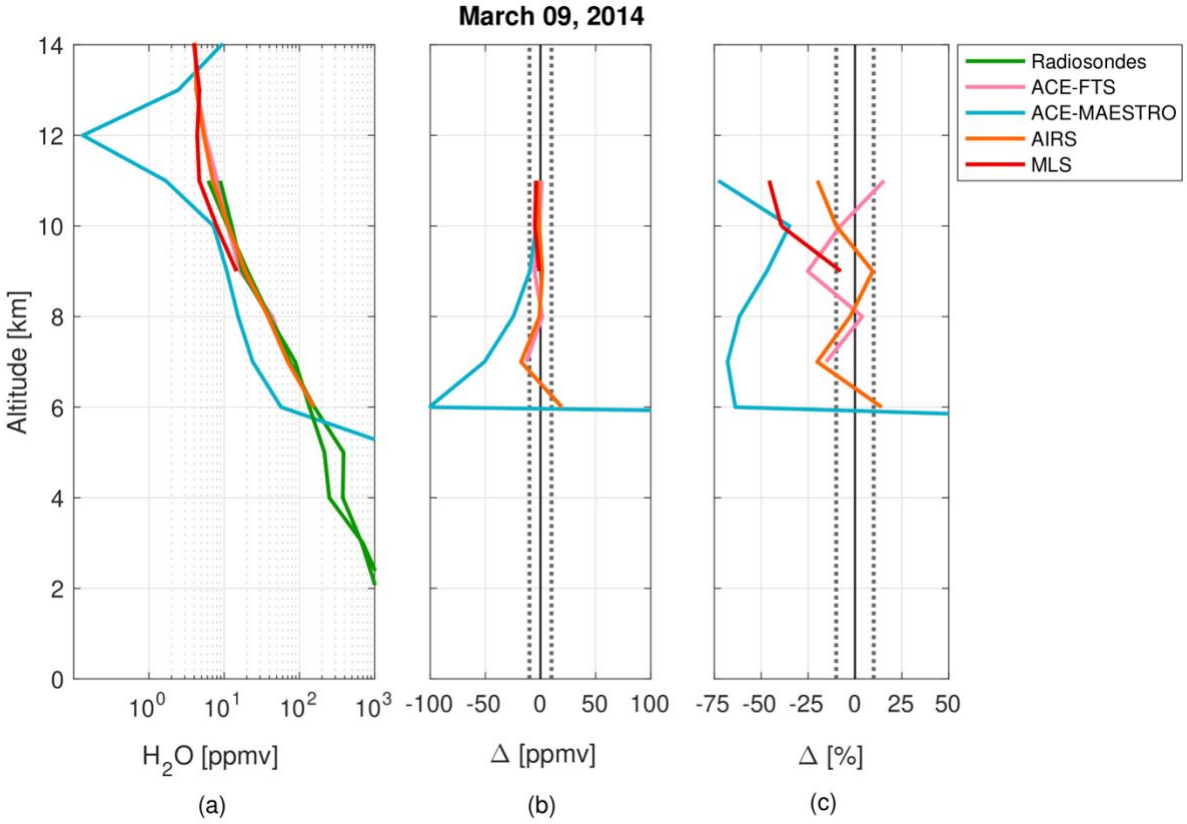

Figure 5: Individual satellite vs. radiosonde comparisons on March 09, 2014.

5

S16/ Two ground based reference datasets

In most studies there is a single defined reference dataset. In this manuscript there are two (FTIR and sonde). I recommend adding a short explanation as to why two reference datasets are used and the consequences of bias between two so called reference datasets (I gather the reason is to get more ground based to satellite coincidences). Given the high vertical resolution and accuracy of the GRUAN sonde dataset (in the UTLS region) should this be the primary (or single) reference dataset?

It is true that most studies use a single reference dataset. Two datasets were used in this study for a few reasons:

- Two datasets are available. The sondes and the FTIR are the only instruments routinely producing water vapour profiles from a standardized methodology at Eureka at the moment.
- The best available reference, the GRUAN-processed radiosondes, does not have ideal overlap with all the satellite measurements. In large part, this is due to the time of day they are launched and their twice-per-day frequency of measurements. In addition, the available raw data files needed for GRUAN processing have gaps and are available only from mid-2008 onwards. Consequently, some comparisons with the radiosondes are limited in time, space, or altitude-ranges. Too few coincidences were found between the radiosondes and MIPAS, SCIAMACHY, and TES for meaningful comparisons, for example.
- The GRUAN processing is not part of an ongoing arrangement, since Eureka is not an official GRUAN site. It is useful to see how well the FTIR comparison results align with the GRUAN results for ongoing monitoring of water vapour profiles produced by satellite instruments that have coincidences with Eureka.

Text has been added to the start of section 3 commenting on the use of two reference datasets.

> "Water vapour profiles from ACE-FTS, ACE-MAESTRO, AIRS, MIPAS, MLS, SCIAMACHY, and TES were compared with Eureka radiosonde and PEARL 125HR measurements following the methodology described below. Two ground-based reference measurements are used in this study to maximize comparisons with available satellite measurements. The radiosondes provide high vertical resolution profiles; however, they had few or no coincidences with MIPAS, SCIAMACHY, and TES. The 125HR, while having more limited vertical resolution, had coincident measurements with all satellite datasets used in this study."

S17/ Sonde measurements at TPH and above and the recommendation to instigate FPH measurements at Eureka.

In section 2.1 I find a bit of ambiguity. It states that RH% sonde measurements are only valid below the TPH, but then explains that the measurements up to 15km can be used. The sentence on pg 4, line 24 could be changed to state that 'historically' or 'usually' data has been limited to below the TPH, and also referenced as it is an important point.

The suggested change, inserting 'usually', has been made.

One of the conclusions of the study is that FPH measurements should be made at Eureka. For UTLS studies, if RH% sonde data is valid up to ~15km then what is gained from FPH measurements, this just needs to be explained a bit more (maybe greater accuracy than the RH% sonde, extended altitude range etc.)? The current sentence on pg 21, line 6 states "FPH measurements would offer the advantage of high accuracy as well as consistent coverage throughout the UTLS". Does this mean sonde data is not consistent? If so, why not?

Where we say that the radiosondes do not offer consistent coverage, that refers to the availability of the radiosonde profiles in the UTLS. Profiles are only sometimes used in that region due to the uncertainty filtering applied. The greater accuracy and lower uncertainty of the FPH measurements would be an advantage, as would their ability to capture information at higher altitudes in the lower stratosphere. The wording has been changed to more clearly articulate that it is the altitude range, in addition to the better accuracy, that would be an advantage of the FPH:

> "FPH water vapour measurements at Eureka would enhance the ongoing satellite validation work there and enable a valuable reference for PEARL water vapour measurements. FPH measurements would offer improved accuracy as well better coverage throughout UTLS altitudes relative to the radiosondes and 125HR. FPH measurements have been used for the validation of other missions such as MLS (Hurst et al. 2016) and MIPAS (Stiller et al., 2012, using the MOHAVE measurements). Adding FPH measurements would be a useful next step for the comparison and validation of water vapour profiles at Eureka."

References

Carr, D. B., Littlefield, R. J., Nicholson, W. L., and Littlefield, J. S.: Scatterplot Matrix Techniques for large N, Journal of the American Statistical Association, 82 (389), pp. 424-436, doi:10.1080/01621459.1987.10478445, 1987.

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.