

Author response to reviewer's comments on

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

5

by Weaver et al.

Reply to Reviewer #1

10 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.

15 *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

20 *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.

25

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

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

30 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.:

35

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

5 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.

10 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.

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

$$\Delta_{\text{abs}}(12 \text{ km}): -0.3 \text{ ppmv}$$

$$\Delta_{\text{rel}}(12 \text{ km}): -1.4\%$$

20 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, i.e., to calculate the percent differences using the difference between the mean profiles, we get:

25
$$\Delta_{\text{abs}}(12 \text{ km}): -0.3 \text{ ppmv}$$

$$\Delta_{\text{mean}}(12 \text{ km}): -4.4\%$$

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

30
$$\text{H}_2\text{O} = \Delta_{\text{abs}} / \Delta_{\text{mean}} = 0.3 \text{ ppmv} / 4.4\% = 6.8 \text{ ppmv}$$

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

35 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:

40
$$X = (1, 3, 5)$$

$$Y = (2, 2, 8)$$

The mean of X is: 3

45 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).

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

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

10 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.:

15 $-1 / -25\% = 4,$

i.e., the original mean of Y.

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

20 *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.*

25 *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.”

30 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.

35 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.

- 5 *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.*

10 *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:

15 “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
20 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.

25

30 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
35 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.

40 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 \sim 10%) in the text, but this does not carry through in the analysis, figures and tables or in comparison commentary.

5 *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.

10 *Minor, but related points:*

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

15 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.

20 *-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).*

25 You are correct: the $\pm 10\%$ relative difference lines were added to aid the reader in interpreting the differences. $\pm 5\%$ 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.

5 *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*
10 *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.*

15 *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.*

20 *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.*

25 *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 H₂O product with the vertical resolution of the retrieved δ D. The H₂O 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.*

30 *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.*

35 *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.*

40 *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.*

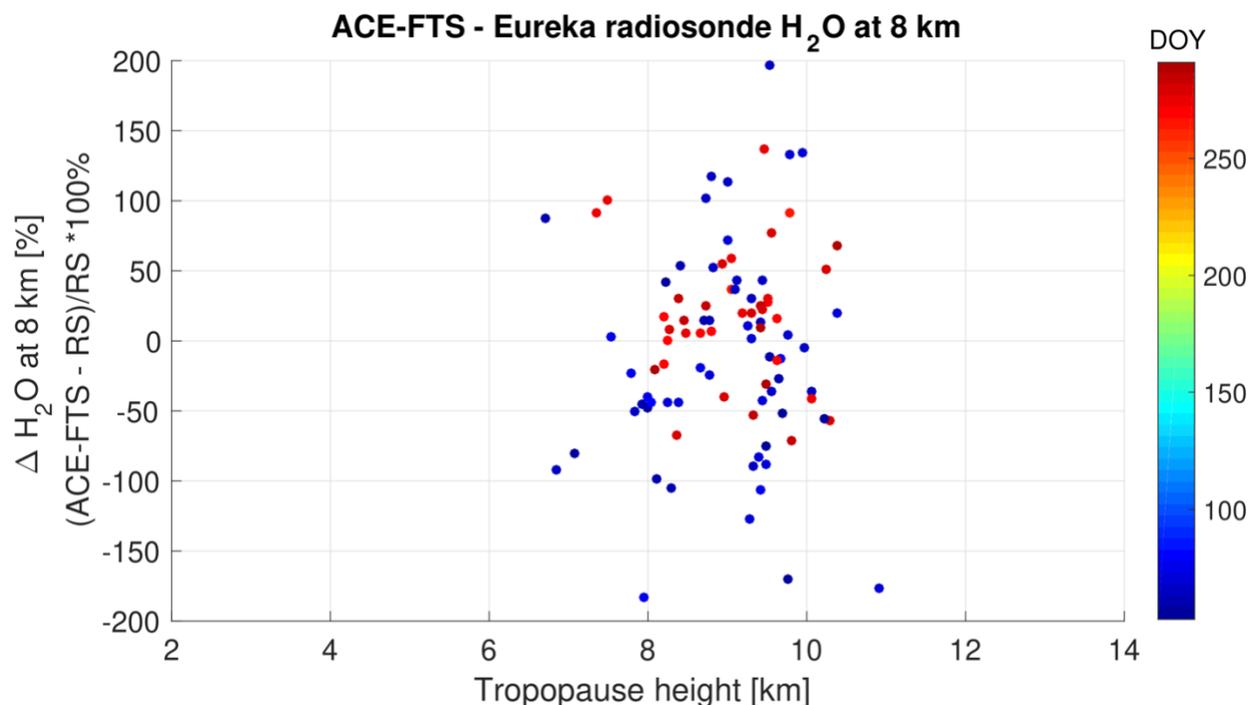
G6/ Seasonal cycle and seasonality in the TPH

5 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).

10 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).

15 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:



20 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:

5 *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.

10 *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.

15 *In eqn 7, 'GF' would be better represented as 'GF(z)'.*

Thank you; this change has been made.

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

20 *I think convolver 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?*

25 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”

30 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.

- 5 *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:

- 10 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.

- 15 *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

- 20 *There are a series of statistics quoted in section 3.2.4 in which the units are ambiguous, for instance pg 16, line 5: $-1.6 \pm 1.5\%$ (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: $-25.3 \pm 5.9\%$ (sigma = 33.5%) is '%' the correct unit for the sigma value (the issue also reappears in line 23, and other instances)?

- 25 *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.

- 30 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:

5 “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

10 *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?*

15 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).

20 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

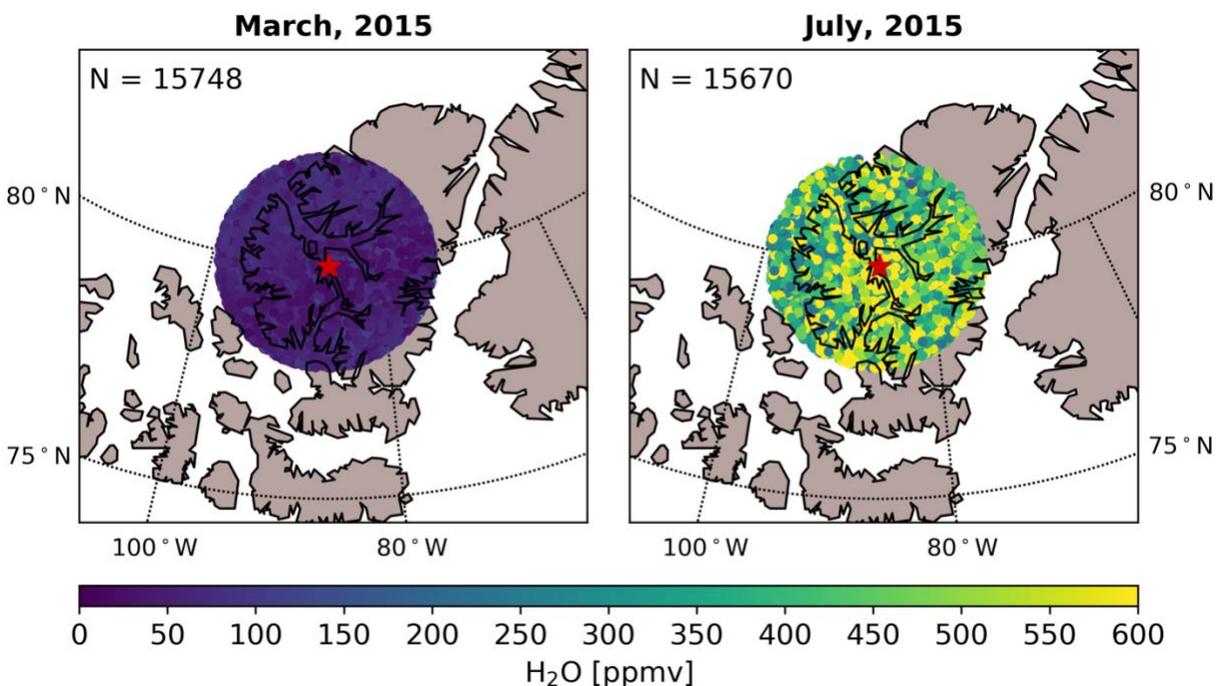
25 *Figure 3 displays a decade of AIRS WV at 400hpa for 2 months: March and July. I'm struggling to find the significance of this figure. On pg 12, line 15 it states that figure 3 shows the spatial variation in water vapour abundance. The data is averaged over 10 years, hence mostly likely averaging 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.*

30 *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.*

35 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.

- 5 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 H₂O abundances in the area.



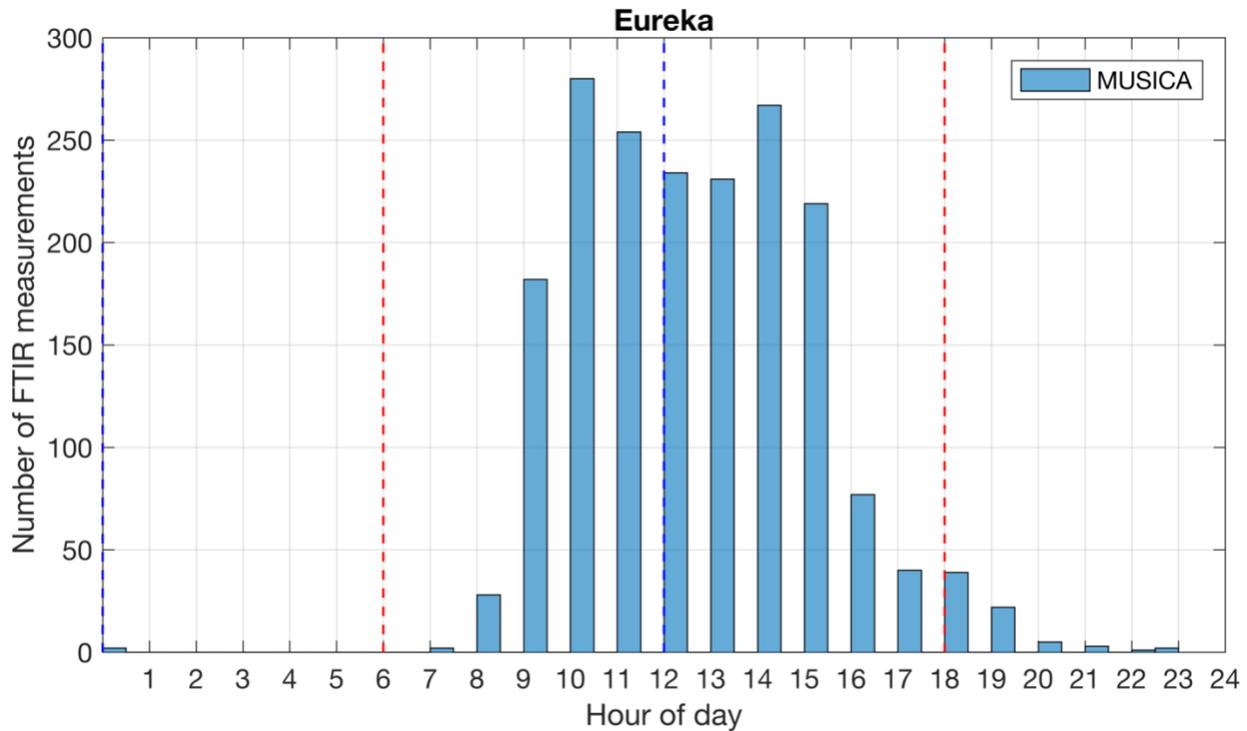
10 Figure 2: New "Figure 3" showing the spatial variability of H₂O AIRS measurements at 400 hPa near Eureka for two example months, March and July, in 2015.

S12/ Figure 5.

- 15 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.

- 20 Yes, there are no comparisons reported between the FTIR and sonde above 11 km. This is because only altitudes with N ≥ 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).
- 25 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.



5

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.

10

S13/ Figure 5, part 2...

15

In the legend it states, $X = \text{sonde}$, $Y = 125\text{HR}$. 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.

20

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?

- 5 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.

10 *S15/ Displaying instrument vertical resolution.*

- 15 *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).*

- 20 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.

- 25 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.

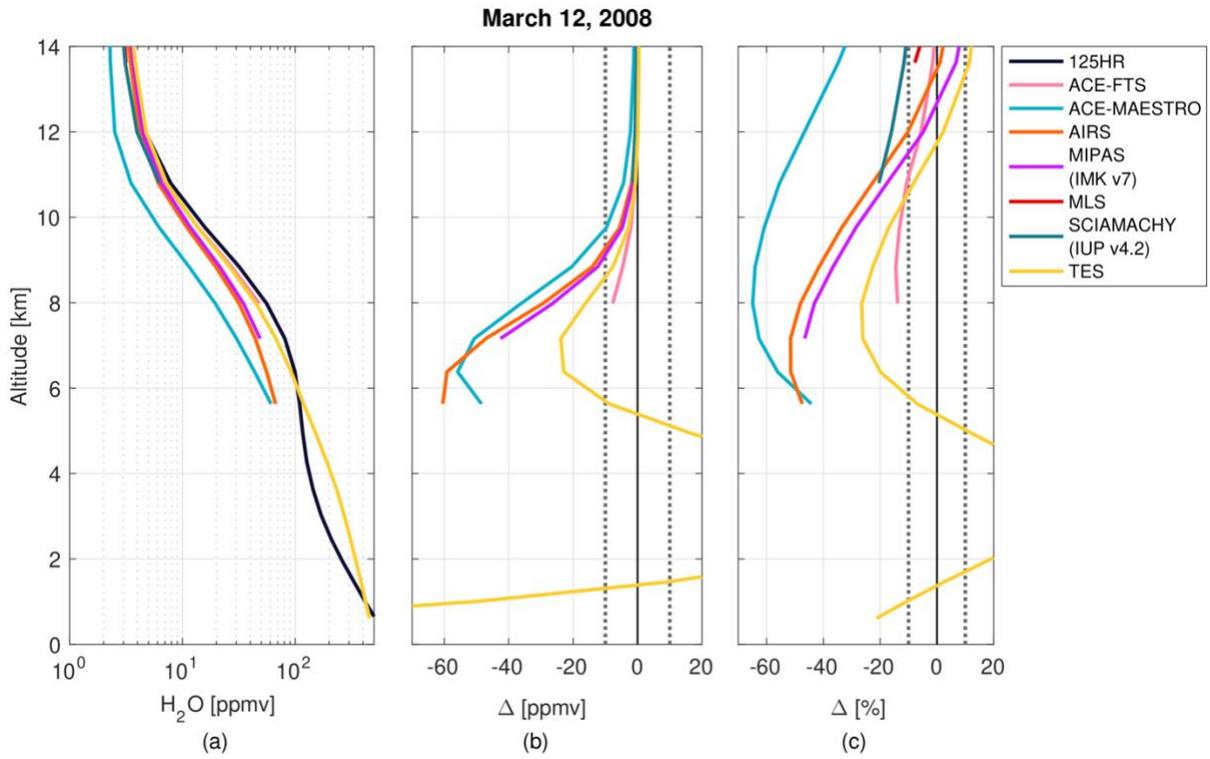


Figure 4: Individual satellite vs. 125HR profile comparisons on March 12, 2008.

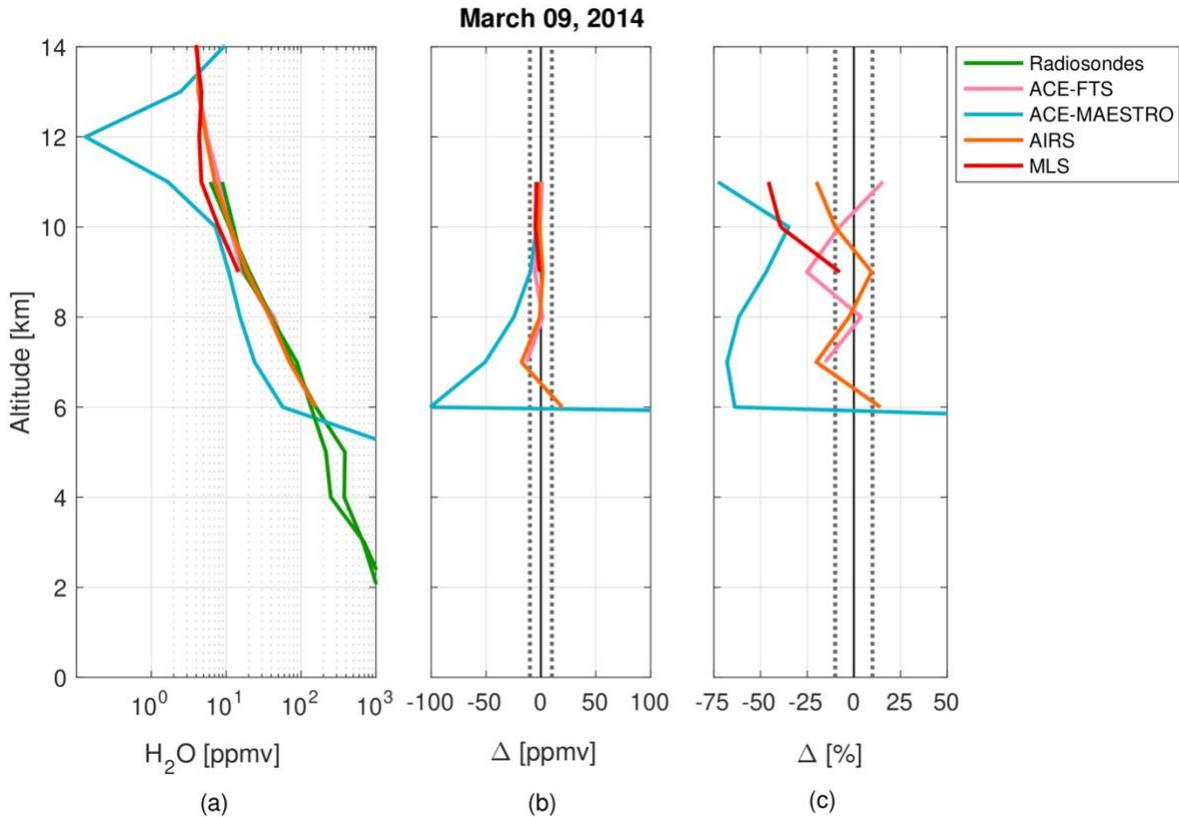


Figure 5: Individual satellite vs. radiosonde comparisons on March 09, 2014.

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.

Author response to reviewer's comments on

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

5

by Weaver et al.

Reply to Reviewer #2

10 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.

15 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 I25HR 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.

35

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 (v_5 and v_7) at 12 km in Table 2. The mean difference from the I25HR 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.

45 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:

5

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

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

$$10 \quad \Delta_{rel}(z) = 100\% \times \frac{1}{N(z)} \sum_{i=1}^{N(z)} \frac{X_i(z) - Y_i(z)}{Y_i(z)}, \quad (2)$$

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

$$15 \quad \Delta_{mean}(z) = 100\% \times \frac{\frac{1}{N(z)} \sum_{i=1}^{N(z)} X_i(z) - \frac{1}{N(z)} \sum_{i=1}^{N(z)} Y_i(z)}{\frac{1}{N(z)} \sum_{i=1}^{N(z)} Y_i(z)} = 100\% \times \frac{\sum_{i=1}^{N(z)} [X_i(z) - Y_i(z)]}{\sum_{i=1}^{N(z)} Y_i(z)}. \quad (3)$$

15

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.

20

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

25

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:

30

$$\begin{aligned} \Delta_{abs}(12 \text{ km}): & -0.3 \text{ ppmv} \\ \Delta_{rel}(12 \text{ km}): & -1.4\% \end{aligned}$$

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

35

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

40

$$\begin{aligned} \Delta_{abs}(12 \text{ km}): & -0.3 \text{ ppmv} \\ \Delta_{mean}(12 \text{ km}): & -4.4\% \end{aligned}$$

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

$$\text{H}_2\text{O} = \Delta_{\text{abs}} / \Delta_{\text{mean}} = 0.3 \text{ ppmv} / 4.4\% = 6.8 \text{ ppmv}$$

5 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.

10 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:

$$X = (1, 3, 5)$$

$$Y = (2, 2, 8)$$

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

The difference between the two means is: -1

20 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:

$$25 \frac{X - Y}{Y} * 100\% = \left(-\frac{1}{2}, \frac{1}{2}, -\frac{3}{8}\right) * 100\% = (-50\%, 50\%, -37.5\%)$$

$$\text{Mean percent difference } (\Delta_{\text{rel}}) = -12.5\%$$

30 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.

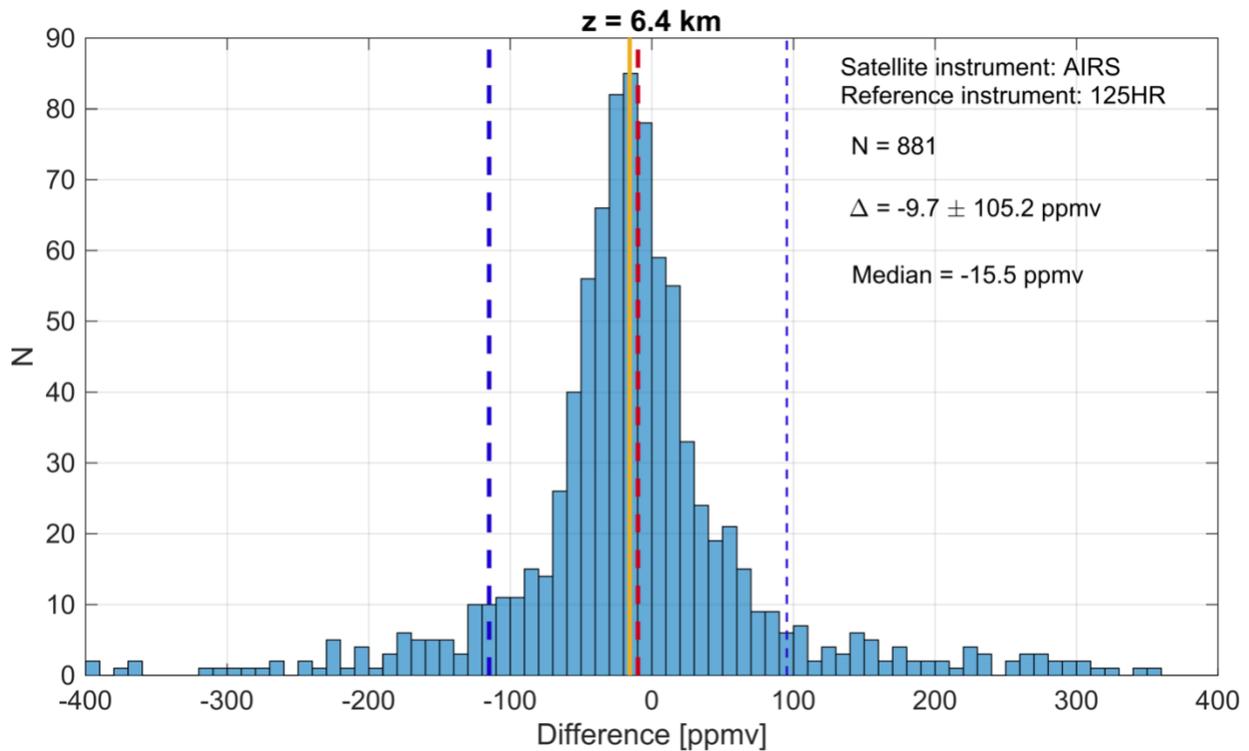
35

.... 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.

5

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.

10



15 *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.*

20

25

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

- 5 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.
- 10 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
- 15 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).
- 20 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.

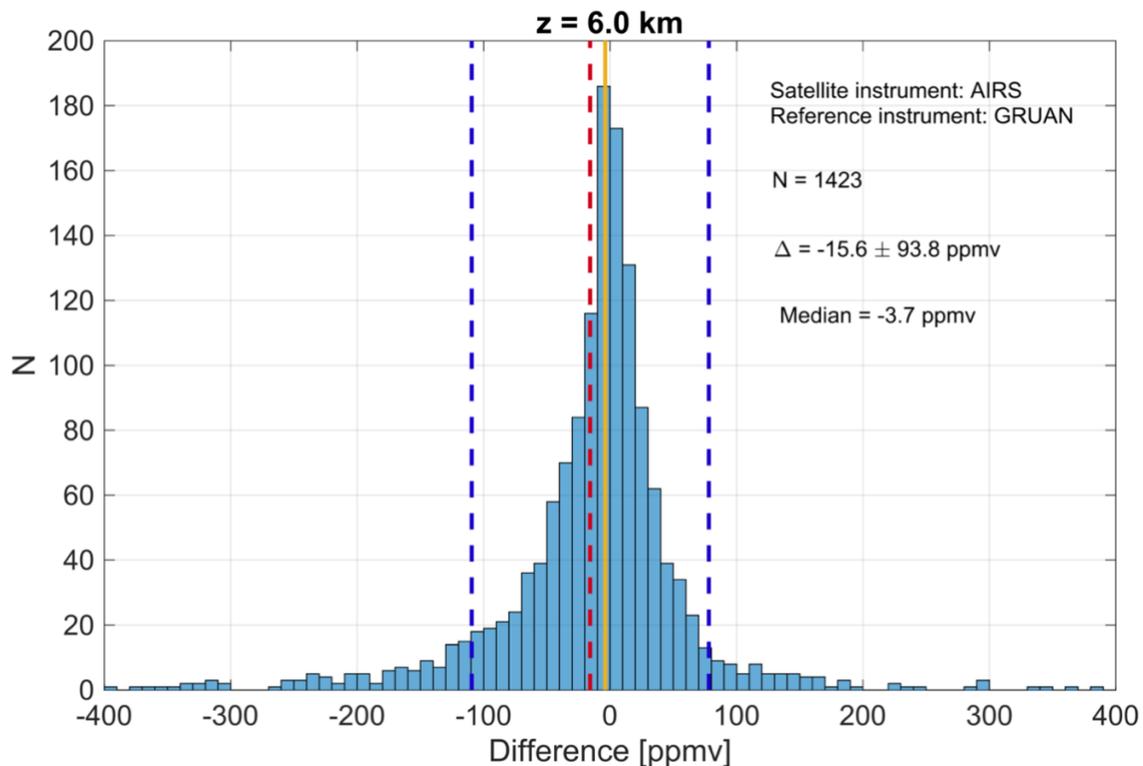
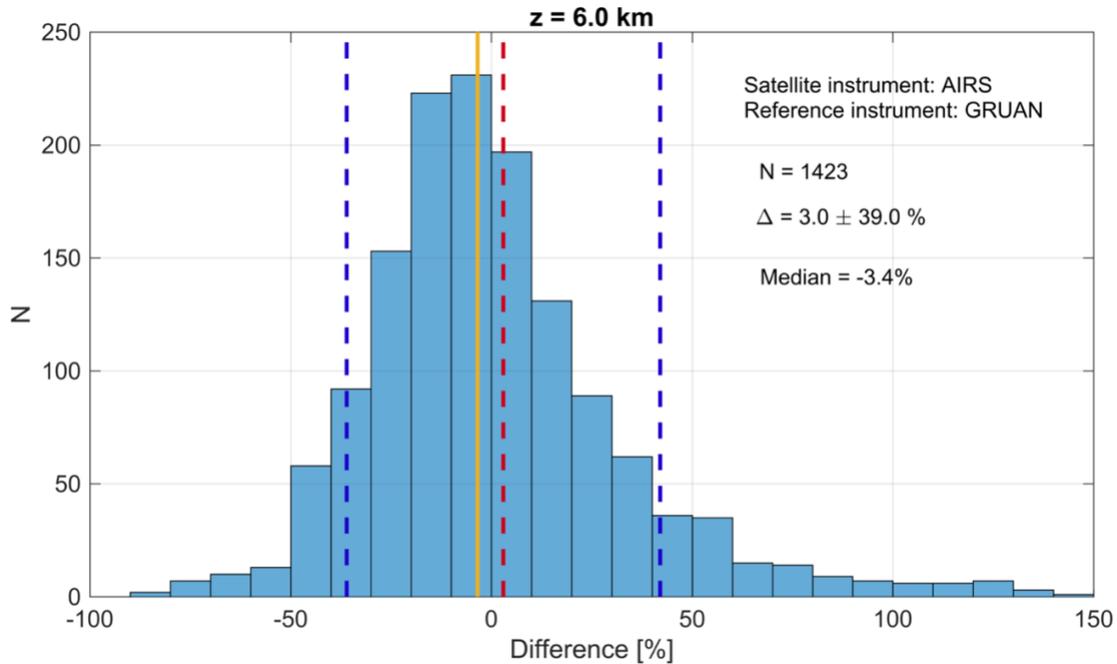


Figure 2: Histogram of absolute differences between AIRS and GRUAN-processed radiosonde water vapour measurements at 6 km. Lines defined as in Figure 1.



5 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.

10 *Correlation coefficients and correlation plots are of limited quantitative value in a paper focused on measurement biases between pairs of instruments. Two sets of measurements can be well correlated even though there are huge biases between them! Correlation 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*

15 *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; Ohya 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).

5 *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*
10 *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.*

15 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
20 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
25 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
30 AIRS comparison results, the paper notes that a much tighter coincidence criteria of 25 km and 2 hours shows similar comparison results.

General Comments:

35 *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.*

40 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.

5 *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.*

10 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.
15 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?

20 "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.

25 This sentence has been reworded to include the upper troposphere.

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

30 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.

35 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?

5 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.

10
15 *P6 L20 this would be a good place to mention the vertical resolution of the MUSICA FTIR WV profiles*

This has been added:

20 “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.

30 *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.

35 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.

40
45 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:

5 “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).”

10 “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).”

15 “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).”

20 *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.

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

30 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.

35 *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.

40 *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?*

45 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.

5 *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?*

10 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.

15 *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:

20 "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?

25 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.

30 *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? If 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.*

35 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%.

40 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?

5

-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%".

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?

15 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.

20 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."

25 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?

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

40 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, respectively. 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.

45 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 I25HR is apparent in Figure 7, so the "scatter" is not "around the zero line" as stated, otherwise there would be no bias.

5 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.

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

References

- 15 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 NO₂ measurements using
20 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.
25 *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.
30 <https://doi.org/10.5194/amt-6-263-2013>, 2013.
- Ohyama, H., Kawakami, S., Shiomi, K., Morino, I., & Uchino, O.: Intercomparison of XH₂O data from the GOSAT TANSO-FTS (TIR and SWIR) and ground-based FTS measurements: Impact of the spatial variability of XH₂O on the intercomparison. *Remote Sensing*, 9(1). <https://doi.org/10.3390/rs9010064>, 2017.
- 35 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.

- 5 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.
- 10 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.
- 15 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.
- 20 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.

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

Dan Weaver¹, Kimberly Strong¹, Kaley A. Walker¹, Chris Sioris², Matthias Schneider³, C. Thomas McElroy⁴, Holger Vömel⁵, Michael Sommer⁶, Katja Weigel⁷, Alexei Rozanov⁷, John P. Burrows⁷, William G. Read⁸, Evan Fishbein⁸, and Gabriele Stiller³

5

¹Department of Physics, University of Toronto, Toronto, Ontario, Canada

²Environment and Climate Change Canada, Toronto, Ontario, Canada

³Institute of Meteorology and Climate Research (IMK-ASF), Karlsruhe Institute of Technology, Karlsruhe, Germany

⁴Department of Earth and Space Science and Engineering, York University, Toronto, Canada

10 ⁵Earth Observing Laboratory, NCAR, Boulder, Colorado, USA

⁶GRUAN Lead Centre, Deutscher Wetterdienst, Lindenberg, Germany

⁷Institute of Environmental Physics, University of Bremen, Bremen, Germany.

⁸Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA

15 *Correspondence to:* Dan Weaver (dweaver@atmosph.physics.utoronto.ca)

Abstract. Improving measurements of water vapour in the lower stratosphere and upper troposphere (UTLS) is a priority for the atmospheric science community. In this work, UTLS water vapour profiles derived from Atmospheric Chemistry Experiment (ACE) satellite measurements are assessed with coincident ground-based measurements taken at a high Arctic
20 observatory at Eureka, Nunavut, Canada. Additional comparisons to satellite measurements taken by AIRS, MIPAS, MLS, SCIAMACHY, and TES are included to put the ACE-FTS and ACE-MAESTRO results in context.

Measurements of water vapour profiles at Eureka are made using a Bruker 125HR solar absorption Fourier transform infrared spectrometer at the Polar Environment Atmospheric Research Laboratory (PEARL) and radiosondes launched from the Eureka
25 Weather Station. Radiosonde measurements used in this study have been processed with software developed by the Global Climate Observing System (GCOS) Reference Upper Air Network (GRUAN) to account for known biases and calculate uncertainties in a well-documented and consistent manner.

ACE-FTS measurements were within 11 ppmv (13%) of 125HR measurements between 6 and 14 km. Between 8 and 14 km ACE-FTS profiles showed a small wet bias of approximately 8% relative to the 125HR. ACE-FTS water vapour profiles had mean differences of 13 ppmv (32%) or better when compared to coincident radiosonde profiles at altitudes between 6 and
30 14 km; mean differences were within 6 ppmv (12%) between 7 and 11 km. ACE-MAESTRO profiles showed a small dry bias relative to the 125HR of approximately 7% between 6 and 9 km and 10% between 10 and 14 km. ACE-MAESTRO profiles agreed within 30 ppmv (36%) of the radiosondes between 7 and 14 km. ACE-FTS and ACE-MAESTRO comparison results show closer agreement with the radiosondes and PEARL 125HR overall than other satellite datasets - except AIRS. Close

agreement was observed between AIRS and the 125HR and radiosonde measurements, with mean differences within 5% and correlation coefficients above 0.83 in the troposphere between 1 and 7 km.

Comparisons to MLS at altitudes around 10 km showed a dry bias, e.g., mean differences between MLS and radiosondes were -25.6%. SCIAMACHY comparisons were very limited due to minimal overlap between the vertical extent of the measurements. TES had no temporal overlap with the radiosonde dataset used in this study. Comparisons between TES and the 125HR showed a wet bias of approximately 25% in the UTLS and mean differences within 14% below 5 km.

1. Introduction

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). Increases in stratospheric water vapour abundances are expected to be largest 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.

Ground-based observations of water vapour are made using a variety of instruments. Many instruments only acquire total column measurements, e.g., Sun photometers. Others acquire profiles as well, such as Fourier transform infrared (FTIR) spectrometers. However, ground-based FTIR observations are limited by the relatively sparse network of sites globally, and current FTIR water vapour profile retrievals have a modest vertical resolution (e.g., Barthlott et al., 2017). Balloon-based radiosonde sensors measure atmospheric humidity profiles with high vertical resolution, typically better than 100 m, and are launched daily from approximately 1000 sites globally (Durre et al., 2006). This geographic coverage nonetheless has many gaps, e.g. in the polar and oceanic regions, and radiosonde launches are typically limited to once or twice a day. While limited in their global coverage, ground-based instruments produce well-characterized measurements that can be used to study specific sites, compare with models, and validate satellite measurements.

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.

Deleted: For example, In the upper troposphere and lower stratosphere (UTLS), water vapour has significant impacts on radiative forcing (Solomon et al. 2010; Soden et al., 2008). Changes in UTLS water vapour abundances have important implications for climate (Dessler and Sherwood, 2009; Held and Soden, 2000).

T are particularly

Formatted: English (UK)

Deleted: affected by changes to water vapour

Deleted: ; Dessler et al., 2013

Deleted: For example, , indicating that stratospheric water vapour abundances affect temperatures at the surface

Deleted: C

Deleted: Modelling results show that the

Deleted: largest cin the stratosphere are expected to be i

Deleted: , due to the stratospheric water vapour feedback al13

Deleted: Indeed, t

Moved down [1]: The Global Climate Observing System (GCOS) considers acquiring measurements of water vapour profiles to an accuracy of 5% essential for understanding the climate system (GCOS, 2016). However, global measurements of UTLS water vapour are not yet acquired routinely at the accuracy sought by the atmospheric science community. Instruments and measurement techniques are being developed to fill this observational need.

Deleted: Since there is a small amount of humidity in the Arctic atmosphere, spectral windows in the infrared region of the spectrum that are usually saturated, i.e. no radiation can pass through the atmosphere at those frequencies without being absorbed, are not fully saturated in the Arctic. Consequently, increases in atmospheric water vapour can have a particularly acute effect on radiative balance of the Arctic atmosphere, since infrared spectral windows that are saturated elsewhere are not fully saturated in the Arctic due to the smaller amount of humidity (Tobin et al., 1999; Turner et al., 2010).

Deleted: T Indeed, temperature records indicate the changes in Arctic surface air temperature are twice as large as the global average (Wendisch et al., 2017). This larger increase in temperatures at Arctic latitudes is known as Arctic amplification, an aspect of the climate system that is driven by multiple factors, including the decline in sea ice, changes to cloud cover, and increases in water vapour abundances (Serreze et al., 2011).

Deleted: (e.g., Schneider et al., 2016).

Deleted: and Vane

Deleted: recently published

Deleted: 6

Formatted: English (US)

In addition, a World Climate Research Programme (WCRP) Stratosphere-troposphere Processes And their Role in Climate (SPARC) activity is currently conducting a comprehensive overview of water vapour satellite measurements between the upper troposphere and lower mesosphere. This effort, the second SPARC water vapour assessment (WAVAS-II), intercompares the available satellite measurements to understand the differences between available datasets, measurement uncertainties, and trends in stratospheric and lower mesospheric water vapour. Results from the WAVAS-II effort are being published in a special inter-journal issue of AMT/ACP/ESSD, e.g., Khosrawi et al. (2018), and is available at: https://www.atmos-chem-phys.net/special_issue830.html.

Deleted: For example

Deleted: stratospheric and lower mesospheric

Developing highly accurate and vertically-resolved UTLS water vapour profile measurements from satellite instruments is a priority of the atmospheric observing community (Müller et al., 2015). However, obtaining sensitivity to the troposphere and producing high vertical resolution profiles is challenging for many satellite instruments. The Global Climate Observing System (GCOS) considers acquiring measurements of water vapour profiles to an accuracy of 5% essential for understanding the climate system (GCOS, 2016). However, global measurements of UTLS water vapour are not yet acquired routinely at the accuracy sought by the atmospheric science community. Instruments and measurement techniques are being developed to fill this observational need. Comparisons to ground-based observations offer an opportunity to assess the accuracy of satellite measurements.

Moved (insertion) [1]

The objective of this study is to assess the Arctic water vapour profiles retrieved from Atmospheric Chemistry Experiment (ACE) satellite observations using comparisons to coincident measurements taken at a Canadian high Arctic observatory in Eureka, Nunavut. In addition, other satellite instruments with Eureka-coincident water vapour profile measurements are compared to put the ACE results in the context of the broader effort to measure water vapour from satellites. This study adds to earlier work that has compared ground-based FTIR measurements to ACE v3.5/3.6 (e.g., Griffin et al., 2017) and studies comparing ACE measurements to those of other satellites (e.g., Sheese et al., 2017). Due to the vertical sensitivity of the available Eureka reference measurements, and the importance of this region for understanding factors influencing the atmosphere's radiative balance, the focus of this work will be on altitudes of the UTLMS, i.e., altitudes between 5 and 15 km. This study is structured as follows. Section 1 introduces the motivation for UTLS water vapour measurements and describes the ground-based measurement site. Section 2 describes the instruments and datasets used in the study. Section 3 compares the satellite and ground-based measurements, noting the methods used to match observations and account for different vertical sensitivities. Section 4 discusses the results of the comparisons. Section 5 offers conclusions about the ability of the ACE and other satellite datasets to contribute to our knowledge of high Arctic water vapour and comments on the implications for future research.

Deleted:

1.1. Ground-based reference site

Eureka, Nunavut is a research site on Ellesmere Island in the Canadian high Arctic. It has an extremely cold and dry environment. Eureka is located at 10 metres above sea level on the shore of Slidre Fjord, 12 km east of Eureka Sound. Open water occurs regionally during summer, but during the rest of the year, the surface of the fjords and sounds are frozen. The

geography of the surrounding area is variable, including ridges, hills, and small mountains. Because of the site's 80° N latitude, there is no sunlight between mid-October and mid-February.

Environment and Climate Change Canada's Eureka Weather Station (EWS) is the primary presence in Eureka (79.98° N, 85.93° W). One of the key measurements taken at the EWS is the twice-daily radiosonde observations of temperature, pressure, wind, and humidity profiles. Radiosonde measurements at Eureka extend back to 1948. These measurements show, for example, that tropospheric temperatures are increasing, water vapour total columns are increasing, and temperature and humidity inversions often form in the lower troposphere above Eureka between fall and spring (Lesins et al., 2010). The EWS is also used as an operational hub for government and academic research conducted in the area. Since 2006, the Canadian Network for the Detection of Atmospheric Change (CANDAC) has operated a large suite of atmospheric monitoring instruments at the Polar Environment Atmospheric Research Laboratory (PEARL) near Eureka (Fogal et al., 2013). The Ridge Lab is the largest of the PEARL facilities, and is located at 80.05° N, 86.4° W on top of a ridge at 610 m elevation, 10 km west of Eureka. The large number of observations taken by the EWS and PEARL instruments offer extensive characterization of atmospheric conditions at the site.

Many polar-orbiting and limb-viewing satellites commonly have overpasses with Eureka. As a result, measurements taken at PEARL have contributed to many validation studies, e.g. of ACE (Griffin et al., 2017), MOPITT (Buchholz et al., 2017), OCO-2 (Wunch et al., 2017), and OSIRIS (Adams et al., 2012).

2. Instruments

This section presents the water vapour datasets from Eureka ground-based instruments and Eureka-coincident satellite instruments that are used in this study. Table 1 summarizes the available datasets, notes the technique, retrieval version, and how often measurements are taken. Figure 1 illustrates the temporal availability of atmospheric water vapour measurement from each instrument. Figure 2 illustrates the vertical ranges of the datasets.

2.1. Radiosondes

Radiosondes are launched by the EWS twice a day (11:15 and 23:15 UT) using hydrogen-filled balloons. Occasionally, additional radiosondes are launched at other times of day for campaigns. The balloons typically reach the middle of the stratosphere (i.e., 30-33 km) before bursting.

The EWS used Vaisala-built RS92 radiosonde models during the timeframe examined in this study. These sensors are widely used by meteorological stations around the world. RS92 relative humidity (RH) measurements are made using thin-film capacitance sensors. The variable of interest for this study is volume mixing ratio (VMR) in parts per million by volume (ppmv). RH measurements from the radiosondes can be converted to mixing ratio using:

$$\text{VMR}(z) = \frac{\text{RH}(z) e_s(T(z))}{P(z)}, \quad (1)$$

where RH is the relative humidity, T and P are the temperature and pressure at a given altitude (z), and e_s is the temperature-dependent saturation vapour pressure of water vapour with respect to liquid water. The e_s equation of Hyland & Wexler (1983) is used for consistency with Vaisala humidity measurement calibration (Miloshevich, 2006).

Deleted: is

As the balloon rises through the atmosphere, there comes a point where the humidity sensor can no longer report a meaningful value. Limiting the radiosonde humidity measurements to below the tropopause height (TPH) or a typical tropopause value usually ensures that only physically meaningful observations are used; however, this potentially removes valid and useful information.

Eureka radiosonde humidity profiles often have clear structure and information about water vapour above the tropopause, which is typically between 8 and 12 km. Miloshevich et al. (2009) found that the tropopause is not a limiting factor for RS92 humidity measurements, and reported close agreement between bias-corrected radiosonde and frostpoint hygrometer (FPH) profiles at temperatures below -70°C and below mixing ratios of 5 ppmv. They recommended limiting radiosondes to pressures greater than 100 hPa during daytime and 75 hPa at night. The mean altitude at which the atmosphere above Eureka has a pressure of 100 hPa is 16.01 km ($\sigma = 0.47$ km), based on radiosonde measurements between 1961 and 2017. We limit radiosonde humidity measurements to altitudes below 15 km for this study as a quality control measure.

RS92 humidity measurements are also known to be affected by solar heating and low temperature calibration error dry biases, as well as errors due to sensor response lag (Vömel et al., 2007a; Miloshevich et al., 2009). The dry bias caused by solar heating of the sensor is not significant in Eureka during winter due to the lack of sunlight; however, it can affect measurements during the sun-lit portion of the year. The calibration error and time-lag error affect low temperature measurements, and are relevant for Eureka conditions. To correct for known biases in a consistent, transparent, and well-documented manner, Eureka radiosonde measurements have been processed with software developed by the GRUAN, described by Dirksen et al. (2014). Eureka is not a formal GRUAN-participating site and the data are not a formal GRUAN data product; however, available raw Eureka radiosonde measurement files were processed by the GRUAN team for use in this study. This processing also calculates uncertainties for reported values and recovers flight details (e.g., latitude, longitude). Only raw files between September 3, 2008 and October 7, 2017 were available for processing. Minor gaps within that timeframe exist. In total, 5515 radiosonde profiles which have been processed using GRUAN methodologies are available for Eureka. They have been quality control-filtered to remove any profile with 'rejected' status.

In the troposphere, the uncertainty of Eureka radiosonde water vapour mixing ratio profiles are typically 3 to 5%. In the LMS, the uncertainty varies from profile-to-profile, ranging from 3% to above 50%. Uncertainty in the water vapour mixing ratio, calculated by propagating uncertainties in Equation 1 by quadrature, is dominated by the relative humidity uncertainty.

Deleted: (and sometimes parts of the lower stratosphere)

Deleted: greatly

Deleted: .

Temperature measurement uncertainties are typically a few tenths of a degree. Pressures similarly have uncertainties on the order of tenths of a hPa. There are occasionally thin dry layers in the middle troposphere that have larger humidity uncertainty. These profile elements are kept. If there are sections of the profile larger than 500 m in the troposphere with high uncertainty values, the entire profile is filtered out.

In the lower stratosphere, the profile reaches a point where the uncertainty increases rapidly. This point changes from profile-to-profile. We limit each individual water vapour profile to the altitude where this rapid increase in uncertainty occurs by finding where the uncertainty first reaches 20%. This is typically a few kilometres above the tropopause. Thus, each radiosonde profile has a different altitude range, depending on the height reached by the balloon and the uncertainty of the measurements.

5 The mean altitude reached by the filtered profiles is 11.3 km ($\sigma = 4.4$ km).

Once launched, radiosonde balloons drift away from the site due to winds. The radiosondes used in this study stayed within a mean distance of 29.8 km ($\sigma = 16.5$ km) from Eureka while under 15 km altitude. The mean time to reach 15 km altitude was 54.4 minutes ($\sigma = 6.2$ minutes).

2.2. PEARL 125HR

10 The Bruker-made IFS 125HR FTIR spectrometer used for this study is located at the Ridge Lab. Installed in July 2006, the 125HR records high-resolution (0.0035 cm^{-1}) mid-infrared (MIR) solar absorption spectra in the framework of the Network for the Detection of Atmospheric Composition Change (NDACC) (Batchelor et al., 2009). Because this technique relies on sunlight, measurements require clear-sky conditions. Due to PEARL's 80° N latitude, there are no 125HR measurements from mid-October to mid-February (i.e., during Polar Night). Even in mid-summer, the high latitude FTIR spectrometer
15 measurements occur at relatively large solar zenith angle (SZA). The minimum SZA at Eureka is 56.5° . This measurement geometry means that the 125HR typically samples the atmosphere south of the Ridge Lab. During the 24-hour sunlight of Polar Day, during the high Arctic summer, the Sun's position is north of the instrument during what is usually night. However, 125HR measurements are not made overnight due to on-site operator limitations and the lack of an automated shut down trigger in the case of problematic weather.

20 The 125HR water vapour dataset used in this study was produced using the retrieval technique summarized in Schneider et al. (2012) and Barthlott et al. (2017), as part of the MULTI-platform remote Sensing of Isotopologues for investigating the Cycle of Atmospheric water (MUSICA) project. MUSICA uses the existing NDACC FTIR spectrometer observations to produce precise and accurate measurement of water vapour isotopologues. This process applies an Optimal Estimation technique based on Rodgers (2000) and the PROFITT retrieval code of Hase et al. (2004) using a combination of strong and weak absorption
25 features on a logarithmic scale. The accuracy of the MUSICA water vapour profiles is about 10% (Schneider et al., 2016). The sensitivity of the retrieval to the atmosphere (i.e., the sum of the averaging kernel rows) varies seasonally due to the dependence on the SZA. The retrieval is typically sensitive throughout the troposphere (i.e., sensitivity above 0.9) and there is some sensitivity in the lower stratosphere (e.g., sensitivity above 0.5). The MUSICA retrieval's sensitivity to the lower stratosphere is maximum during March, which is also when ACE coincidences occur with Eureka. [The mean degrees of freedom for signal \(DOFS\) of the Eureka MUSICA retrievals is 2.9.](#)

30 MUSICA ground-based FTIR products nominally exclude measurements recorded at SZAs greater than 78.5° . This filter has been removed for this study. Due to Eureka's high-latitude location, this filter removes all measurements between February and the end of March, as well as between September and mid-October. A study of the MUSICA water vapour total column

Deleted: 2016

Deleted: 2016

Deleted:

dataset derived from the PEARL 125HR showed that the SZA limit was likely unnecessarily strict, as agreement did not change between the 125HR and other instruments when the SZA limit was relaxed (Weaver et al., 2017). Standard quality control of the MUSICA dataset, which was applied to the data used here, is described in detail by Barthlott et al. (2016).

2.3. ACE on SCISAT

5 The Canadian Space Agency's (CSA's) SCISAT was launched into a high-inclination (74°) 650 km altitude Earth orbit on August 12, 2003. This orbit enables limb-viewing measurements over the polar regions, as well as other latitudes. There are two primary Atmospheric Chemistry Experiment (ACE) instruments aboard SCISAT, ACE-Fourier transform spectrometer (ACE-FTS) and ACE-Measurement of Aerosol Extinction in the Stratosphere and Troposphere Retrieved by Occultation (ACE-MAESTRO). They share a sun-tracker. ACE solar occultation limb-viewing observations involve keeping the sun-tracker pointed at the Sun as the satellite approaches a sunrise or sunset during its orbit and taking sequences of atmospheric and exo-atmospheric absorption spectra.

Coincidences between ACE and Eureka occur during the months of February, March, September, and October. 348 out of 551 coincidences between ACE and Eureka between August 2006 and March 2017 occurred during February and March.

2.3.1. ACE-FTS

15 ACE-FTS is an FTIR spectrometer built by ABB Inc. It acquires spectra between 750 and 4400 cm^{-1} at a resolution of 0.02 cm^{-1} (Bernath et al., 2005). This series of measurements, taken every 2 seconds, is used to retrieve trace gas profiles between the mid-troposphere and 150 km with a vertical resolution ranging between 3 and 4 km (Boone et al., 2013). This technique has a horizontal resolution of ~300 km (Bernath, 2017).

Deleted: 2005;

This study uses ACE-FTS v3.6 data, provided on the 1-km altitude grid in water vapour mixing ratio. Measurements with quality control flags identifying outliers, high percent errors, or instrument/processing errors were filtered out, following recommendations in Sheese et al. (2015). The water vapour retrieval is limited to altitudes between 5 and 100 km.

The validation of an earlier version (v2.2) of ACE-FTS (and to a limited extent, ACE-MAESTRO) water vapour retrievals was examined by Carleer et al. (2008). They concluded that ACE-FTS measurements provide accurate H₂O measurements in the stratosphere (better than 5% from 15–70 km) but expressed no firm conclusions about its water vapour measurements in the upper troposphere. Comparisons to FPH measurements showed a possible small dry bias in ACE-FTS measurements at altitudes near 10 km.

Sheese et al. (2017) examined the current ACE-FTS v3.6 H₂O product (as well as other molecules) by comparing it with co-located MIPAS and MLS measurements by hemisphere. Correlations between ACE-FTS and MLS were observed to be greater than between ACE-FTS and MIPAS. Their analysis examined stratospheric altitudes, where a mean relative difference in the ACE-FTS water vapour product was observed above 16 km ranging from -12 to 2%. In addition, tight coincidence criteria of 15 minutes and 25 km were applied to examine agreement near the hygropause. A mean dry bias of 20% was observed in ACE-FTS profiles relative to MIPAS v5 and MLS v3.3/3.4 at 13 km altitude.

2.3.2. ACE-MAESTRO

ACE-MAESTRO is a dual spectrometer with a wavelength range of 285–1015 nm and a resolution of 1.5 - 2.5 nm (McElroy et al., 2007). 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).

- 5 Water vapour profiles are retrieved from ACE-MAESTRO optical depth spectra. The tangent height registration of the optical depth spectra relies on matching simulated O₂ slant columns obtained from air density profiles, based on ACE-FTS temperature and pressure, with slant columns observed by ACE-MAESTRO using the O₂ A band. The water vapour profiles are retrieved on an altitude grid that matches the vertical sampling. Within 500 km of Eureka, ACE-MAESTRO water vapour profiles include altitudes ranging between 4 and 25 km.
- 10 The ACE-MAESTRO dataset is sparser than the ACE-FTS dataset for two main reasons. ACE-MAESTRO pointing determination requires the existence of ACE-FTS data, so the available ACE-MAESTRO occultation events are a subset of the ACE-FTS occultations. In addition, ACE-MAESTRO ozone is a necessary input to the ACE-MAESTRO water vapour retrieval. The ACE-MAESTRO ozone retrieval fails occasionally, causing most of the measurements missing from the ACE-MAESTRO water vapour product relative to ACE-FTS product.

15 2.4. Aqua

The U.S. National Aeronautic and Space Administration (NASA) launched the Aqua satellite into a 705 km altitude Sun-synchronous near-polar orbit on May 4, 2002. Aqua's orbit has a 1:30 pm equatorial crossing time and an inclination of 98.2°. It is part of the A-Train constellation of Earth observation satellites. The primary mission of Aqua instruments is to study the atmospheric component of the global water cycle (Parkinson, 2003).

20 2.4.1. AIRS

The AIRS instrument is a hyperspectral thermal infrared grating spectrometer on board Aqua. Its detector observes Earth-emitted radiance from a nadir-orientation using 2378 channels between 3.7 and 15.7 µm. AIRS acquires an enormous number of measurements, collecting about three million spectra per day (Chahine et al., 2006).

- 25 AIRS water vapour retrievals have been used to study processes such as the water vapour feedback (Dessler et al., 2008), to evaluate climate models (Pierce et al., 2006), and to improve numerical weather forecasting (Chahine et al., 2006). AIRS aims to produce dense global measurements of temperature and humidity at an accuracy comparable to radiosondes. This study uses level 2 AIRS retrieval v6 data, described in detail by Susskind et al. (2003, 2014). The standard temperature product contains 28 pressure levels, while the standard water vapour product has 15 pressure levels from 1100 to 50 hPa (e.g., between the surface and approximately 20 km in altitude near Eureka).

- 30 Only altitudes that meet the “best” level of quality are used for this study, following the guidelines in the AIRS v6 user guide (Olsen et al., 2017). The altitude range for which AIRS profiles are available varies significantly, with fewer passing the quality control filter at low-tropospheric altitudes. The AIRS retrieval is insensitive to water vapour layers with less than 0.01

Deleted: a

mm of integrated water vapour. This approximately translates to water vapour abundances less than 15 ppmv (Olsen et al., 2017), typically affecting profile elements above 15 km near Eureka. AIRS is also limited to altitudes with pressures greater than 100 hPa, and has diminishing sensitivity at altitudes with pressures less than 300 hPa (approximately 9 km near Eureka) (Olsen et al., 2017). As mentioned in the discussion of the radiosondes' altitude range, 100 hPa occurs at approximately 16 km in altitude above Eureka. The relative abundance of AIRS profiles ensures measurements are nonetheless available for comparisons.

2.5. Aura

NASA's Aura satellite was launched into a near-polar Sun-synchronous 705 km orbit on July 15, 2004. It is part of the A-train constellation of Earth observing satellites, orbiting 15 minutes behind Aqua. Aura's orbit has a 98.2° inclination and an equatorial crossing time near 1:45 pm local solar time. Instruments aboard Aura, such as the Microwave Limb Sounder (MLS) and TES, study atmospheric chemistry and dynamics.

2.5.1. MLS

MLS measures radiation emitted from the atmosphere from a limb-viewing geometry. The atmosphere is scanned twice each minute as the satellite progresses through an orbit that offers a nearly global coverage, between 82° N and 82° S. MLS measurements have been used to assess ACE as well as other satellite measurements, e.g., Hegglin et al. (2013) and Sheese et al. (2017). This study uses MLS v4.2 data.

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 \(Livesey et al., 2016\)](#). At Eureka, MLS's lower altitude limit of 316 hPa corresponds to altitudes near 8 km. MLS water vapour profiles agree within 1% of FPH measurements in the stratosphere, i.e. at $P < 100$ hPa (Hurst et al., 2014). Hurst et al. (2016) showed that agreement between MLS v4.2 and the FPH measurements began to diverge in 2010 at a rate of approximately 1% per year. At 215 hPa and 316 hPa, MLS v1.5 was observed to have a dry bias of 11 to 23% relative to 10 geographically dispersed FPH measurement sites (Vömel, 2007b).

2.5.2. TES

TES is an FTS aboard Aura that observes emitted radiance between 650 and 3050 cm^{-1} spectral resolution of 0.10 cm^{-1} when observing in nadir mode and 0.025 cm^{-1} limb viewing mode (Beer et al., 2001). Limb scanning measurements were performed only until May 2005. The TES water vapour retrieval uses nadir observations, which have a footprint of 5 km by 8 km. Routine measurements involve a series of observations continuously for 16 orbits (26 hours).

Measurements are only available near Eureka's high Arctic latitude until September 2008. The latitudinal range of TES measurements was limited to latitudes between 50° S and 70° N in summer 2008 to conserve instrument life (Herman and Osterman, 2014). Measurements were further limited to between 30° S and 50° N in spring 2010. However, high latitude measurements were taken in July 2011 as part of a special observation set.

TES retrieval v6 is used for this study. It is based on an optimal estimation non-linear least-squares approach described by

Bowman et al. (2006). 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).

- 5 Comparisons between TES v5 water vapour and global radiosonde measurements have shown a wet bias of 15% in the middle troposphere (Herman and Kulawik, 2013). Shephard et al. (2008) compared TES water vapour v3 with radiosondes, finding a wet bias in TES retrievals of between 5% in the lower troposphere and 15% in the upper troposphere.

2.6. EnviSat

10 The European Space Agency (ESA)'s Environmental Satellite (EnviSat) was a large platform for Earth observation instruments. Launched into a polar orbit on March 1, 2002, with an inclination of 98.5° and an equatorial crossing time of 10:00 am mean local solar time. Observations from its ten instruments ended in April 2012. On board were two atmospheric limb sounders, the Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) and the Scanning Imaging Absorption Spectrometer for Atmospheric Cartography (SCIAMACHY). The decade of measurements taken by MIPAS and SCIAMACHY have been widely used to study atmospheric composition, and are often used in comparisons to other limb
15 sounders.

2.6.1. MIPAS

MIPAS is an FTIR spectrometer that observes mid-infrared atmospheric emission from a limb-viewing geometry (Fischer et al., 2008). The spectral resolution of MIPAS was reduced from 0.025 cm⁻¹ to 0.0625 cm⁻¹ in 2004 due to technical problems. The timeframe examined in this study, 2006 onwards, is entirely during the reduced spectral resolution period. This
20 measurement mode has improved spatial resolution. In polar regions, the nominal tangent altitude spacing is 1.5 km in the UTLS region.

This study uses MIPAS retrieval v5 and v7 from the Institute of Meteorology and Climate Research (IMK). Both retrieval versions cover the same temporal range. This retrieval technique is described by von Clarmann et al. (2009) and uses Tikhonov regularization. In the UTLS, the profiles are provided on a 1-km grid. At 10 km, the vertical resolution (v5) is 3.3 km the
25 horizontal resolution is estimated to be 206 km (von Clarmann et al., 2009). Quality control filtering is applied according to recommended values. MIPAS water vapour data is recommended for use only above 12 km altitude. However, in this study all available altitudes provided in the official data release are used. MIPAS water vapour profile retrievals reach altitudes as low as 5 km.

30 Stiller et al. (2012) compared an earlier version of the MIPAS IMK retrieval (v4) with cryogenic frostpoint hygrometer (CFH) measurements of water vapour profiles during the Measurements of Humidity in the Atmosphere and Validation Experiments (MOHAVE) campaign near Pasadena, California in October 2009. Above 12 km, MIPAS showed agreement within 10%. Results suggest MIPAS v4 water vapour might be 20-40% wet biased around 10 km.

Deleted: ,

2.6.2. SCIAMACHY

SCIAMACHY is an imaging spectrometer that has limb, nadir, and occultation viewing modes (Bovensmann et al., 1999). Limb measurements of scattered sunlight are the basis for the Institut für Umwelphysik (IUP) v3.01 and v4.2 water vapour retrievals used in this study. Both retrieval versions cover the same temporal range. It is based on the optimal-estimation approach described by Rodgers (2000) using a first-order Tikhonov constraint. The vertical resolution is approximately 3 km. The retrieval calculates a scaling factor for the tropospheric water vapour profile; altitudes below 10 km are not recommended for use and are not used here. The details of this retrieval are described in Weigel et al. (2016) for v3.01. For v4.2 several changes were implemented first of all to improve the aerosol correction and the vertical resolution. Additionally, v4.2 uses all appropriate SCIAMACHY measurements, v3.01 only a subset. One issue for limb sensing is the number of cloud free scenes. This is limited by the sampling approach, which was constrained by the data rate available on Envisat. Weigel et al. (2016) compared MIPAS v3.01 to MIPAS v5, MLS v3.3, and other satellite datasets, in 30° latitudinal bands. Results showed SCIAMACHY limb measurements between 10 and 25 km in altitude were reliable between 11 and 23 km, and accurate to about 10% between 14 and 20 km. Below 14 km, differences with other datasets increase to up to 50%, showing a possible SCIAMACHY v3.01 wet bias, which is most pronounced in the tropics and least in the polar latitudes.

3. Comparison of water vapour measurements

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 profiles at high vertical resolution; 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.

3.1. Method

Coincident profile measurements have been compared using difference and correlation plots. Absolute differences and percent relative differences are calculated using:

$$25 \text{ difference} = X - Y, \quad (2)$$

$$\% \text{ difference} = \frac{(X-Y)}{Y} \times 100\%, \quad (3)$$

where X is the satellite measurement and Y is the reference measurement, e.g., 125HR or radiosondes.

To show the overall agreement observed between the measurements, the absolute and percentage means of coincident profile differences are calculated, i.e., using:

$$\Delta(z) = \frac{1}{N(z)} \sum_{i=1}^{N(z)} [X_i(z) - Y_i(z)], \quad (4)$$

and

$$\Delta_{\%}(z) = 100\% \times \frac{1}{N(z)} \sum_{i=1}^{N(z)} \frac{|X_i(z) - Y_i(z)|}{Y_i(z)}, \quad (5)$$

Altitude ranges for which there are measurements available vary for each contributing matched pair of profiles, resulting in a variable number of profiles contributing to comparisons at each altitude. The number of contributing matches at each altitude level is reported in the comparison figures.

In addition to showing profile comparisons, comparisons at specific representative altitudes are presented. These illustrate the extent of the variability in the overall mean agreement between the datasets.

A minimum number of 15 coincidences is required, i.e. $N > 15$, for results to be reported and shown in the tables and figures.

This aimed to balance the reality that there are limited number of coincidences available and the need to ensure there are a meaningful number of comparison results available at each altitude.

3.1.1. Coincidence criteria

A three-hour temporal coincidence criterion was used for all comparisons and applied in two ways. Firstly, if multiple coincidences were found within this interval, only the closest pair was kept. Each pair of coincident measurements is thus independent of others contributing to the overall assessment of different measurement techniques. This method often results in a smaller time difference between measurements than is otherwise permitted by the criterion. The comparisons were also performed using all possible coincidences within this criterion. While increasing the number of matches, in some cases significantly, the observed agreement between instruments was similar to that for the first method, which is summarized in Table 2 and Table 3. Results using the first method are discussed below. Results of comparisons where all possible coincidence pairs are used are available in Supplementary Tables 1 and 2.

A 500 km spatial coincidence criterion was also applied. The spatial criterion is similar in scale to the horizontal area covered by a limb-viewing satellite measurement. When calculating the distance between PEARL and an ACE observation, the 30 km (calculated geometrically) tangent height of the ACE measurement was used as the satellite measurement's position. This approach has been used for validation, e.g., Fraser et al., 2008.

The difference in measurement geometries, and the long path of a limb-viewing measurement in particular, can result in ACE-FTS measuring a different airmass than the 125HR and radiosondes. Fig. 3 illustrates the variation of water vapour abundances in the region around Eureka using AIRS measurements at 400 hPa (corresponding to altitudes between 6.1 and 7.5 km, with a mean altitude of 6.7 km and a standard deviation of 0.2 km) for two sample months, March and July, in a representative year (2015). Variability in the water vapour abundances in the region around Eureka is seen to be larger in the summer than in the winter. October resembles the results shown for March.

Deleted: .

Deleted: needed

Deleted: ure

Deleted: binned

3.1.2. Smoothing

When comparing satellite profiles with the PEARL 125HR, the comparison instrument's profile was smoothed by the MUSICA averaging kernel of the 125HR measurement to account for the vertical resolution differences between the instruments. The procedure for smoothing followed Rodgers and Connor (2003):

$$5 \quad \mathbf{x}_{smoothed} = \mathbf{A}(\mathbf{x} - \mathbf{x}_a) + \mathbf{x}_a, \quad (6)$$

where \mathbf{x}_a is the MUSICA a priori profile, \mathbf{x} is the comparison instrument profile, and \mathbf{A} is the averaging kernel matrix. Since the MUSICA water vapour retrievals are performed on a logarithmic scale, the smoothed profile is calculated using:

$$\mathbf{x}_{smoothed} = e^{\mathbf{A}(\mathbf{x} - \mathbf{x}_a) + \mathbf{x}_a}, \quad (7)$$

where \mathbf{x} , \mathbf{x}_a , and \mathbf{A} are in \log_e space.

10 Before smoothing, the satellite profile was interpolated to the MUSICA retrieval grid and the MUSICA a priori profile was used to fill gaps in the comparison profile (e.g., altitudes beneath the lower limit of satellite measurements). After smoothing, altitudes for which there were no original data were removed. [Altitude-specific comparisons between satellite measurements and the FTIR are thus presented on the MUSICA retrieval grid, e.g., 6.4 km, 8.0 km, and 9.8 km.](#)

When comparing satellite measurements to the radiosonde profiles, radiosonde profiles were smoothed using the satellite's averaging kernels where possible, i.e., for SCIAMACHY and TES, following the same procedure described for the 125HR. MIPAS retrievals do not use an a priori profile, so the smoothed radiosonde profile is calculated using:

$$\mathbf{x}_{smoothed} = e^{\mathbf{A}\mathbf{x}}. \quad (8)$$

In the cases of ACE-FTS, ACE-MAESTRO, AIRS, and MLS, the radiosonde profiles have been smoothed using Gaussian weighting functions with a full width half maximum (FWHM) that approximates the vertical resolution of the satellite measurement. This procedure is used because ACE instruments do not have averaging kernels. MLS has an averaging kernel for use in the polar regions; however, the user's guide states that the use of the water vapour averaging kernel at the lowest valid altitude levels (i.e., lower stratosphere at 316 hPa and 262 hPa) is not recommended (Livesey et al., 2016). Since these altitudes are of particular interest to this study, the MLS averaging kernels are not suitable. AIRS also has averaging kernels, distributed in supplementary data files; however, the AIRS averaging kernels only capture the information added during the final physical retrieval, but not information extracted from the AIRS radiances during the neural network step. We use of the width of the AIRS weighting functions to estimate a Gaussian [smoothing width that](#) generally overestimates the amount of smoothing. Thus, weighting functions are used in these cases as a reasonable approximate method of smoothing the vertical resolution of these profiles.

To create weighting functions, first, Gaussian functions are calculated using:

$$30 \quad GF(z) = \left(\sqrt{2\pi} \cdot \frac{FWHM}{2\sqrt{2\ln 2}}\right)^{-1} \cdot \exp\left(\frac{-(z-z_0)^2}{2\left(\frac{FWHM}{2\sqrt{2\ln 2}}\right)^2}\right), \quad (9)$$

Deleted: 4

Deleted: 5

Deleted:

Deleted: 6

Deleted: The AIRS retrieval uses many channels effectively synthesizing a narrow weighting function, then is possible from any one channels.

Deleted: smoother

Deleted: 7

where $FWHM$ is the full-width half-maximum, z is the new low-resolution grid point, z_o are the original altitude levels.

Weighting functions were calculated by sampling the GF at the original radiosonde measurement altitude levels and normalizing the GF so that the total weight assigned to all profile elements is equal to one. The weighting functions are different for each pair of coincident profiles because the vertical sampling of each radiosonde profile varies.

5 Lastly, the vertical resolution of radiosonde water vapour VMR profiles were downgraded using the weighting functions (w_f):

$$x_{smoothed}(z_i) = \sum_{i=1}^N w_{fi} \cdot VMR(z). \quad (10)$$

An example of weighting functions used to align the radiosonde measurement with the approximate vertical resolution of ACE-FTS is shown in Figure 4 (a). Fig. 4 (b) shows an example of a radiosonde profile before and after smoothing. Weighting functions with a $FWHM$ equal to 3.0 km have been used to approximate the vertical resolution of ACE-FTS, while comparisons to ACE-MAESTRO, AIRS, and MLS used weighting functions with a $FWHM$ of 1.0 km.

3.2. Comparison results

Differences between individual coincident profiles were calculated. The means of those differences are presented. When reporting a mean agreement in the text, \pm values refer to the standard error in the mean (SEM). Profile results are presented, as well as comparison results at select altitude levels. Results between the satellites and the 125HR at 6.4 km are highlighted because the 125HR has very good sensitivity at that altitude, and this is near the lowermost altitude reached by the ACE measurements. Comparison results between the satellites and the radiosondes are highlighted at 10 km because radiosondes have sensitivity at that altitude and this is the lowermost altitude of other comparison studies, e.g., Sheese et al. (2017), and it is near the lower limit of many satellite datasets.

Some combinations of instruments did not have significant overlap in time, location, or vertical sensitivity. MIPAS and the radiosondes had no coincidences due to a mismatch in the time of day of the measurements as well as the quality control filtering. The temporal ranges of the TES and radiosonde datasets did not overlap. SCIAMACHY did not have any coincidences with the radiosondes, unless the coincidence criterion was expanded to 6 hours. Even then, only 8 matches were found. SCIAMACHY and the 125HR had 201 coincidences; however, SCIAMACHY is limited to altitudes above 10 km, where the 125HR has limited sensitivity.

3.2.1. Ground-based reference measurements

As illustrated in Fig. 5, comparison between the 125HR and 137 coincident radiosonde profiles smoothed by 125HR averaging kernels shows agreement within 5% between 8 and 14 km; the 125HR has a wet bias relative to the radiosonde profiles below 8 km of approximately 8% (with closer agreement below 2 km). This is similar to the 6% wet bias in the PEARL 125HR total columns relative to the Eureka radiosondes reported by Weaver et al. (2017). If all possible coincident pairs are used, rather than limiting comparisons to unique pairs, the number of contributing matches increases to 270 and the agreement is very similar.

Deleted: smoothed profiles were calculated by convolving

Deleted: profile of the radiosonde

Deleted: with

Deleted: 8

Deleted:

3.2.2. ACE-FTS

76 pairs of coincident ACE-FTS and PEARL 125HR measurements show close agreement. Between 6 and 9 km agreement was within 9 ppmv and 13%; between 8 and 14 km, agreement is within 1.4 ppmv and 10%. Full profile comparisons are shown in Fig. 6. The mean difference of 18 coincident profiles at 6.4 km was -6.3 ± 8.4 ppmv ($0.2 \pm 6.8\%$); the time series of differences at 6.4 km are shown in Fig. 7. At 8.0 km, 46 coincident measurements agreed to within 1.4 ± 2.6 ppmv ($7.2 \pm 6.6\%$). Differences at 8.0 km are illustrated in Fig. S1 (a). Correlation plots at 6.4 km, 8.0 km, and 9.8 km are presented in Fig. 8. Between 6 and 14 km, correlation coefficients (R) are between 0.48 and 0.80. Expanding the time criterion to 6 hours nearly doubles the number of coincidences but results in similar agreement. Overall, relative to the 125HR, ACE-FTS shows a wet bias between 8 and 14 km of 7 to 10% and small differences of approximately 10 ppmv (2%) near 6 km (Fig. 6).

108 coincident measurements were found between ACE-FTS and Eureka radiosondes. Profile differences are shown in Fig. 9, alongside results from other comparisons. These differences are also shown in Fig. S2, where ACE-FTS and ACE-MAESTRO comparison results are presented without other satellites for easier reading. Between 7 and 11 km, differences are within 6 ppmv (12%). At 6 km, ACE-FTS and radiosonde profiles mean differences are -13.3 ± 12.1 ppmv ($22.8 \pm 9.2\%$). Differences at 10 km, -5.4 ± 2.0 ppmv ($-9.1 \pm 6.9\%$), are shown in Fig. 10 (a). Differences at 6 km and 8 km are illustrated in the supplementary materials, Fig. S3 (a) and Fig. S4 (a). Correlation plots at 6.4 km, 8.0 km, and 9.8 km are shown in Fig. 11. Correlation coefficients between 6 and 12 km range between 0.52 and 0.94.

In addition, comparisons have been done between the ACE-FTS using AIRS as a reference. Differences at 10 km were -1.5 ± 0.3 ppmv ($-6.1 \pm 1.7\%$), increasing at lower altitudes to -17.0 ± 3.7 ppmv ($39.6 \pm 4.3\%$) at 6 km. Correlation coefficients for altitudes between 6 and 12 km were between 0.62 and 0.81. Correlation plots of ACE-FTS vs. AIRS at 6, 8, and 10 km are shown in Fig. 12.

3.2.3. ACE-MAESTRO

27 coincident measurements found between ACE-MAESTRO and the PEARL 125HR show agreement within 12 ppmv (7%) between 6 and 8 km and within 3 ppmv (12%) between 9 and 14 km. Overall, between 6 km and 14 km, ACE-MAESTRO shows a dry bias of approximately 10% relative to the 125HR (Fig. 6). Examining the agreement at specific altitudes in the middle and upper troposphere shows scatter around the zero line, illustrated in Fig. 7 and Fig. S1.

103 coincident ACE-MAESTRO and radiosonde profiles were found with overlap between 5 and 11 km. Mean differences were large at 5 km, e.g., -84.0 ± 121.1 ppmv ($123.4 \pm 71.1\%$). Percent differences oscillate around -10% between 7 and 10 km. At 8 km, ACE-MAESTRO had 90 coincidences with the radiosondes, with differences of -16.3 ± 8.7 ppmv ($-7.6 \pm 9.4\%$), shown in Fig. S4. At 10 km, absolute and relative mean differences were -2.6 ± 3.2 ppmv ($-5.9 \pm 10.9\%$), respectively, shown in Fig. 10.

In addition, comparisons have been done between the ACE-MAESTRO using AIRS as a reference. Differences at 10 km were -0.7 ± 0.9 ppmv ($-10.5 \pm 3.7\%$), decreasing at lower altitudes to -13.7 ± 7.5 ppmv ($69.9 \pm 13.5\%$) at 6 km. Correlation

Deleted: 11

Deleted:

Deleted:

Deleted: and

Deleted: and

Formatted: Justified, Space Before: 6 pt, Line spacing: 1.5 lines

coefficients for altitudes between 6 and 12 km were about 0.45. Correlation plots of ACE-MAESTRO vs. AIRS at 6, 8, and 10 km are presented in Fig. 12.

3.2.4. Other satellite measurements vs. ground-based references

AIRS

- 5 Close agreement was observed between 3189 coincident AIRS and 125HR measurements and between 2489 coincident AIRS and radiosonde profiles. AIRS profiles agree with the 125HR within 5% between 1 km and 14 km, as shown in Fig. 6. A mean difference of -9.7 ± 3.5 ppmv ($-1.6 \pm 1.5\%$) was observed between AIRS and 125HR measurements at 6.4 km, where both instruments have good sensitivity. This is shown in Fig. 7 (b). In the mid-troposphere, agreement is within 4%. Correlation coefficients at all altitudes are above 0.84. Correlation plots for AIRS vs. 125HR at 6.4, 8.0, and 9.8 km are shown in Figure 8.
- 10 Mean agreement within 5% is observed between AIRS and the radiosondes between 1 and 7 km, as shown in Fig. 9. Differences as large as 13% are observed between 8 km and 14 km. Differences at 10 km are shown in Fig. 9 (b), where scatter around zero is seen. As well, the time series of differences shows a potential seasonality to the agreement, with a low (dry) bias maximum in summer. Tightening the coincidence criteria to 2 hours and 25 km significantly reduces the number of matches, with 45 contributing to comparisons at 1 km and 1255 contributing to comparisons at 8 km. Results from these tighter
- 15 matches show differences of less than 4% between 2 and 7 km, with slightly larger differences at 1 km. Differences remained similar between 8 and 14 km with these stricter coincidence criteria.

MIPAS

- MIPAS v5 and v7 comparisons with the PEARL 125HR show a dry bias of approximately 15% in the upper troposphere. At 6.4 km, the lowest altitude available for comparisons with a reasonable number of coincident measurements ($N = 64$), mean differences using MIPAS v5 were -38.2 ± 11.9 ppmv ($-22.4 \pm 7.8\%$). MIPAS v7 showed similar differences as v5 with respect to the 125HR at 6.4 km, i.e. -46.9 ± 11.2 ppmv ($-25.3 \pm 5.9\%$). The time series of differences between the 125HR and MIPAS datasets at 6.4 km is illustrated in Fig. 7 (c), showing large scatter. Correlation at 6.4 km was moderate ($R = 0.50$). Between 7 and 14 km a good correlation was observed for both retrieval versions ($R > 0.81$). Agreement improves between 7 and 10 km. MIPAS v5 reaches a mean difference of -3.6 ± 0.4 ppmv ($-10.1 \pm 1.1\%$) at 9.8 km. Above 10 km, differences are small, better than 2 ppmv and 7%.
- 20
- 25

No MIPAS measurements were coincident with radiosondes. In part due to the partial overlap of the datasets (September 2008 to April 2012), and also because MIPAS only had Eureka coincidences during mid-day and mid-night, limiting matches within 3 hours of radiosonde launches.

If AIRS is used as a reference, MIPAS v5 and v7 have hundreds or thousands of matches for comparison at each altitude level.

- 30 The results show that MIPAS has a dry bias relative to AIRS of approximately 15% between 6 and 10 km, comparable to the 125HR results.

MLS

Deleted: Page Break

Deleted: agreement

Deleted: $-1.6 \pm 1.5\%$ (

Deleted: $\sigma = 45.9$

Deleted: , e.g., at 3.0 km, the mean difference is $-3.8 \pm 1.6\%$ ($\sigma = 32.7$)...

Deleted: , a mean difference of $7.5 \pm 0.8\%$ ($\sigma = 30.4$)

Deleted: % ($\sigma = 38.1$)

Deleted: % ($\sigma = 33.5$)

Deleted: around the zero line

Deleted: % ($\sigma = 27.7$)

Formatted: Justified, Space Before: 6 pt, Line spacing: 1.5 lines

Deleted: Page Break

Relative to the 125HR, an MLS dry bias is observed in the UTLMS, where mean differences range from -8.8 ± 0.4 ppmv ($-18.5 \pm 0.8\%$) at 8.8 km to -0.0 ± 0.0 ppmv (-42.8 ± 17.8 ppbv; $-0.3 \pm 0.4\%$) at 13.6 km. This can be seen in Fig. 6. At 9.8 km, mean differences between 2443 coincidences were -4.8 ± 0.2 ppmv ($-12.5 \pm 0.6\%$); at 12.0 km, mean differences between 2445 coincidences were -0.4 ± 0.0 ppmv ($-4.6 \pm 0.5\%$).

Deleted: 6

Deleted: ±

Deleted: 8.

Deleted:

Deleted: 2443

Deleted:

Deleted: 0

- 5 MLS comparisons with the radiosondes have overlap only between 9 and 13 km; comparisons are shown in Fig. 9. At altitudes between 9 and 12 km the matched measurements are highly correlated, with R values between 0.83 and 0.92. Comparisons between MLS and radiosondes showed a dry bias at altitudes between 8 and 12 km. At 10 km, MLS had 447 coincidences with radiosonde measurements, with a mean differences of -5.1 ± 1.2 ppmv ($-25.6 \pm 1.4\%$). The time series of differences between MLS and the radiosondes at 10 km is shown in Fig. 10 (c).

Deleted: % ($\sigma = 29.4$)

10 SCIAMACHY

SCIAMACHY could be compared only with the 125HR, as its measurements did not have coincidences with the radiosonde dataset used in this study. 201 SCIAMACHY v3.01 and 1506 SCIAMACHY v4.2 profiles had coincidences with the 125HR; however, these are limited to altitudes above 10 km. Profile comparison results are shown in Fig. 6. For both retrieval versions, a small dry bias is seen with respect to the 125HR at 10.8 and 12.0 km, i.e., 5% for v3.01 and 10% for v4.2. At 13.6 km, mean differences were about 1%.

Deleted: (e.g. 5% for v3.01 and 10% for v4.2)

TES

TES shows moderate agreement with the PEARL 125HR, but TES had only a single coincidence with the Eureka radiosonde dataset. The latter is largely because TES had no coincidences with Eureka after September 2008, except for a few during mid-July 2011 (Fig. 1). 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. The time series of differences at 6.4 km is shown in Fig. 7 (c), where large scatter is seen, e.g., $\sigma = 75.1\%$.

Deleted:

Deleted: -

Deleted: around the zero line

3.3. Summary of profile comparisons

A summary of comparisons between the satellites and the PEARL 125HR is presented in Table 2. Table 3 provides a summary of the comparisons between the satellites and the Eureka radiosondes. In addition to the number of measurements, means, standard deviations, and SEMs at each altitude, these tables also include the medians of the differences. If the distance criterion was reduced to 350 km, similar differences were observed, but with a much smaller number of coincident measurements in some cases. There is no apparent temporal trend in the differences between satellite datasets and the Eureka-based reference measurements.

- 30 In addition to the comparison results presented in Fig. 6 through Fig. 12, six figures are presented in the supplementary materials. Fig. S1 shows the time series of differences for the satellite datasets and 125HR at 8 km. Figs. S3 through S5 show differences between the satellite datasets and the radiosondes at 6, 8, and 12 km altitudes. Two addition figures, formatted in

Deleted: four

the same manner as Fig. 6 and Fig. 9, show profile comparison results for example days where all satellite datasets had coincident measurements with the 125HR (Fig. S6) and with the radiosondes (Fig. S7).

In some comparisons, e.g., the comparison between AIRS and the radiosondes at 12 km, the reported mean of the absolute differences and percent differences were different signs, e.g., the mean of the absolute differences was negative while the mean of the percent differences was positive. This is the result of reporting the mean of individual comparisons, rather than comparing the mean profiles of each instrument. The latter would ensure the sign is always the same in both cases. Percent differences are weighted differently than the absolute differences when the mean is calculated. Histograms were plotted for the differences between each instrument comparison at each altitude discussed in this study. These results (not shown) showed that the differences are typically distributed in a nearly Gaussian manner, justifying the use of the mean, SEM, and standard deviation to characterize the results.

4. Discussion

This study's moderately tight temporal criterion, 3 hours, aimed to minimize the impact of water vapour's variability on the observed agreement. The variability of water vapour over the 500 km distance criterion likely contributes to the differences observed between measurements. This is especially true for lower-tropospheric measurements, given the variability of surface terrain in the region around Eureka. The seasonally-changing tropopause height also introduces a source of variability, particularly for altitudes between 8 and 10 km. In the summer, the TPH is often above 8 km at Eureka, and sometimes is above 10 km. The TPH can be as low as 6 km. H₂O abundances and variability are typically larger at altitudes below the TPH. However, no seasonal pattern in the differences were observed, or pattern with respect to the TPH.

Measurement techniques also result in differences in the air sampled. While radiosondes measure air close to Eureka throughout their profile, the 125HR's solar-viewing geometry primarily samples air south of Eureka due to the large SZA of high-latitude measurements. Limb-sounding satellite measurement techniques used by ACE-FTS, ACE-MAESTRO, MIPAS, MLS, and SCIAMACHY yield vertical profiles by observing across long horizontal stretches of atmosphere. While this technique enables the retrieval to resolve vertical structure, this horizontal path results in profiles containing information about the atmosphere across an extended area. Thus, exact agreement between the satellite and ground-based measurements is not expected. It is worth noting that all of the instruments' measurement techniques observe the atmosphere only in cloud-free conditions, except the Eureka radiosondes.

Since ACE coincidences with Eureka are limited to periods of time when water vapour abundances are relatively similar across the region, the distance criterion is expected to have less impact on the observed agreement than if year-round measurements were compared. Typical March and July water vapour abundances in the area around Eureka are shown in Fig. 3.

Agreement between both ACE instruments and the Eureka reference measurements was closer than that observed in comparisons conducted by Carleer et al. (2008), which examined an earlier version of these datasets (e.g. ACE-FTS v2.2) and reported differences on the order of 40% at altitudes lower than 15 km and a possible dry bias at around 10 km altitude. Sheese et al. (2017) reported an ACE-FTS negative bias ranging between 3 and 20% relative to MLS and MIPAS at around

Deleted: Average

Deleted: ure

Deleted: . These have been calculated by creating 50 x 50 km grid boxes, centred at Eureka, and taking the mean value of all AIRS profile elements at 400 hPa near Eureka within each box between 2006 and 2016...

14 km; however, the Sheese et al. analysis involves measurements taken over a broad range of global geographic locations and did not discuss altitudes below 13 km.

The ACE-FTS comparisons presented here show a positive (wet) bias of between 7 and 10% relative to the 125HR in the 8 to 14 km altitude range. Relative to the Eureka radiosondes, ACE-FTS shows very close agreement (within 4% or 6 ppmv) in the upper troposphere (7 to 9 km). At altitudes above 10 km, a positive (wet) bias relative to the radiosondes is observed, ranging between 12 and 32%, although this corresponds to very small mean differences, i.e. of about 1 ppmv. If AIRS is taken as a reference, a larger number of coincidences are found and similar results are observed, although with closer agreement around 10 km. These results indicate ACE-FTS offers accurate H₂O profiles in the Arctic UTLS region, e.g. down to 7 km.

ACE-MAESTRO profiles show a dry bias relative to the 125HR of approximately 10% down to 7 km. Comparisons to the radiosondes also showed a dry bias, ranging from -3% at 7 km to -21% at 11 km. At 6 km and below, large differences between ACE-MAESTRO and the radiosonde profiles are large, as was the case in the 125HR comparison; however, in both cases there are too few coincidences for firm conclusions. Using AIRS as a reference results in hundreds of coincidences and similar results, e.g. similar magnitudes with an increasingly large difference at altitudes below 7 km.

ACE-MAESTRO shows weak correlations with the Eureka 125HR and radiosonde datasets in Figs. 8 and 11. However, this is likely due to the combination of water vapour's variability, seen in the Figs. 8 and 11 correlation plots involving AIRS, and the relatively low number of coincidences found. As shown in Fig. 12, the number of coincidences and the correlations between ACE-MAESTRO and AIRS are much larger, e.g. $N = 233$ and $R = 0.64$ at 10 km, while the differences are similar to other comparisons, e.g., there were large differences at 6 km. In addition, the correlation and best-fit line are impacted by outlier points at low altitudes (e.g., at 6.4 km in the comparison with the 125HR) that influence the overall statistics because of the relatively small number of coincidences at those altitudes. ACE-FTS correlation plots are also affected by outliers.

For both ACE-FTS and ACE-MAESTRO, measurements at altitudes below approximately 5 km are often not possible because ACE's sun-tracker is unable to lock onto the Sun reliably due to cloud effects and refraction (Boone et al., 2005). This issue may contribute to the larger differences observed at low altitudes. This is especially the case with ACE-MAESTRO, whose retrieval produces profiles extending as low as 4 km with tangent heights determined by extrapolation based on the vertical sampling above 5 km.

AIRS and TES are the only satellite instruments in this study whose measurements are performed in nadir-viewing modes and whose retrieval products reach the lower troposphere. Humidity inversions typically occur near Eureka between 500 m and 2 km in altitude. Sometimes, major structure is seen in the water vapour profile between 2 and 4 km as well. Individual profile-to-profile comparisons with the Eureka radiosondes shows AIRS retrievals do not fully capture structure in the humidity inversion feature, explaining much of the individual profile differences at the lowest altitude levels. This is expected because the vertical resolution of AIRS is not always sufficient to resolve these vertical structures (Susskind et al., 2014). The AIRS user guide warns of occasional 'strange results' in proximity to near-surface humidity inversions, however, the AIRS profiles coincident with Eureka showed no features that were oddly shaped or clearly erroneous. The magnitude of the inversion was

often inaccurate or the inversion was not seen in the AIRS profile. This could also be in part due to a geographic or temporal mismatch between the measurements.

Similarly, individual profile-to-profile comparisons with the nearest radiosonde profile show TES profiles often capture the general shape of the lower tropospheric humidity profiles structure; however, the smoothing operation is not enough to bring the measurements into agreement. Where radiosondes from earlier or later in the day reveal a humidity profile with less fine vertical structure, agreement between TES and the 125HR was much closer.

5. Conclusions

This study compared high Arctic UTLS water vapour measurements taken by seven satellite-based instruments with measurements acquired by the Eureka radiosondes and the PEARL 125HR. The focus of the work was to assess the UTLS water vapour retrieved from ACE-FTS and ACE-MAESTRO measurements. The ACE instruments' ability to observe UTLS water vapour is a valuable contribution to global atmospheric monitoring, as its profiles extend to lower altitudes than many other satellite-based measurements, particularly those retrieved from limb-viewing observations.

ACE-FTS and ACE-MAESTRO showed good agreement with both the radiosondes and the 125HR in the UTLS. No obvious temporal trend is apparent in the differences. ACE-FTS showed a wet bias of approximately 7 to 10% relative to the 125HR.

An ACE-FTS dry bias of 2 to 9% was observed relative to the radiosondes between 8 and 10 km. While agreement is observed in the upper troposphere, the observed agreement did not reach the 5% accuracy goal set by GCOS. ACE-MAESTRO profiles at altitudes below 7 km had large differences relative to both the radiosondes and the 125HR; between 8 and 10 km, a dry bias between 6 and 18% is observed relative to both the radiosondes and the 125HR. Nonetheless, ACE water vapour measurements showed closer agreement overall with the Eureka reference measurements in the UTLS than did the other satellite datasets examined in this study, with the exception of AIRS.

AIRS water vapour profiles showed close agreement with both the 125HR and radiosonde measurements, i.e. within the 5% GCOS target. The observed accuracy of the AIRS measurements suggests they can be used for analysis of humidity conditions near Eureka. Given the high density and frequency of AIRS measurements, it would be worthwhile to use AIRS measurements to create climatologies of water vapour conditions near the site, and also to examine patterns of water vapour abundances in the region. AIRS data may also be useful for validation studies in cases where radiosonde and 125HR measurements do not offer sufficient numbers of coincident measurements. In addition, global UTLS comparisons between AIRS and ACE water vapour measurements could also be examined to better understand the accuracy of the ACE-FTS and ACE-MAESTRO water vapour datasets.

MIPAS and SCIAMACHY comparisons at altitudes where the data is recommended (i.e., above 10 km) showed agreement within 6% of the 125HR. Coincidences with the radiosondes were not available. At UTLS altitudes where the MIPAS data is not recommended for use, but is included in the publicly available data product, large differences and variability were observed. This supports the recommendation to limit the use of MIPAS v5 and v7 water vapour profiles to 12 km and above. MIPAS v5 and v7 and SCIAMACHY v3.01 and v4.2 comparison results were very similar.

Deleted: the WMO

Deleted: WMO

MLS comparisons with the radiosondes and 125HR between 8 and 12 km showed a dry bias. This aligns with UTLS-region MLS dry biases observed by Hurst et al. (2016) and Vömel et al. (2007b) using FPH measurements.

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.

Deleted: the advantage of high

Deleted: as consistent

Deleted: the

Deleted: These

Data availability:

The satellite datasets used in this study are available for download through their respective websites. All require registration except TES and MUSICA.

ACE-FTS and ACE-MAESTRO: <http://www.ace.uwaterloo.ca/data.php>

5 AIRS: https://airs.jpl.nasa.gov/data/get_data

MIPAS (IMK retrieval): <https://www.imk-asf.kit.edu/english/308.php#org0f1a3a1>

MLS: <https://mls.jpl.nasa.gov/>

SCIAMACHY: <http://www.iup.uni-bremen.de/scia-arc/>

TES: <https://tes.jpl.nasa.gov/data/>

10 The PEARL 125HR water vapour data are available through the online MUSICA repository at:

<ftp://ftp.cpc.ncep.noaa.gov/ndacc/MUSICA/>.

However, the dataset used in this study has relaxed the usual solar zenith angle criterion to expand available measurements at the high-latitude site of Eureka. Please contact Dan Weaver (dweaver@atmosp.physics.utoronto.ca) regarding access to this dataset.

15 Radiosonde data used in this study are owned by Environment and Climate Change Canada and are not currently available online. Please contact Dan Weaver (dweaver@atmosp.physics.utoronto.ca) regarding access to this dataset.

Deleted: Satellite

Deleted: pages

Deleted: is

Deleted:

Deleted: the authors

Deleted: is

Deleted: is

Deleted: the authors

Author contributions:

D. Weaver led the study, performed 125HR measurements, gathered the datasets, calculated the comparisons between the datasets, created the figures and tables, and wrote the manuscript. K. Strong advised and guided the work and provided significant editing, advice, and comments. K. A. Walker contributed insight into the ACE-FTS dataset, as well as helpful
5 comments on the manuscript. C. Sioris contributed the ACE-MAESTRO water vapour dataset, as well as helpful comments on the manuscript and discussions about the meteorological conditions at Eureka. M. Schneider performed the 125HR retrievals following the MUSICA technique and offered detailed comments. H. Vömel offered insight into the radiosonde data and its use. M. Sommer processed the raw Eureka radiosonde measurement files according to the GRUAN technique. C. T. McElroy contributed insight and comments regarding the ACE-MAESTRO measurement comparisons.

10 J. P. Burrows proposed and leads the SCIAMACHY project; in this context, initiating the concept used for the limb H₂O profiles algorithm, he has contributed to and advised on its evolution, as used in this study. A. Rozanov leads the limb retrieval group which has developed key parts of the H₂O retrieval algorithm. K. Weigel is an expert on limb remote sensing having led the development the V3.1 and V.4 H₂O limb product, validated the data products and initiating and coordinating the IUP contribution to this study.

15 W. G. Read contributed expertise about the MLS dataset. E. Fishbein offered helpful comments on the manuscript and insight regarding the AIRS dataset. G. Stiller offered helpful comments about the manuscript and insight on the MIPAS dataset.

Competing interests:

The authors declare that they have no conflict of interest.

Deleted: ,

Deleted: Alexei

Deleted: Katja

Acknowledgements.

This work was primarily funded by the National Sciences and Engineering Research Council (NSERC) through the Probing the Atmosphere of the High Arctic (PAHA) project. Spring visits to PEARL were made as part of the Canadian Arctic ACE/OSIRIS Validation Campaigns funded by the Canadian Space Agency (CSA), with additional support from Environment and Climate Change Canada (ECCC), NSERC, and the Northern Scientific Training Program.

CANDAC/PEARL funding partners are the Arctic Research Infrastructure Fund, Atlantic Innovation Fund/Nova Scotia Research Innovation Trust, Canadian Foundation for Climate and Atmospheric Science, Canada Foundation for Innovation, CSA, ECCC, Government of Canada International Polar Year, NSERC, Ontario Innovation Trust, Ontario Research Fund, Indian and Northern Affairs Canada, and the Polar Continental Shelf Program. This work also received funding from the NSERC CREATE Training Program in Arctic Atmospheric Science, and the CSA-supported Canadian FTIR Observing Network (CAFTON) and Arctic Validation and Training for Atmospheric Research in Space (AVATARS) projects.

The authors would like to thank PEARL PI James Drummond, PEARL Site Manager Pierre Fogal and the CANDAC operators for logistical and operational support at Eureka; ECCC for providing the radiosonde data; Rodica Lindenmaier, Rebecca Batchelor, and Joseph Mendonca for 125HR measurements; and CANDAC Data Manager Yan Tsehtik.

The authors wish to thank the staff at ECCC's Eureka Weather Station for logistical and on-site support.

MUSICA has been funded by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement number 256961.

The Atmospheric Chemistry Experiment is a Canadian-led satellite mission mainly supported by the CSA.

We thank the satellite data retrieval and validation teams for the ACE, AIRS, MIPAS, MLS, SCIAMACHY, and TES missions, as well as their funding agencies.

Part of the research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

The SCIAMACHY limb water vapour data sets v3.01 and v4.2 are a result of the DFG (German Research Council) Research Unit "Stratospheric Change and its Role for Climate Prediction" (SHARP) and the ESA SPIN (ESA SPARC Initiative) and SQWG (SCIAMACHY Quality Working Group) projects and were calculated using resources of the German HLRN (High-Performance Computer Center North).

Figures 3, 5, 8, 11, and 12 were produced using python and the matplotlib and numpy packages. The authors would like to thank those development communities.

6. 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., Pazmino, A., Roth, C., Savastiouk, V., Walker, K., Weaver, D., and Zhao, X.: Validation of ACE and OSIRIS ozone and NO₂ measurements using ground-based instruments at 80°N. *Atmos. Meas. Tech.*, 5, 927-953, doi:10.5194/amt-5-927-2012, 2012.
- Barthlott, S., Schneider, M., Hase, F., Blumenstock, T., Kiel, M., Dubravica, D., García, O. E., Sepúlveda, E., Mengistu Tsidu, G., Takele Kenea, S., Grutter, M., Plaza, E. F., Stremme, W., Strong, K., Weaver, D., Palm, M., Warneke, T., Notholt, J., Mahieu, E., Servais, C., Jones, N., Griffith, D. W. T., Smale, D., and Robinson, J.: Tropospheric water vapour isotopologue data (H₂¹⁶O, H₂¹⁸O and HD¹⁶O) as obtained from NDACC/FTIR solar absorption spectra. *Earth Syst. Sci. Data*, 9, 15-29, <https://doi.org/doi:10.5194/essd-9-15-2017>, 2017.
- Batchelor, R., Strong, K., Lindenmaier, R., Mittermeier, R., Fast, H., Drummond, J. R., and Fogal, P.: A New Bruker IFS 125HR FTIR Spectrometer for the Polar Environment Atmospheric Research Laboratory at Eureka, Nunavut, Canada: Measurements and Comparison with the Existing Bomem DA8 Spectrometer. *J. Atmos. Ocean. Tech.*, 26, 1328-1340, 2009.
- Beer, R., Glavich, T. A., and Rider, D. M.: Tropospheric emission spectrometer for the Earth Observing System's Aura Satellite. *Appl. Opt.*, 40, 2356-2367, 2001.
- Bernath, P., McElroy, C., Abrams, M., Boone, C., Butler, M. C.-P., Camy-Peyret, C., Carleer, M., Clerbaux, M., Coheur, P.-F., Colin, R., DeCola, P., DeMaziere, M., Drummond, J.R., Dufour, D., Evans, W.F.J., Fast, H., Fussen, D., Gilbert, K., Jennings, D.E., Llewellyn, E.J., Lowe, R.P., Mahieu, E., McConnell, J.C., McHugh, M., McLeod, S.D., Michaud, R., Midwinter, C., Nassar, R., Nichitiu, F., Nowlan, C., Rinsland, C.P., Rochon, Y.J., Rowlands, N., Semeniuk, K., Simon, P., Skelton, R., Sloan, J.J., Soucy, M.-A., Strong, K., Tremblay, P., Turnbull, D., Walker, K.A., Walker, I., Wardle, D.A., Wehrle, V., Zander, R., and Zou, J.: Atmospheric Chemistry Experiment (ACE): Mission overview. *Geophys. Res. Lett.*, 32, L15S01, 2005.
- Bernath, P. F.: The Atmospheric Chemistry Experiment (ACE). *J. Quant. Spectrosc. Ra.*, 186, 3-16, <https://doi.org/10.1016/j.jqsrt.2016.04.006>, 2017.
- Boone, C. D., Nassar, R., Walker, K. A., Rochon, Y., McLeod, S. D., Rinsland, C. P., and Bernath, P. F.: Retrievals for the Atmospheric Chemistry Experiment Fourier-transform spectrometer. *Appl. Opt.*, 44, 7218-7231, doi:10.1364/AO.44.007218, 2005.
- Boone, C.D., Walker, K.A., and Bernath, P.F.: Version 3 Retrievals for the Atmospheric Chemistry Experiment Fourier Transform Spectrometer (ACE-FTS) in The Atmospheric Chemistry Experiment ACE at 10: A Solar Occultation Anthology, Peter F. Bernath (Editor), A. Deepak Publishing, Hampton, Virginia, 2013.

Deleted: ACIA: Arctic Climate Impact Assessment. Cambridge University Press, 1042p, 2005.†

Deleted: Discuss.

Deleted: doi:10.5194/essd-2016-9

Deleted: 2016

Formatted: English (CAN)

Formatted: Space Before: 6 pt

Formatted: Bibliography, Space Before: 6 pt

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Default Paragraph Font, English (UK)

Deleted: Bloch, M. and Karasinski, G.: Water vapour mixing ratio profiles over Hornsund, Arctic. Intercomparison of lidar and AIRS results. *Acta Geophys.*, 62, Iss. 2, 290-301, doi:10.2478/s11600-013-0168-3, 2014.†

Formatted: English (CAN)

Bovensmann, H., Burrows, J.P., Buchwitz, M., Frerick, J., Noël, S., Rozanov, V.V., Chance, K.V., and Goede, A.P.: SCIAMACHY: Mission Objectives and Measurement Modes. *J. Atmos. Sci.*, 56, 127–150, doi.org/10.1175/1520-0469(1999)056<0127:SMOAMM>2.0.CO;2, 1999.

Formatted: English (CAN)

Formatted: Bibliography

Bowman, K.W., Rodgers, C. D., Sund-Kulawik, S., Worden, J., Sarkissian, E., Osterman, G., Steck, T., Lou, M., Eldering, A., Shepherd, M., Worden, H., Lampel, M., Clough, S., Brown, P., Rinsland, C., Gunson, M., and Beer, R.: Tropospheric Emission Spectrometer: Retrieval Method and Error Analysis. *IEEE Trans. Geosci. Remote Sensing*, 44, (5) 1297- 1307, 2006.

[Buchholz, R. R., Deeter, M. N., Worden, H. M., Gille, J., Edwards, D. P., Hannigan, J. W., Jones, N. B., Paton-Walsh, C., Griffith, D. W. T., Smale, D., Robinson, J., Strong, K., Conway, S., Sussmann, R., Hase, F., Blumenstock, T., Mahieu, E., and Langerock, B.: Validation of MOPITT carbon monoxide using ground-based Fourier transform infrared spectrometer data from NDACC. *Atmospheric Measurement Techniques*, 10\(5\), 1927–1956. <https://doi.org/10.5194/amt-10-1927-2017>, 2017.](#)

Carler, M., Boone, C., Walker, K., Bernath, P., Strong, K., Sica, R., Randall, C.E., Vömel, H., Kar, J., Hopfner, M., Milz, M., von Clarmann, T., Kivi, R., Valverde-Canossa, J., Sioris, C., Izawa, M.R.M., Dupuy, E., McElroy, C.T., Drummond, J.R., Nowlan, C.R., Zou, J., Nichitiu, F., Lossow, S., Urban, J., Murtagh, D., and Dufour, D.G.: Validation of water vapour profiles from the Atmospheric Chemistry Experiment (ACE). *Atmos. Chem. Phys. Discuss.*, 8, 4499-4559, 2008.

15 [Chahine, M. T.: GEWEX: The Global Energy and Water Cycle Experiment. *Eos*, 73 \(2\), 13–14, doi:10.1029/91EO00007, 1992.](#)

Deleted: ¶

Formatted: English (CAN)

Formatted: Bibliography

Formatted: Default Paragraph Font, English (UK)

Deleted: Dessler, A. E. and Sherwood, C. S.: A Matter of Humidity. *Science*, 323, 1020-1021, 2009. doi:10.1126/science.1171264¶

[Chahine, M. T., Pagano, T.S., Aumann, H. H., Atlas, R., Barnett, C., Blaisdell, J., Chen, L., Divakarla, M., Fetzer, E.J., Goldberg, M., Gauthier, C., Granger, S., Hannon, S., Irion, F.W., Kakar, R., Kalnay, E., Lambrigtsen, B.H., Lee, S., Le Marshall, J., McMillan, W. W., McMillin, L., Olsen, E. T., Revercomb, H., Rosenkranz, P., Smith, W.L., Staelin, D., Strow, L.L., Susskind, J., Tobin, D., Wolf, W., and Zhou, L.: AIRS. *Bull. Amer. Meteor. Soc.*, 87, 911–926, doi:10.1175/BAMS-87-7-911, 2006.](#)

Dessler, A. E., Zhang, Z., and Yang, P.: Water vapor climate feedback inferred from climate fluctuations, 2003-2008. *Geophys. Res. Lett.*, 35, L20704, 2008.

[Dessler, A. E., Schoeberl, M. R., Wang, T., Davis, S. M., and Rosenlof, K. H.: Stratospheric water vapor feedback. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.1310344110>, 2013.](#)

Dirksen, R.J., Sommer, M., Immler, F.J., Hurst, D.F., Kivi, R., and Vömel, H.: Reference quality upper-air measurements: GRUAN data processing for the Vaisala RS92 radiosonde. *Atmos. Meas. Tech.*, 7, 4463-4490, 2014.

[Durre, I., Vose, R. S., Wuertz, D. B.: Overview of the Integrated Global Radiosonde Archive. *J. Climate*, 19, 53-68, doi:10.1175/JCLI3594.1, 2006.](#)

Deleted: Dupuy, E., Walker, K.A., Kar, J., Boone, C.D., McElroy, C.T., Bernath, P.F., Drummond, J.R., Skelton, R., McLeod, S.D., Hughes, R.C., Nowlan, C.R., Dufour, D.G., Zou, J., Nichitiu, F., Strong, K., Baron, P., Mevilacqua, R.M., Blumenstock, T., Bodeker, G.E., Borsdorff, T., Bourassa, A.E., Bovensmann, H., Boyd, I.S., Bracher, A., Brogniez, C., Burrows, J.P., Catoire, V., Ceccherini, S., Chamberlat, S., Christensen, T., Coffey, M.T., Cortesi, U., Davies, J., De Clerq, C., Degenstein, D.A., De Mazière, M., Demoulin, P., Dodion, J., Firanski, B., Fischer, H., Forbes, G., Fruidevaux, L., Fussen, D., Gerard, P., Godin-Beekmann, S., Goutail, F., Granville, J., Griffith, D., Haley, C.S., Hannigan, J.W., Hopfner, M., Jin, J.J., Jones, A., Jones, N.B., Jucks, K., Kagawa, A., Kasai, Y., Kerzenmacher, T.E., Kleinbohl, A., Klekociuk, A.R., Kramer, I., Kullmann, H., Kuttippurath, J., Kyrola, E., Lambert, J.-C., Libesey, N.J., Llewellyn, E.J., Lloyd, N.D., Mahieu, E., Manney, G.L., Marshall, B.T., McConnell, J.C., McCormick, M.P., McDermid, I.S., McHugh, M., LeLinden, C.A., Mellqvist, J., Mizutani, K., Murayama, Y., Murtagh, D.P., Oelhaf, H., Parrish, A., Petelina, S.V., Piccolo, C., Pommereau, J.-P., Randall, C.E., Robert, C., Roth, C., Schneider, M., Senten, C., Steck, T., Strandberg, A., Strawbridge, K.B., Sussman, R., Urban, J., Vanhellemont, F., Vigouroux, C., von Clarmann, T., von der Gathen, P., von Savigny, C., Waters, J.W., Witte, J.C., Wolff, M., and Zawodny, J.M.: Validation of ozone measurements from the Atmospheric Chemistry Experiment (ACE). *Atmos. Chem. Phys.*, 9, 287-343, 2009.¶

Fraser, A., Adams, C., Drummond, J.R., Goutail, F., Manney, G., and Strong, K.: The Polar Environment Atmospheric Research Laboratory UV-visible ground-based spectrometer: First measurements of O₃, NO₂, BrO, and OCIO columns. *J. Quant. Spectrosc. Ra.*, 110:986 – 1004, doi:10.1016/j.jqsrt.2009.02.034, 2008.

Deleted: ¶

5 Fischer, H., Birk, M., Blom, C., Carli, B., Carlotti, M., von Clarmann, T., Delbouille, L., Dudhia, A., Ehhalt, D., Endemann, M., Flaud, J. M., Gessner, R., Kleinert, A., Koopman, R., Langen, J., López-Puertas, M., Mosner, P., Nett, H., Oelhaf, H., Perron, G., Remedios, J., Ridolfi, M., Stiller, G., and Zander, R.: MIPAS: an instrument for atmospheric and climate research. *Atmos. Chem. Phys.*, 8, 2151-2188, doi:10.5194/acp-8-2151-2008, 2008.

Formatted: English (CAN)

Fogal, P., LeBlanc, L.M., and Drummond, J.: The Polar Environment Atmospheric Research Laboratory (PEARL): Sounding the Atmosphere at 80° North. *Arctic*, 66 (3), 377-386, 2013.

Formatted: English (CAN)

10 Global Climate Observing System (GCOS): The Global Observing System for Climate: Implementation Needs. GCOS-200. Available at: <https://public.wmo.int/en/programmes/global-climate-observing-system>, last access: 14 June 2018, 2016.

Griffin, D., Walker, K. A., Conway, S., Kolonjari, F., Strong, K., Batchelor, R., Boone, C. D., Dan, L., Drummond, J. R., Fogal, P. F., Fu, D., Lindenmaier, R., Manney, G. L., and Weaver, D.: Multi-year comparisons of ground-based and spaceborne Fourier transform spectrometers in the high Arctic between 2006 and 2013, *Atmos. Meas. Tech.*, 10, 3273-3294, doi:10.5194/amt-10-3273-2017, 2017.

Formatted: Bibliography, Space Before: 6 pt

Hase, F., Hannigan, J. W., Coffey, M. T., Goldman, A., H'opfner, M., Jones, N. B., Rinsland, C. P., and Wood, S. W.: Intercomparison of retrieval codes used for the analysis of high-resolution, ground-based FTIR measurements, *J. Quant. Spectrosc. Ra.*, 87, 25–52, 25-52, 2004.

Deleted: ¶

20 Hegglin, M. I., Tegtmeier, S., Anderson, J., Froidevaux, L., Fuller, R., Funke, B., Jones, A., Lingenfelser, G., Lumpe, J., Pendlebury, D., Remsberg, E., Rozanov, A., Toohey, M., Urban, J., von Clarmann, T., Walker, K. A., Wang, R., and Weigel, K.: SPARC Data Initiative: Comparison of water vapor climatologies from international satellite limb sounders. *J. Geophys. Res. Atmos.*, 118, 11,824-11846, doi:10.1002/jgrd.50752, 2013.

Formatted: English (CAN)

Formatted: Bibliography

25 Herman, R., and Kulawik, S. (Eds.): Earth Observing System (EOS) Tropospheric Emission Spectrometer (TES) Data Validation Report (Version F07_10 data). JPL Internal Report D-33192, Available online at: https://eosweb.larc.nasa.gov/sites/default/files/project/tes/readme/tes_validation_report_v6.pdf, November 5, 2013.

Deleted: Held, I. M. and Soden, B. J.: Water vapour feedback and global warming. *Annu. Rev. Energy Environ.*, 25, 441-475, 2000.¶

Herman, R., and Osterman, G. (Eds.): Earth Observing System (EOS) Tropospheric Emission Spectrometer (TES) Level 2 (L2) Data User's Guide (Up to & including Version 6 data), Version 6.0. JPL Internal Report D-38042, Available online at: <https://tes.jpl.nasa.gov/documents/>, June 20, 2014.

30 Hurst, D. F., Lambert, A., Read, W. G., Davis, S. M., Rosenlof, K. H., Hall, E. G., Jordan, A. F., Oltmans, S. J.: Validation of Aura Microwave Limb Sounder stratospheric water vapor measurements by the NOAA frost point hygrometer. *J. Geophys. Res. Atmos.*, 119, 1612-1625, doi:10.1002/2013JD020757, 2014.

- Hurst, D. F., Read, W. G., Vömel, H., Selkirk, H. B., Rosenlof, K. H., Davis, S. M., Hall, E. G., Jordan, A. F., Oltmans, S. J.: Recent divergences in stratospheric water vapor measurements by frost point hygrometers and the Aura Microwave limb Sounder. *Atmos. Meas. Tech.*, 9, 4447-4457, doi:10.5194/amt-9-4447-2016, 2016.
- Khosrawi, F., Lossow, S., Stiller, G. P., Rosenlof, K. H., Urban, J., Burrows, J. P., Damadeo, R. P., Eriksson, P., Garcia-Comas, M., Gille, J. C., Kasai, Y., Kiefer, M., Nedoluha, G. E., Noël, S., Raspollini, P., Read, W. G., Rozanov, A., Sioris, C. E., Walker, K. A. and Weigel, K.; The SPARC water vapour assessment II: Comparison of stratospheric and lower mesospheric water vapour time series observed from satellites. *Atmos. Meas. Tech.*, 11, 4435-4463. <https://doi.org/10.5194/amt-11-4435-2018>, 2018.
- Lesins, G., Duck, T. J., and Drummond, J. R.: Climate Trends at Eureka in the Canadian High Arctic, *Atmos.-Ocean*, 48 (2), 59–80, 2010.
- Livesey, N. J., Read, W. G., Wagner, P. A., Froidevaux, L., Lambert, A., Manney, G. L., Millán Valle, L. F., Pumphrey, H. C., Santee, M. L., Schwartz, M. J., Wang, S., Fuller, R. A., Jarnot, R. F., Knosp, B. W.: Aura Microwave Limb Sounder Version 4.2x Level 2 data quality and description document, JPL D-333509 Rev. B, Version 4.2x-2.0, May 9, 2016. Available at: <https://mls.jpl.nasa.gov/data/datadocs.php>, last access: 14 June 2018, 2016.
- McElroy, C.T., Nowlan, C.R., Drummond, J.R., Bernath, P.F., Barton, D.V., Dufour, D.G., Midwinter, C., Hall, R.B., Ogyu, A., Ullberg, A., Wardle, D.I., Kar, J., Zou, J., Nichitiu, F., Boone, C.D., Walker, K.A., and Rowlands, N.: The ACE-MAESTRO instrument on SCISAT: description, performance, and preliminary results. *Appl. Optics*, 46 (20), 4341-4356, 2007.
- Miloshevich, L. M., H. Vömel, D. N. Whiteman, B. M. Lesht, F. J. Schmidlin, and Russo, F.: Absolute accuracy of water vapor measurements from six operational radiosonde types launched during AWEX-G and implications for AIRS validation. *J. Geophys. Res.*, 111, D09S10, doi:10.1029/2005JD0060832006, 2006.
- Miloshevich, L., Vömel, H., Whiteman, D., and Leblanc, T.: Accuracy assessment and correction of Vaisala RS92 radiosonde water vapor measurements. *J. Geophys. Res.*, 114, D11305, 2009.
- Müller, R., Kunz, A., Hurst, D. F., Rolf, C., Krämer, M., and Riese, M.: The need for accurate long-term measurements of water vapor in the upper troposphere and lower stratosphere with global coverage. *Earth's Future*, 4, 25-32, doi:10.1002/2015EF000321, 2015.
- Olsen, E. T., Fetzer, E., Hulley, G., Kalmus, P., Manning, E., and Wong, S.: AIRS/AMSU/HSB Version 6 Level 2 product user guide, Jet Propulsion Laboratory, Version 1.6, available at: <https://airs.jpl.nasa.gov/resources/guides>, last access: 14 June 2018, 2017.
- Parkinson, C. L.: Aqua: An earth-observing satellite mission to examine water and other climate variables. *IEEE Trans. Geosci. Remote Sens.*, 41, 173-183, doi:10.1109/TGRS.2002.808319, 2003.

Formatted: Bibliography, Space Before: 6 pt, After: 0 pt, Line spacing: single, Widow/Orphan control, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Formatted: Font: (Default) Times New Roman, 10 pt, Font color: Auto, English (CAN)

Deleted: Hwang, Y.-T., and Frierson, D.M.W.: Increasing atmospheric poleward energy transport with global warming. *Geophys. Res. Lett.*, 37, 24, L24807, 2010. [↕](#)
Immler, F.J., Dykema, J., Gardiner, T., Whiteman, D.N., Thorne, P.W., and Vömel, H.: Reference quality upper-air measurements: guidance for developing GRUAN data products. *Atmos. Meas. Tech.*, 3, 1217-1231, 2010. [↕](#)

Formatted: Font: Times New Roman, English (CAN)

Formatted: Font: Times New Roman, English (CAN)

Formatted: Font: Times New Roman, English (CAN)

Formatted: Font: Times New Roman, Not Italic, English (CAN)

Formatted: Font: Times New Roman, Not Italic, English (CAN)

Formatted: Font: Times New Roman, Not Italic, English (CAN)

Formatted: Font: Times New Roman, English (CAN)

Formatted: Font: Times New Roman, Not Italic, English (CAN)

Formatted: Font: Times New Roman, English (CAN)

Deleted: Kurylo, M. J.: Network for the detection of stratospheric change (NDSC). Proceedings of SPIE. The International Society for Optical Engineering. Remote Sensing of Atmospheric Chemistry, 1491, 168-174, 1991. [↕](#)

Formatted: English (CAN)

Formatted: Bibliography

Moved (insertion) [3]

Deleted: [↕](#)

Moved up [3]: Miloshevich, L. M., H. Vömel, D. N. Whiteman, B. M. Lesht, F. J. Schmidlin, and Russo, F.: Absolute accuracy of water vapor measurements from six operational radiosonde types launched during AWEX-G and implications for AIRS validation. *J. Geophys. Res.*, 111, D09S10, doi:10.1029/2005JD0060832006, 2006. [↕](#)

Formatted: Bibliography

Pierce, D. W., Barnett, T. P., Fetzer, E. J., and Gleckler, P.J.: Three-dimensional tropospheric water vapor in coupled climate models compared with observations from the AIRS satellite system. *Geophys. Res. Lett.*, 33, L21701, doi:10.1029/2006GL027060, 2006.

Formatted: English (CAN)

Formatted: Font: (Default) Cambria Math, English (CAN)

Formatted: English (CAN)

Formatted: English (CAN)

Riese, M., Ploeger, F., Rap, A., Vogel, B., Konopka, P., Dameris, M., and Forster, P.: Impact of uncertainties in atmospheric mixing on simulated UTLS composition and related radiative effects. *J. Geophys. Res.*, 117, D16305, doi:10.1029/2012JD017751, 2012.

Rodgers, C. D.: *Inverse Methods for Atmospheric Sounding: Theory and Practice*. World Scientific Publishing Co., Hackensack, N.J., 2000.

Rodgers, C., and Connor, B.: Intercomparison of remote sounding instruments. *J. Geophys. Res.*, 108, 4116, 2003.

Schneider, M., Barthlott, S., Hase, F., González, Y., Yoshimura, K., García, O. E., Sepúlveda, E., Gomez-Pelaez, A., Gisi, M., Kohlhepp, R., Dohe, S., Blumenstock, T., Wiegeler, A., Christner, E., Strong, K., Weaver, D., Palm, M., Deutscher, N. M., Warneke, T., Notholt, J., Lejeune, B., Demoulin, P., Jones, N., Griffith, D. W. T., Smale, D., & Robinson, J. Ground-based remote sensing of tropospheric water vapour isotopologues within the project MUSICA. *Atmospheric Measurement Techniques*, 5(12), 3007–3027, <https://doi.org/10.5194/amt-5-3007-2012>, 2012.

Formatted: Bibliography

Formatted: Default Paragraph Font, English (UK)

Schneider, M., Wiegeler, A., Barthlott, S., González, Y., Christner, E., Dyroff, C., García, O. E., Hase, F., Blumenstock, T., Sepúlveda, E., Mengistu Tsidu, G., Takele Kenea, S., Rodríguez, S., and Andrey, J.: Accomplishments of the MUSICA project to provide accurate, long-term, global and high-resolution observations of tropospheric $\{H_2O, \delta D\}$ pairs – a review. *Atmos. Meas. Tech.*, 9, 2845–2875, doi:10.5194/amt-9-2845-2016, 2016.

Formatted: Not Superscript/ Subscript

Schröder, M., Lockhoff, M., Shi, L., August, T., Bennartz, R., Borbas, E., Brogniez, H., Calbet, X., Crewell, S., Eikenberg, S., Fell, F., Forsythe, J., Gambacorta, A., Graw, K., Ho, S.-P., Höschel, H., Kinzel, J., Kursinski, E.R., Reale, A., Roman, J., Scott, N., Steinke, S., Sun, B., Trent, T., Walther, A., Willen, U., Yang, Q., 2017: GEWEX water vapor assessment (G-VAP). WCRP Report 16/2017; World Climate Research Programme (WCRP): Geneva, Switzerland; 216 pp., 2017.

Formatted

Sheese, P.E., Boone, C.D., and Walker, K.A.: Detecting physically unrealistic outliers in ACE-FTS atmospheric measurements. *Atmos. Meas. Tech.*, 8, 741–750, 2015.

Deleted: ¶

Sheese, P.E., Walker, K.A., Boone, C.D., Bernath, P.F., Froidevaux, L., Funke, B., Raspollini, P., von Clarmann, T.: ACE-FTS ozone, water vapour, nitrous oxide, nitric acid, and carbon monoxide profile comparisons with MIPAS and MLS. *J. Quant. Spectrosc. Ra.*, 186, 63–80, doi:10.1016/j.jqsrt.2016.06.026, 2017.

Deleted: diat. Transfer

Deleted: ¶

Sioris, C.E., Zou, J., McElroy, C.T., McLinden, C.A., and Vömel, H.: High vertical resolution water vapour profiles in the upper troposphere and lower stratosphere retrieved from MAESTRO solar occultation spectra. *Adv. Space Res.*, 46, 642–650, 2010.

Deleted: Shindell, D.T.: Climate and ozone response to increased stratospheric water vapor. *Geophys. Res. Lett.*, 28, 1551–1554, 2001. ¶

- Sioris, C. E., Zou, J., Plummer, D. A., Boone, C. D., McElroy, C. T., Sheese, P. E., Moeini, O., and Bernath, P. F.: Upper tropospheric water vapour variability at high latitudes – Part 1: Influence of the annular modes. *Atmos. Chem. Phys.*, 16, 3265–3278, 2016.*
- 5 Soden, B. J., Held, I. M., Colman, R., Shell, K. M., Kiehl, J. T., and Shields, C. A.: Quantifying climate feedbacks using radiative kernels. *J. Clim.*, 21, 3504-3520, doi:10.1175/2007JCLI2110.1, 2008.
- Solomon, S., Rosenlof, K. H., Portmann, R. W., Daniel, J. S., Davis, S. M., Sanford, T. J., and Plattner, G.-K.: Contributions of Stratospheric Water Vapor to Decadal Changes in the Rate of Global Warming. *Science*, 327, doi:10.1126/science/1182488, 2010.
- 10 Stiller, G. P., Kiefer, M., Eckert, E., von Clarmann, T., Kellmann, S., Garcia-Comas, M., Funke, B., Leblanc, T., Fetzer, E., Froidevaux, L., Gomez, M., Hall, E., Hurst, D., Jordan, A., Kämpfer, N., Lambert, A., McDermid, I. S., McGee, T., Miloshevich, L., Nedoluha, G., Read, W., Schneider, M., Schwartz, M., Straub, C., Toon, G., Twigg, L. W., Walker, K., and Whiteman, D. N.: Validation of MIPAS IMK/IAA temperature, water vapour, and ozone profiles with MOHAVE-2009 campaign measurements. *Atmos. Meas. Tech.*, 5, 289-320, doi:10.5194/amt-5-289-2012, 2012.
- 15 Susskind, J., Barnett, C. D., and Blaisdell, J. M.: Retrieval of atmospheric and surface parameters from AIRS/AMSU/HSB data in the presence of clouds. *IEEE Trans. Geosci. Remote Sens.*, 41, doi:10.1109/TGRS.2002.808236, 390-409, 2003.
- Susskind, J., Blaisdell, J. M., and Iredell, L.: Improved methodology for surface and atmospheric soundings, error estimates, and quality control procedures: the atmospheric infrared sounder science team version-6 retrieval algorithm. *J. Appl. Rem. Sens.*, 8 (1), doi:10.1117/1.JRS.8.084994, 2014.
- 20 Vömel, H., Selkirk, H., Miloshevich, L., Valverde-Canossa, J., Valdes, J., Kyrö, E., Kivi, R., Stolz, W., Peng, G. and Diaz, J.: Radiation dry bias of the Vaisala RS92 humidity sensor. *J. Atmos. Oceanic Tech.*, 24, 953-963, 2007a.
- Vömel, H., Barnes, J. E., Forno, R. N., Fujiwara, M., Hasebe, F., Iwasaki, S., Kivi, R., Komala, N., Kyrö, E., Leblanc, T., Morel, B., Ogino, S.-Y., Read, W. G., Ryan, S. C., Saraspriya, S., Selkirk, H., Shiotani, M., Canossa, J. V., and Whiteman, D. N.: Validation of Aura MLS water vapor by balloon-borne Cryogenic Frostpoint Hygrometer measurements. *J. Geophys. Res.*, 112, D24S37, doi:10.1029/2007D008698, 2007b.
- 25 Von Clarmann, T., Höpfner, M., Kellmann, S., Linden, A., Chauhan, S., Funke, B., Grabowski, U., Glatthor, N., Kiefer, M., Schieferdecker, T., Stiller, G. P., and Versick, S.: Retrieval of temperature, H₂O, O₃, HNO₃, CH₄, N₂O, ClONO₂ and ClO from MIPAS reduced resolution nominal mode limb emission measurements. *Atmosph. Meas. Tech.*, 2, 159-175, 2009.
- Weaver, D., Strong, K., Schneider, M., Rowe, P. M., Sioris, C., Walker, K. A., Mariani, Z., Uttal, T., McElroy, C. T., Vömel, H., Spassiani, A., and Drummond, J. R.: Intercomparison of atmospheric water vapour measurements at a Canadian High Arctic site. *Atmos. Meas. Tech.*, 10, 2851-2880, https://doi.org/10.5194/amt-10-2851-2017, 2017.
- 30

Moved (insertion) [2]

Deleted: a

Deleted: Soden, B., Jackson, D., Ramaswamy, V., Schwarzkopf, M., and Huang, X.: The radiative signature of upper tropospheric moistening. *Science*, 310, 841–844, 2005.*

Formatted: English (CAN)

Deleted: Sioris, C. E., Zou, J., McElroy, C. T., Boone, C. D., Sheese, P. E., and Bernath, P. F.: Water vapour variability in the high-latitude upper troposphere – Part 2: Impact of volcanic eruptions. *Atmos. Chem. Phys.*, 16, 2207–2219, 2016.*
Stevens, B. and Bony, S.: What are climate models missing? *Science*, 340, 1053-1054, 2013.*

Formatted: Bibliography

Formatted: English (CAN)

Moved up [2]: Soden, B., Jackson, D., Ramaswamy, V., Schwarzkopf, M., and Huang, X.: The radiative signature of upper tropospheric moistening. *Science*, 310, 841–844, 2005.*

Deleted: Trenberth, B.J., Wetherald, R.T., Stenchikov, G.L., Robock, A.: Global cooling after the eruption of Mount Pinatubo: a test of climate feedback by water vapor. *Science*, 296, 5568, 727-730, 2002.*

Trenberth, K.E., Dai, A., van der Schrier, G., Jones, D.P., Barichivich, J., Briffa, K.R., and Sheffield, J.: Global warming and changes in drought. *Nature Climate Change*, 4, 17-22, 2013.*
Trenberth, K. E.: Challenges for Observing and Modeling the Global Water Cycle, in *Remote Sensing of the Terrestrial Water Cycle*. (eds. V. Lakshmi, D. Alsdorf, M. Anderson, S. Biancamaria, M. Cosh, J. Entin, G. Huffman, W. Kustas, P. van Oevelen, T. Painter, J. Parajka, M. Rodell and C. Rüdiger), John Wiley and Sons, Inc, Hoboken, NJ, 2014.*

Formatted: English (CAN)

Formatted: Bibliography

Formatted: English (CAN), Not Superscript/ Subscript

Formatted: English (CAN)

Formatted: English (CAN), Not Superscript/ Subscript

Formatted: English (CAN)

Formatted: English (CAN), Not Superscript/ Subscript

Formatted: English (CAN)

Formatted: English (CAN), Not Superscript/ Subscript

Formatted: English (CAN)

Formatted: English (CAN), Not Superscript/ Subscript

Formatted: English (CAN)

Formatted: English (CAN), Not Superscript/ Subscript

Formatted: English (CAN)

Weigel, K., Rozanov, A., Azam, F., Bramstedt, K., Damadeo, R., Eichmann, E.-U., Gebhardt, C., Hurst, D., Kraemer, M., Lossow, S., Read, W., Spelten, N., Stiller, G. P., Walker, K. A., Weber, M., Bovensmann, H., and Burrows, J. P.: UTLS water vapour from SCIAMACHY limb measurements V3.01 (2002-2012). *Atmos. Meas. Tech.*, 9, 133-158, doi:10.5194/amt-9-133-2016, 2016.

- 5 [Worden, J., Kulawik, S. S., Shephard, M. W., Clough, S. A., Worden, H., Bowman, K., and Goldman, A.: Predicted errors of tropospheric emission spectrometer nadir retrievals from spectral window selection, *J. Geophys. Res.*, 109, D09308, doi:10.1029/2004JD004522, 2004.](#)

Formatted: Bibliography

Formatted: English (CAN)

[Wunch, D., Wennberg, P. O., Osterman, G., Fisher, B., Naylor, B., Roehl, M. C., O'Dell, C., Mandrake, L., Viatte, C., Kiel, M., Griffith, D. W. T., Deutscher, N. M., Velasco, V. A., Notholt, J., Warneke, T., Petri, C., De Maziere, M., Sha, M. K.,](#)

- 10 [Sussmann, R., Rettinger, M., Pollard, D., Robinson, J., Morino, I., Uchino, O., Hase, F., Blumenstock, T., Feist, D. G., Arnold, S. G., Strong, K., Mendonca, J., Kivi, R., Heikkinen, P., Iraci, L., Podolske, J., Hillyard, P. W., Kawakami, S., Dubey, M. K., Parker, H. A., Sepulveda, E., Garcia, O. E., Te, Y., Jeseck, P., Gunson, M. R., Crisp, D., and Eldering, A.: Comparisons of the Orbiting Carbon Observatory-2 \(OCO-2\) XCO₂ measurements with TCCON. *Atmos. Meas. Tech.*, 10, 2209–2238, <https://doi.org/10.5194/amt-10-2209-2017>, 2017.](#)

Formatted: Default Paragraph Font, English (CAN)

Formatted: English (CAN)

Formatted: Bibliography, Space Before: 6 pt

Formatted: Font: 10 pt, Not Bold, English (CAN)

15

Table 1: Summary of water vapour datasets used in this study.

Type	Satellite/ Location	Instrument	Instrument type	Measurement geometry	Retrieval version	Time range used	Number of measurements within 500 km of Eureka (Aug. 2006 onwards)	Valid altitude range
Satellite	SCISAT	ACE-FTS	Fourier transform spectrometer	limb	v3.6	Aug. 2006 - Mar. 2017	551	mid-troposphere to mesosphere
		ACE- MAESTRO	grating spectrometer	limb	v30	Aug. 2006 - Sept. 2016	388	mid-troposphere to lower stratosphere
	AURA	TES	Fourier transform spectrometer	nadir	v6	Aug. 2006 - July 2011	5,630	P > 100 hPa
		MLS	radiometer	limb	v4.20-v4.22	Aug. 2006 - Dec. 2015	108,072	P <= 316 hPa
	ENVISAT	MIPAS	Fourier transform spectrometer	limb	IMK v5 & v7	Aug. 2006 - Apr. 2012	v5: 10428; v7: 10712	12 - 50 km recommended; profiles retrieved as low as 4.5 km
		SCIAMACHY	imaging spectrometer	limb	v3.01 & v4.2	Aug. 2006 - March 2012	v3.01: 1638; v4.2: 14530	11 - 25 km
Aqua	AIRS	grating spectrometer	nadir	v6	Aug. 2006 - Dec. 2016	1,892,348	P >= 100 hPa	
Ground-based	PEARL Ridge Lab	125HR	Fourier transform spectrometer	sun-viewing	MUSICA v2015	Aug. 2006 - Sept. 2014 (excluding Polar Night)	1889 (standard); 2713 (no SZA filter)	surface to 14 km (upper altitude varies)
	Eureka Weather Station	Radiosondes	capacitance sensor	balloon-borne in situ	Bias corrected and reprocessed	Sept. 2008 - Sept. 2017	5515	surface to 8 - 15 km (upper altitude varies)

Table 2: Summary of satellite vs. 125HR comparison results. SEM refers to the standard error in the mean, i.e. $\frac{\sigma}{\sqrt{N}}$

Instrument (retrieval version)	Altitude [km]	N	mean difference ± SEM [ppmv]	σ [ppmv]	median difference [ppmv]	mean difference ± SEM [%]	σ [%]	median difference [%]
ACE-FTS (v3.6)	6.4	18	-6.3 ± 8.4	35.7	-2.0	+0.2 ± 6.8	28.9	-4.0
	8.0	46	+1.4 ± 2.6	17.6	-1.4	+7.2 ± 6.6	44.8	-1.4
	9.8	65	+0.5 ± 0.4	3.3	-0.1	+6.1 ± 3.9	31.8	-1.4
	12.0	74	+0.4 ± 0.1	0.9	+0.3	+9.7 ± 2.8	23.8	7.7
ACE-MAESTRO (v30)	6.4	18	-11.9 ± 16.7	71.0	-20.9	-6.7 ± 19.2	81.5	-30.0
	8.0	23	-5.6 ± 6.5	31.4	-9.0	-6.1 ± 18.7	89.6	-29.5
	9.8	25	-2.0 ± 1.5	7.3	-3.4	-10.8 ± 14.7	73.6	-28.8
	12.0	26	-0.6 ± 0.3	1.7	-0.8	-11.4 ± 9.5	48.5	-22.3
AIRS (v6)	3.0	434	-92.8 ± 17.0	354.2	-61.4	-3.8 ± 1.6	32.7	-9.2
	6.4	881	-9.7 ± 3.5	105.1	-15.5	-1.6 ± 1.5	75.9	-10.5
	8.0	1448	-11.1 ± 1.5	56.6	-6.8	-2.9 ± 1.0	38.8	-8.7
	9.8	2517	-2.7 ± 0.2	11.4	-1.6	-3.5 ± 0.6	31.5	-8.6
	12.0	2798	-0.1 ± 0.0	1.8	-0.0	+1.8 ± 0.4	23.4	-0.8
MIPAS (IMK v5)	6.4	24	-38.2 ± 11.9	58.3	-24.4	-22.4 ± 7.8	38.1	-25.5
	8.0	93	-15.8 ± 3.0	29.2	-9.3	-18.7 ± 3.2	30.4	-22.0
	9.8	604	-3.6 ± 0.4	9.5	-2.3	-10.1 ± 1.1	27.7	-14.2
	12.0	897	-0.3 ± 0.0	1.7	-0.2	-1.4 ± 0.7	21.9	-4.2
MIPAS (IMK v7)	6.4	32	-46.9 ± 11.2	63.4	-25.3	-25.3 ± 5.9	33.5	-24.4
	8.0	96	-17.0 ± 3.1	30.0	-9.8	-20.1 ± 3.1	29.7	-22.5
	9.8	634	-3.8 ± 0.4	9.7	-2.3	-10.3 ± 1.1	27.5	-13.9
MLS (v4.2)	8.0	13	-15.9 ± 2.8	10.2	-12.4	-33.1 ± 3.6	12.9	-34.2
	9.8	2443	-4.8 ± 0.2	11.9	-3.2	-12.5 ± 0.6	29.7	-17.1
	12.0	2445	-0.4 ± 0.0	1.9	-0.4	-4.6 ± 0.5	23.0	-7.7
SCIAMACHY (IUP v3.01)	12.0	201	-0.1 ± 0.1	1.9	-0.4	-1.8 ± 1.9	26.7	-7.8
SCIAMACHY (IUP v4.2)	12.0	1506	-0.4 ± 0.0	1.5	-0.4	-5.7 ± 0.5	21.2	-7.5
TES (v6)	3.0	361	-168.2 ± 45.2	859.6	-112.6	-1.0 ± 2.3	43.4	-9.0
	6.4	361	+66.9 ± 15.6	296.7	+19.5	+23.8 ± 3.9	75.1	+6.6
	8.0	361	+30.2 ± 5.8	110.8	+13.0	+27.6 ± 4.0	76.2	+11.9
	9.8	361	+6.4 ± 1.0	19.4	+4.6	+26.0 ± 3.2	60.4	+16.2
	12.0	361	+1.5 ± 0.2	3.0	+1.3	+23.5 ± 2.1	39.2	+19.7

Instrument (retrieval version)	Altitude [km]	N	mean d ± SEM
ACE-FTS (v3.6)	6.4	18	-6.3
	8.0	46	+1.4
	9.8	65	+0.5
	12.0	74	+0.4
ACE-MAESTRO (v30)	6.4	18	-11.9
	8.0	23	-5.6
	9.8	25	-2.0
	12.0	26	-0.6
AIRS (v6)	3.0	434	-92.8
	6.4	881	-9.7
	8.0	1448	-11.1
	9.8	2517	-2.7
	12.0	2798	-0.1
MIPAS (IMK v5)	6.4	24	-38.2
	8.0	93	-15.8
	9.8	604	-3.6
	12.0	897	-0.3
MIPAS (IMK v7)	6.4	32	-46.9
	8.0	96	-17.0
	9.8	634	-3.8
MLS (v4.2)	8.0	13	-15.9
	9.8	2443	-4.8
	12.0	2445	-0.4
SCIAMACHY (IUP v3.01)	12.0	201	-0.1
	12.0	1506	-0.4
TES (v6)	3.0	361	-168.2
	6.4	361	+66.9
	8.0	361	+30.2
	9.8	361	+6.4
	12.0	361	+1.5

Deleted:

Table 3: Summary of satellite vs. radiosonde comparison results. SEM refers to the standard error in the mean, i.e. $\frac{\sigma}{\sqrt{N}}$.

Instrument (retrieval version)	Altitude [km]	N	mean difference \pm SEM [ppmv]	σ [ppmv]	median difference [ppmv]	mean difference \pm SEM [%]	σ [%]	median difference [%]
ACE-FTS (v3.6)	6.0	57	-13.3 \pm 12.1	91.5	+13.6	+22.7 \pm 9.2	69.1	+10.3
	8.0	92	-1.8 \pm 3.6	34.7	+1.8	-1.8 \pm 7.2	69.5	+4.7
	10.0	51	-5.4 \pm 2.0	14.0	-0.4	-9.1 \pm 6.9	49.4	-3.2
	12.0	19	+1.2 \pm 0.4	1.6	+1.7	+32.0 \pm 6.6	28.6	+36.2
ACE-MAESTRO (v30)	6.0	54	-62.4 \pm 36.8	270.7	-29.3	+27.0 \pm 24.8	181.9	-25.6
	8.0	90	-16.3 \pm 8.7	82.3	-11.7	-7.6 \pm 9.4	89.5	-34.4
	10.0	41	-2.6 \pm 3.2	20.3	-1.2	-5.9 \pm 10.9	89.5	-12.8
AIRS (v6)	3.0	584	-27.5 \pm 16.8	407.0	-38.3	+5.4 \pm 1.9	46.2	-4.8
	6.0	1423	-15.6 \pm 2.5	93.7	-3.7	+3.0 \pm 1.0	39.0	-3.4
	8.0	2127	+3.1 \pm 0.9	42.3	+3.8	+12.7 \pm 0.7	34.2	+8.4
	10.0	868	-11.2 \pm 0.6	18.6	-4.2	-12.4 \pm 0.9	27.5	-14.0
	12.0	50	-2.0 \pm 1.2	8.3	+0.4	+5.2 \pm 4.1	28.8	+8.9
MLS (v4.2)	8.0	12	-34.1 \pm 28.3	98.2	-6.6	-25.6 \pm 14.8	51.1	-35.6
	10.0	447	-5.1 \pm 1.2	25.0	-3.5	-25.6 \pm 1.4	29.4	-28.0
	12.0	42	-2.4 \pm 1.2	7.7	+0.0	-4.9 \pm 4.0	26.1	+0.1

Instrument (retrieval version)	Altitude [km]	N	mean difference SEM [ppmv]
ACE-FTS (v3.6)	6.0	57	-13.3
	8.0	92	-1.8
	10.0	51	-5.4
	12.0	19	+1.2
ACE-MAESTRO (v30)	6.0	54	-62.4
	8.0	90	-16.3
	10.0	41	-2.6
	12.0	12	-1.3
AIRS (v6)	3.0	584	-27.5
	6.0	1423	-15.6
	8.0	2127	+3.1
	10.0	868	-11.2
	12.0	50	-2.0
MLS (v4.2)	8.0	12	-34.1
	10.0	447	-5.1
	12.0	42	-2.4

Deleted:

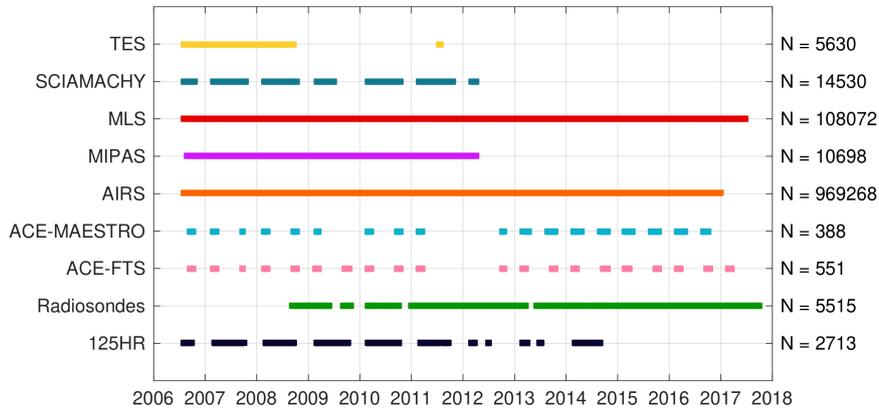


Figure 1: Temporal range of datasets used in this study. N is the number of measurements.

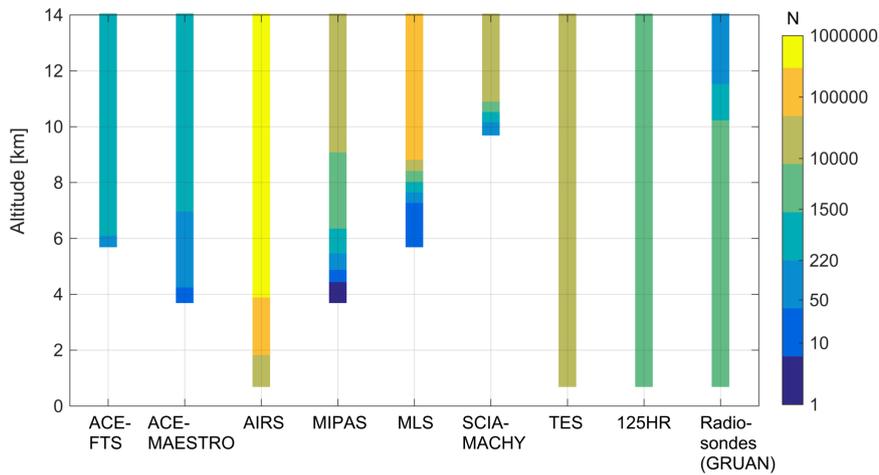


Figure 2: Vertical range of datasets used in this study. Colour range showing the number of profiles at each altitude level shows the $\log(N)$.

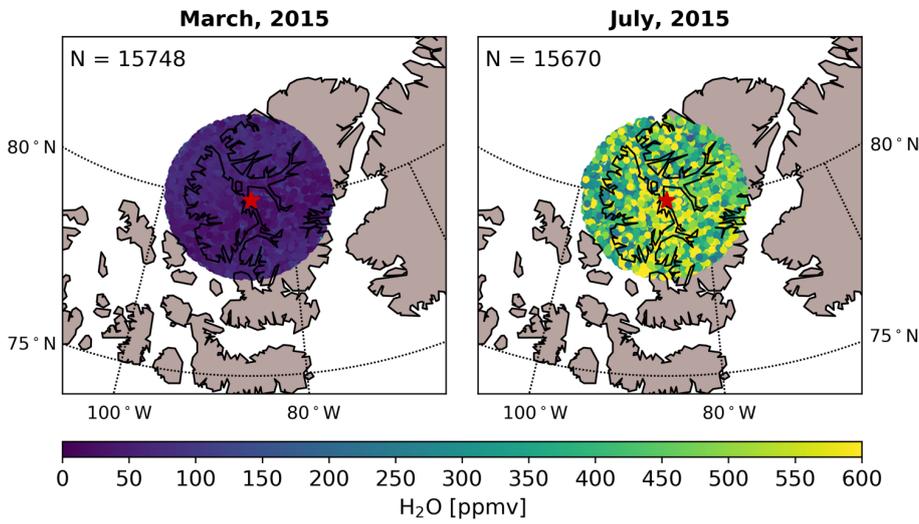
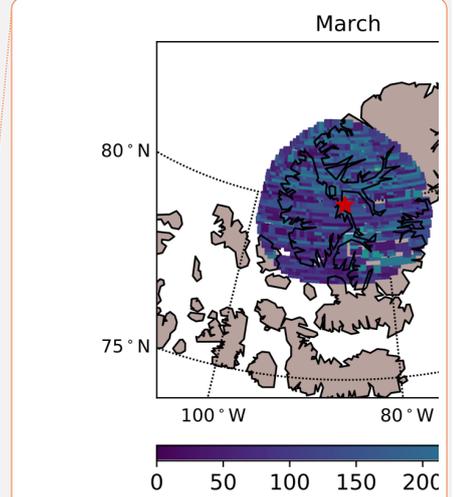


Figure 3: AIRS water vapour abundances at 400 hPa near Eureka (indicated by the red star) in March and July, 2015.



Deleted:

Deleted: Average

Deleted: , calculated by taking the mean value of all AIRS measurements between 2006 and 2016 within 50 x 50 km grid boxes...

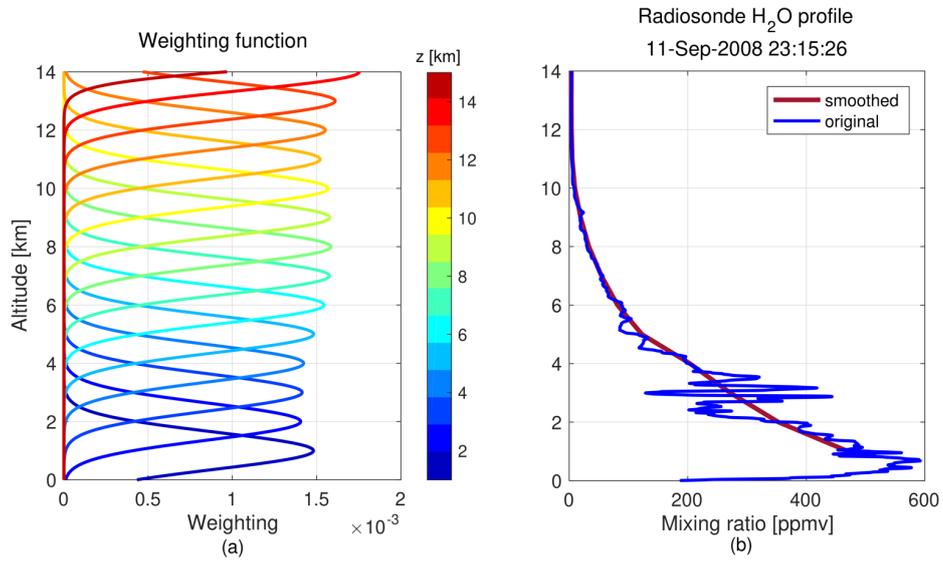


Figure 4: (a) shows an example of weighting functions used to smooth the radiosonde profiles to ACE-FTS vertical resolution. (b) shows the corresponding radiosonde profile, both as measured (blue line) and after smoothing (maroon line) with the weighting function shown in (a).

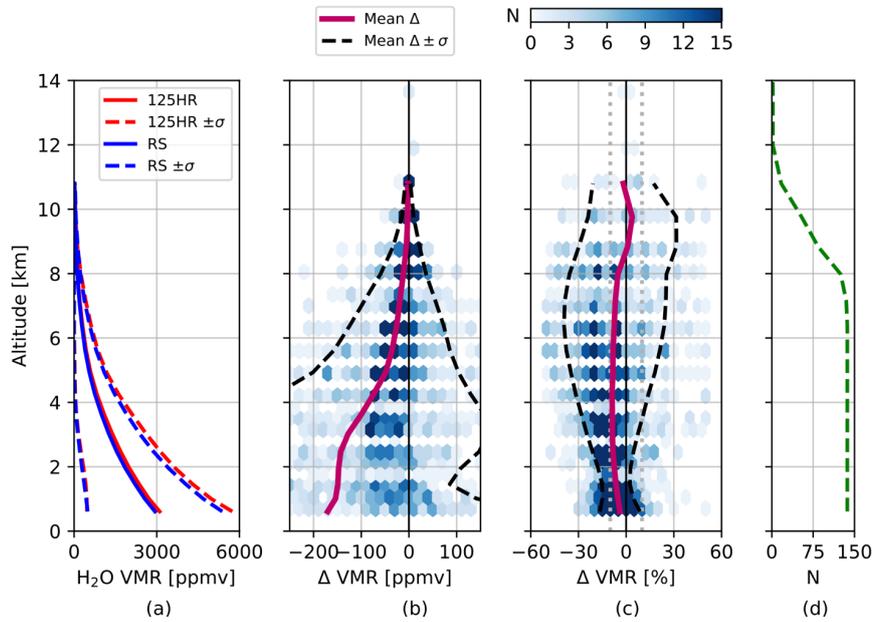


Figure 5: Comparison between Eureka (GRUAN-processed) radiosonde and PEARL 125HR water vapour VMR. (a) Mean profiles (solid lines) \pm the standard deviation (dashed lines). (b) Mean VMR difference (where X = radiosonde and Y = 125HR), using Equation 4. (c) Mean percent difference, using Equation 5. Grey dotted lines show $\pm 10\%$. In (b) and (c), the colour shading shows the number (N) of differences in each hexagon. (d) Number of coincident profile pairs at each altitude level. Note that comparisons are shown up to a maximum altitude of 11 km because the number of coincident pairs above that level do not meet the $N > 15$ threshold.

Deleted: 5.1
Deleted: 5.2

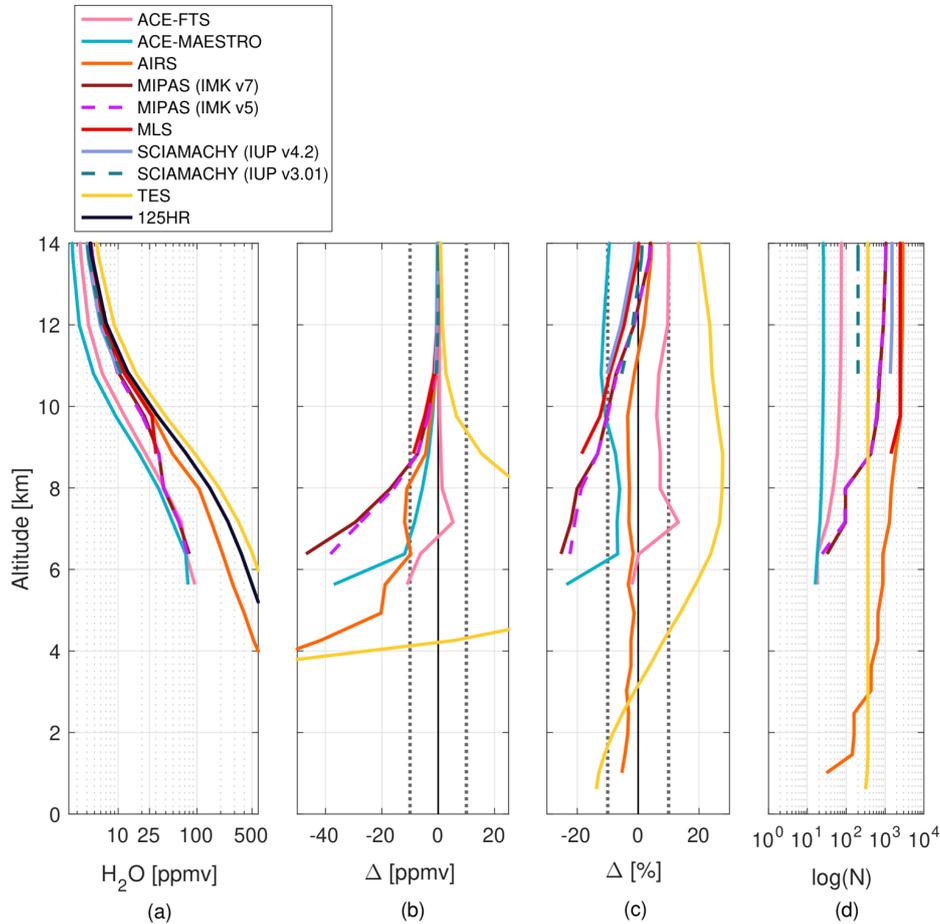


Figure 6: Summary of differences between satellite measurements and PEARL 125HR. (a) The mean of profiles used in the comparison. (b) The mean VMR difference between the satellite profiles and the 125HR profiles, using Equation 4. (c) The mean percent difference between the satellite profiles and the 125HR profiles, using Equation 5. (d) The number of coincident profile pairs contributing to the comparison at each altitude level. Grey dotted lines in (b) and (c) show ± 10 ppmv and $\pm 10\%$, respectively.

Deleted: 5.1

Deleted: 5.2

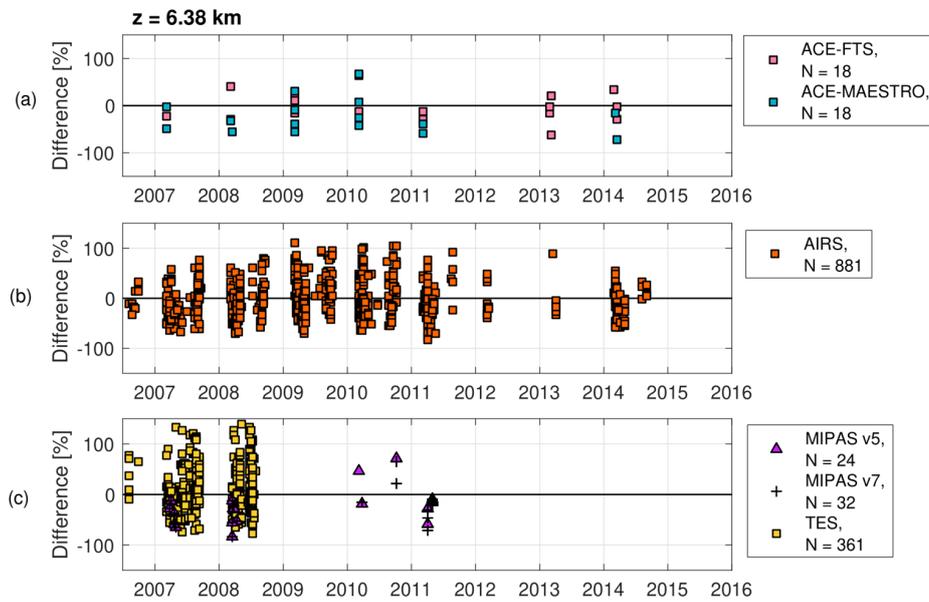


Figure 7: Time series of percent differences between satellite and 125HR water vapour measurements at 6.4 km altitude for (a) ACE-FTS and ACE-MAESTRO, (b) AIRS, and (c) MIPAS and TES. In each case, the differences follow Equation 3, where the satellite is X and the PEARL 125HR is Y .

Deleted: .

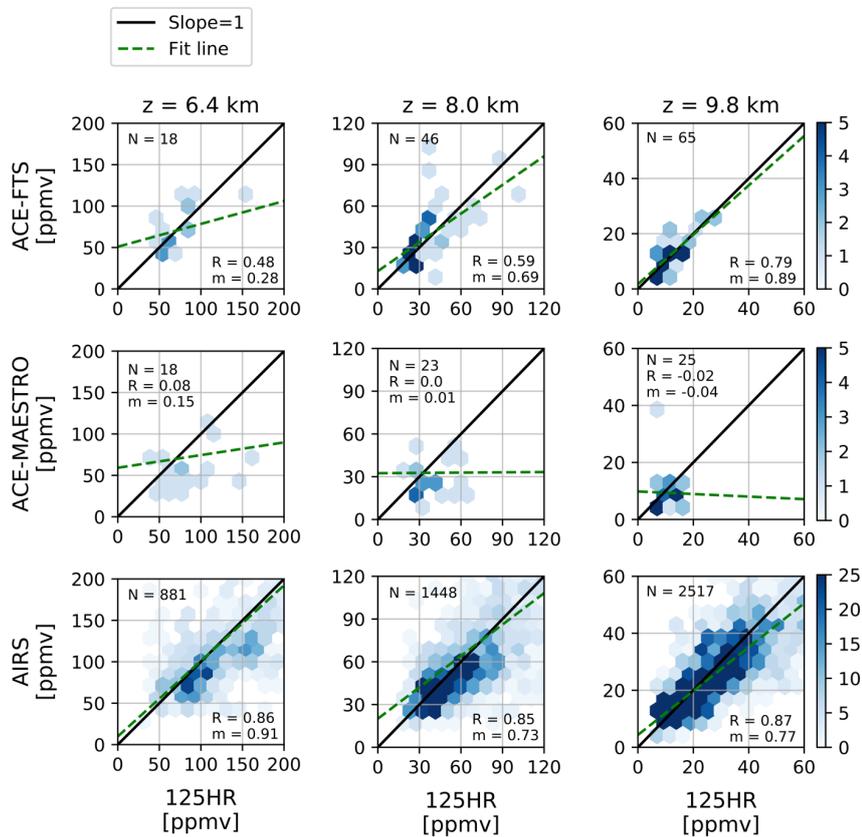


Figure 8: Correlation plots for the ACE-FTS, ACE-MAESTRO, and AIRS satellite measurements vs. 125HR. The number of points in a given hexagon is color-coded to show the density of the points. The scale at each end of a row shows the colour map used for that row. Solid black lines are 1:1 reference lines (i.e. slope = 1); green dashed lines are lines of linear best fit. N is the number of coincident measurements for comparisons between the instruments at that altitude. R is the correlation coefficient. m is the slope of the best fit line.

5

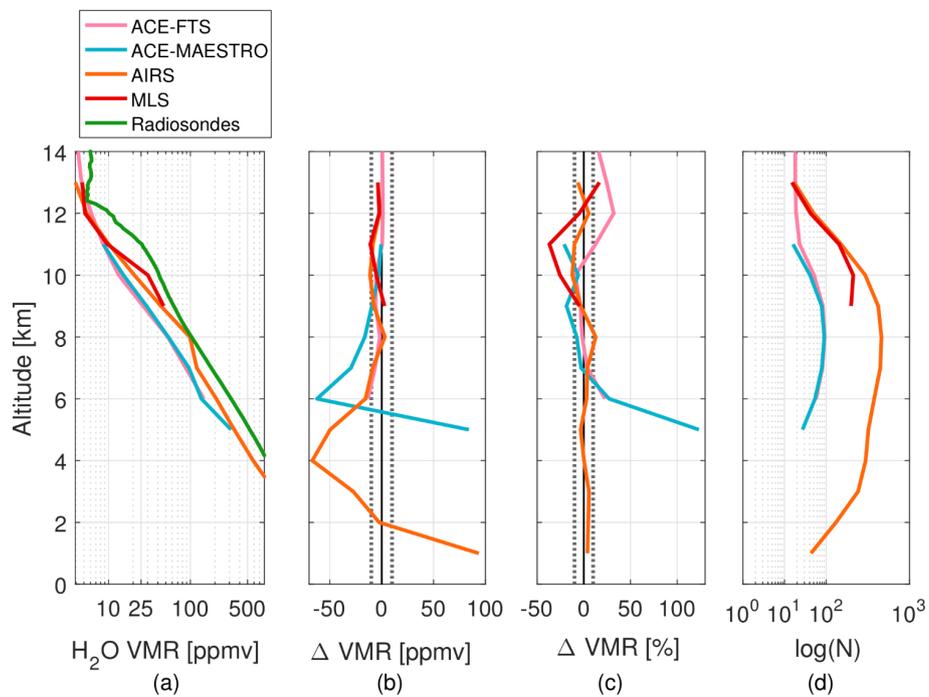


Figure 9: Same as Fig. 6, but a summary of differences between satellite measurements and Eureka radiosondes. A version of this figure with only the ACE-FTS and ACE-MAESTRO is available in the supplementary materials as Fig. S2.

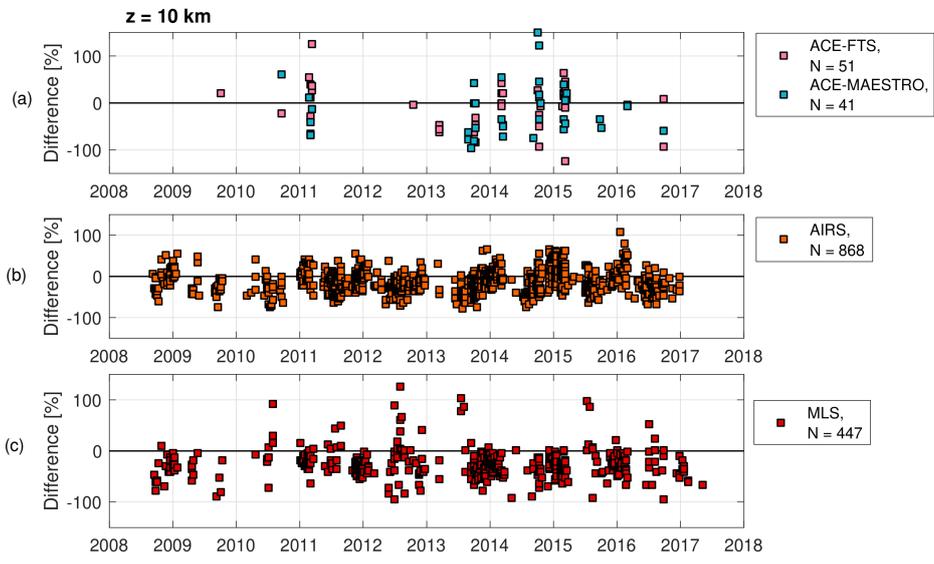


Figure 10: Time series of percent differences between satellite measurements and the Eureka radiosondes at 10 km altitude for (a) ACE-FTS and ACE-MAESTRO, (b) AIRS, and (c) MLS. In each case, the differences follow Equation 3, where the satellite is X and the Eureka radiosondes is Y .

Deleted: .

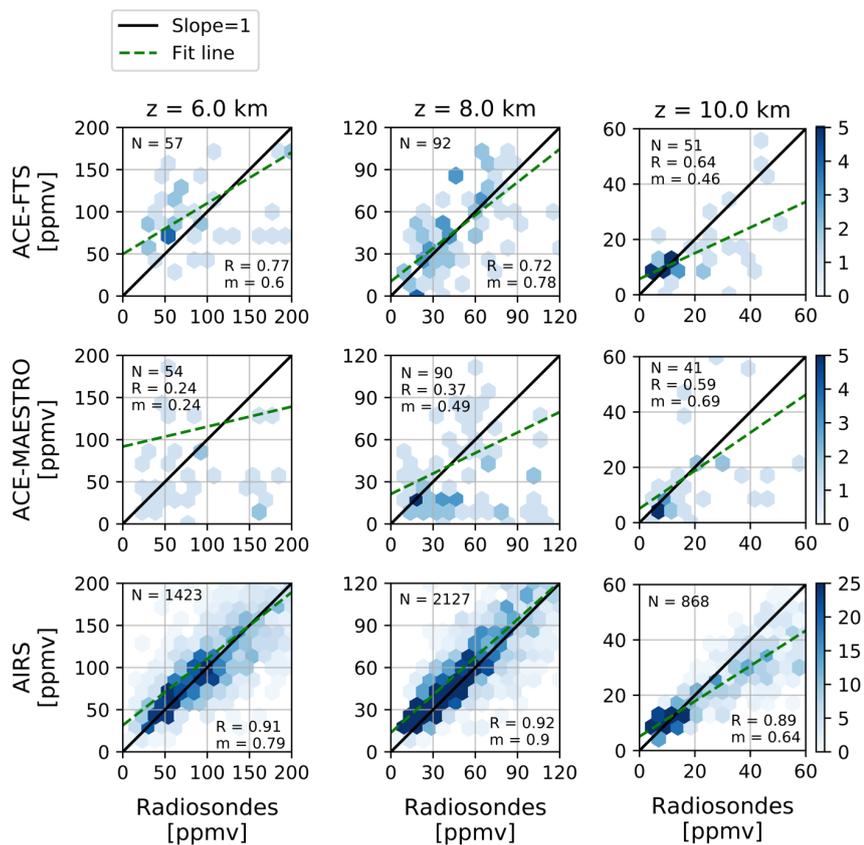


Figure 11: Same as Fig. 8, but of correlation plots for the ACE-FTS, ACE-MAESTRO, and AIRS satellite measurements vs. the Eureka radiosondes.

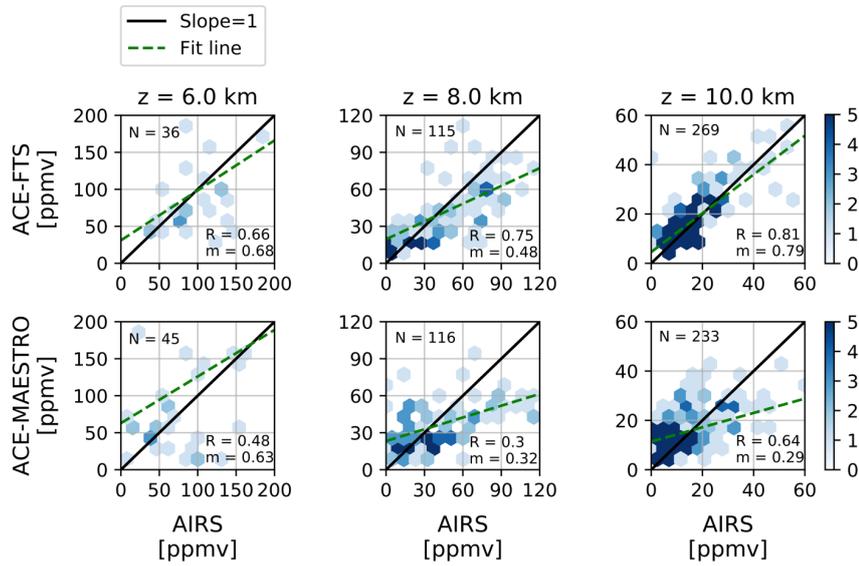


Figure 12: Same as Fig. 10, but of correlation plots for the ACE-FTS and ACE-MAESTRO vs. the AIRS satellite measurements.