Overview:

The primary goal of this research is to compare remotely sensed upper troposphere lower stratospheric water vapour retrievals (UTLS-WV) from the ACE-FTS (ACEF) and ACE-MAESTRO (ACEM) satellite based instruments with that of two ground based water vapour measurement datasets (MIR-FTIR and RH% radiosondes) located at Eureka, Nunavat in the Canadian high arctic. The high artic is a data sparse region and this study provides new additional dataset comparisons. The aim is to see if the ACEF & ACEM datasets are sufficiently accurate (defined as meeting the GCOS 5% accuracy goal for profile measurements), relative to the so called reference ground-based measurements. The secondary aim is also to compare other satellite based UTLS-WV measurements (AIRS, MIPAS, MLS, SCIAMACHY and TES) to the ground based WV datasets and also against ACEF & ACEM. The satellite and MIR-FTIR UTLS-WV datasets are publically available and have been presented in previous published peer reviewed literature and also have been used extensively in other studies. Additionally, in this comparison exercise an extended MIPAS dataset is used, using data from outside the recommended altitude range limits to increase the number of coincidences. Importantly, the RH% radiosonde UTLS WV dataset has been processed using GRUAN procedures. Unfortunately, the current GRUAN dataset is not publically available. There are a total of 9 datasets spanning the time period 2008-2015, with each dataset covering different time periods and altitude ranges. The altitude range comparisons are conducted over are not consistent but with the majority of dataset profile comparisons conducted over the range ~6 to ~12km in ~1km steps. Ubiquitous comparison algorithms are employed based upon spatial and temporal coincidence criteria. Differences relative to the ground based datasets are reported (along with ad-hoc inter-satellite comparisons). Comparisons between the two ground based datasets are also performed. The main findings are that ACEF & ACEM differences relative to the ground based reference measurements are outside the 5% GCOS accuracy goal, but have better comparison statistics compared to other satellites (except for AIRS). The authors conclude AIRS measurements are sufficiently accurate and should be used for future UTLS-WV studies at Eureka (and surrounding region). The authors recommend FPH-WV measurements at Eureka for future ground based-satellite comparisons activities.

The novelty of this study is that is it a comprehensive comparison of UTLS-WV from multiple remote sensing instruments, across multiple platforms in the high artic, and is the first time the Eureka GRUAN processed RH% radiosonde dataset has been used in satellite comparisons. This study will be a welcome addition to literature, especially for the current SPARC WAVAS-II activity. It offers a comparison template which could be extended to other sites with ground-based FTIR or RH% radiosonde measurements. The manuscript meets the scope and requirements of the AMT journal. The manuscript is logically structured. Overall it is well referenced and the writing style is fluent and easily understood.

Improvements can be made. In its current form I do not recommend publication without revision due to a number of issues which are listed below. Such issues need to be addressed: either fixed or with sufficient rebuttal.

General (major) comments:

G1/ Possible erroneous values in tables 2 and 3.

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

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

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

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

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

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

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

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

G3/ Context

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

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

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

G4/ Inclusion of measurement uncertainties.

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

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

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

Minor, but related points:

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

-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).

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

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

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

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

G6/ Seasonal cycle and seasonality in the TPH

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

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

More specific comments (no particular order):

S1/ References and referencing:

There is an instance where a paper is referenced in the manuscript, but not in the reference list (Khosrawi, 2018) and conversely there are papers in the reference list, Kurylo, 1991, Sioris, 2016b, &

Stevens, 2013 that are not referenced in the manuscript. Can the authors please recheck the manuscript and reference list to make sure all cross-referencing is correct.

S2/ GCOS and WMO are used interchangeably.

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

S3/ Equation 7.

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

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

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

S5/ Equations 2 and 3.

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

S6/ Equation 1.

For completeness, the term $e_s(T)$ should be $e_s(T(z))$.

S7/ Sigma (σ) values in section 3.2.4

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

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

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

S8/ Quantifying small dry bias.

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

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

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

S10/ Tables 2 and 3

For completeness could SEM be explained in the table captions?

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

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

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

S12/ Figure 5.

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

S13/ Figure 5, part 2…

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

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?

S15/ Displaying instrument vertical resolution.

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

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?

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.

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?