Author response to reviewer's comments on

"Intercomparison of atmospheric water vapour measurements in the Canadian high Arctic" by Weaver et al.

**Reply to Anonymous Referee #1**

*We would like to thank the reviewer for their detailed and helpful comments. Their suggestions have helped improve the manuscript's clarity. Replies are given below in blue italics.*

The manuscripts deals with ground-based measurements of water vapour at a high latitude (80N) site. Some new data are presented and a broad comparison is made. The work is of special importance due to the sensitivity of the Arctic region to climate change and the crucial role of water vapour as a main feedback mechanism, in combination with the fact that there are few permanent research sites inside the Arctic region. The main strengths of the manuscript are the high number of observation techniques considered, the introduction of new AERI retrievals and a clear demonstration that the MUSICA retrievals have deteriorated for dry sites.

The manuscript also treats the difference at the site between standard radiosonde data and the GRUAN version (not clear if these are new results or not).

*Text has been added to make clear that the Eureka GRUAN data are new (i.e. to Section 2.4.1), and have not been previously presented in published literature. The results of the comparisons presented in this work are indeed new.*

These topics fit well with the scope of AMT. I have no basic doubts about the amount of work, or its quality, that has gone into manuscript, but the results lack generality and the presentation should be improved.

The only general results I can identify is the conclusion around the MUSICA retrievals. Other contributions are either (1) quite specific for the site or (2) are not described in sufficient detail. The basic results are of course obtained at the site (and have intrinsic value), but there is very little discussion on how the results agree with other comparisons. For example, are these results consistent with the ones presented by Buehler et al (2012) and Palm et al (2010) for other high latitude sites? That is, to solve point (1) the results should be properly compared to what has been found elsewhere.

*We agree it is important to articulate how the results of a study compare with those of others. Both the Buehler et al. (2012) and Palm et al. (2010) studies are cited in the introduction for this reason. However, there are significant limitations to doing so in this case because of the differences in instrumentation. Comments about each study are offered in turn. Text has been added to the discussion section to make connections between these other studies and the results of ours in a new sub-section (4.6).*

*The Palm et al. (2010) study examines data also taken at a high latitude measurement site. However, its focus is on comparing satellite measurements to Ny Ålesund FTIR and radiosonde measurements. Limited similarities in instrumentation exist between their study and ours. While they also use a microwave radiometer, their instrument observes a different spectral band (142 GHz), as it is optimized for ozone measurements, and its water vapour measurements are not directly compared to the Ny Ålesund FTIR and radiosondes. For these reasons, connections between the results of the Palm et al. study and ours are limited.*

*The Buehler et al. (2012) study examined water vapour measurements at Kiruna, Sweden from an FTIR, radiosondes, microwave radiometer, GPS, ERA-reanalysis, and the Advanced Microwave Sounding Unit-B satellite. Comparisons using the first three of those instruments have relevance to this study. However, the Buehler et al. study does not present the results of comparisons using any combination of instruments used in our study. Further, they do not fully describe the details of their FTIR retrieval (although stated characteristics have similarities to the MUSICA retrieval set up, e.g. InSb detector measurements and the use of PROFFIT software). This limits our ability to discuss the comparability of results seen in the current study. Nonetheless, they find that their FTIR results have a dry bias with respect to the GPS at low-PWV and wet bias at high-PWV, and state that these results are 'roughly similar to the radiosonde data' comparisons to the GPS.*

*Table 5 in Buehler et al. (2012) summarizes the instruments examined in water vapour comparison studies across their literature review. Notably, this table shows very limited overlapping instrumentation with the current study.*

*In addition, similarities and differences between our results and those of a sub-tropical study, Schneider et al. (2012), are noted in the discussion section. While the Schneider et al. study examines measurements in a very different geographical context (e.g. sub-tropical), the instruments examined (FTIR, sun photometer, GPS, and radiosonde) are similar to the combination examined at Eureka. They observe the sun photometer to underestimate PWV relative to the FTIR and RS (similar to the results of our study); the FTIR, using a retrieval similar to the MUSICA v2012 retrieval discussed in this work, is shown to have a small (3.3%) underestimation of the water vapour column with respect to the radiosondes.*

Regarding point (2) I am mainly thinking about the new AERI retrievals. I simply don't think it is possible to recreate the implemented algorithm from the description provided.

*To address this concern we have considerably expanded the description of the water vapour retrieval from the AERI. To avoid breaking up the flow of the main text, we have moved this expanded text to an appendix. It includes considerably more detail about the retrieval, including a list of retrieval steps with a description of each step, six new equations, and seven new references. Furthermore, we have improved consistency between the retrieval from the E-AERI and the P-AERI. We believe these greatly strengthen the description, making it possible to replicate, and we thank the reviewer for pointing this out. Specific changes include:*

*Page 8, lines 18-19: (now moved to appendix): To improve consistency, we no longer average any E-AERI spectra; we have therefore removed the sentence, "In this work, E-AERI (when located at 0PAL) spectra are averaged over an hour (typically 8 spectra) to reduce noise."*

*Page 9, line 23, after "Rowe et al. (2008)" we now provide more information in the appendix, adding the following:*

*"Errors are similar in spectral shape to the case shown in Rowe et al. (2008; Fig. 3) and vary seasonally with similar magnitudes in winter (maximum of about 0.15 RU), to roughly double in summer (maximum of about 0.30 RU). The noise level for the E-AERI is lower than 0.4 RU (Mariani et al. 2013). For the P-AERI, the noise level near 900 $cm^{-1}$, after noise-filtering, is estimated to be ~0.070 RU in winter, increasing to ~0.15 RU in summer. The noise level in spectral regions used to retrieve water vapor, where the downwelling radiance is considerably stronger, may be up to twice as large as in the window region (Turner et al. 2006)."*

*Page 10, lines 1-2 (now moved to appendix: added radiance uncertainty estimates, replacing "uncertainties in units of …" with "combined errors, excluding noise, that vary spectrally and seasonally between 0.07 and ~2.5 RU.*

*Text was added to the end of the appendix regarding quality control and uncertainties:*

*"Cases with minimum weighted-mean-differences greater than expected uncertainties are removed as part of quality control (here, a threshold of 2 RU removed 150 cases). The frequency-dependent uncertainties and the weights are also used to determine the uncertainty in the retrieved scale factor following standard uncertainty analysis, resulting in uncertainties in PWV of 3% to 11% for winter to summer cases."*

For example, the "changes" discussed around line 4 of page 10, are these applied on the total column, or for layers? I assume the later, but there is no information on what layers that are used. (If the total column is perturbed, the description makes a simple thing very complex.)

*Yes, the total column is perturbed. It is necessary to describe this in terms of a perturbation to the water vapour profile because the two are related but distinct: two different water vapour profiles can result in the same total column. To clarify without omitting this distinction, we have modified the description on page 10 lines 3-8 (now moved to appendix) to phrase it simply as suggested by the reviewer, followed by a longer, more complex description that allows replication. In addition, we have made changes to better frame the explanation.*

*To identify the best frequencies for retrieving water vapour, the expected uncertainty in the total column water vapour that would be retrieved at each frequency is determined. The uncertainty is calculated as follows. First, the*

*sensitivity of the downwelling radiance to changes in the total column of water vapour is determined by calculating the change in simulated radiance for a 5% percent change in the first-guess water vapour profile (i.e. a 5% change in all layers). These values are then used to determine the percent change in water vapour per change in radiance. Next, the percent change in water vapour per change in radiance is multiplied by the combined uncertainty in the measured and simulated downwelling radiances, giving the expected uncertainty in water vapour at each frequency. Note that this uncertainty has units of percentage of the first-guess water-vapour profile, or, equivalently, percentage of the first-guess total column of water-vapour. Finally, the 100 frequencies where this uncertainty is the lowest are selected for monthly … between 1100 and 1400 $cm^{-1}$.*

I can not really understand the treatment of the self-broadening part (end of page 10) of the water vapour optical depth.

*This is a good point. A description of the water vapour continuum is beyond the scope of this work, so we have not included it. However, we added a reference to work that describes it. The important point here is the self-broadened continuum scales with the square of the water vapour concentration and thus must be handled separately. This has been clarified. Furthermore, as all of this is now in the appendix, it does not disrupt the flow of the main text.*

I would also like to see a demonstration that the cloud optical thickness can be as high as 0.6 without having a significant impact on the water vapour retrieval.

*This is an important point. To provide consistency between the P-AERI and the E-AERI datasets, we have revised our results to use a radiance threshold rather than a cloud optical depth threshold for both instruments, as described in the following paragraph, which replaces the paragraph on page 10, lines 12-14, and is now in the appendix.*

*To identify clear-sky time periods, we compare observed to simulated clear-sky radiances between 898.0 and 905.7 $cm^{-1}$, where emission from trace gases is extremely weak and therefore any cloud emission should be evident. To choose a threshold for the comparison, we use as a guide cases where the simulated radiance is lower than the observed radiance, assuming such cases are cloud-free. For the P-AERI, the standard deviation of the difference between observed and simulated radiances for such cases is 0.4 RU, while for the E-AERI, it is 0.8 RU (perhaps due in part to the larger noise of the E-AERI). Thus these values serve as radiance thresholds for clear-sky: if the observation is within 0.4 (P-AERI) or 0.8 (E-AERI) $mW*(m^2\ sr\ cm^{-1})^{-1}$ between 898.0 and 905.7 $cm^{-1}$ the scene is assumed to be cloud-free. If these biases are due to clouds, maximum errors in retrieved water vapour would be 0.6 to 3% for P-AERI (summer to winter), and 5-7% for E-AERI (summer to winter). However, given the magnitude of the negative errors, these differences are more likely due to differences in the water vapour itself as well as sources of error that have already been accounted for.*

That is, Section 3.1.2 should be revised to ensure that the retrieval algorithm can be understood and duplicated by others. However, I don't know if this can be achieved without making the manuscript too long, and an additional article must be considered.

*We believe that with the additional detail given in the appendix and the referenced literature, the method is now reproducible.*

Some general comments around the presentation of the results:

Obviously incorrect/bad comparisons should not be included. This is not the case now and the reader is easily confused by a high share of irrelevant numbers. I am here thinking about comparing columns with different start altitudes (i.e. Ridge lab vs. 0PAL), and results coming from the microwave radiometer after fall of 2010 (when its calibration started to fail).

*We disagree that incorrect comparisons have been presented. However, major revisions have been made to address this comment.*

*There is value in comparing nearby instruments at different altitudes. The results of such comparisons can inform other work. For example, satellite measurements and model outputs typically cover a wider area than the distance between the Ridge Lab and 0PAL (15 km apart). The differences observed between instruments at the two altitudes*

*can inform the validation and use of satellite measurements and model results that cover the highly diverse terrain surrounding Eureka.*

*Instead of showing the direct comparisons of instruments at different sites, the following major changes have been made:*

- *Direct comparisons between instruments at different sites have been removed.*
- *Comparisons between instruments at the Ridge Lab and the radiosondes (standard dataset and GRUAN) now use the radiosonde profile integrated only down to the Ridge Lab altitude (i.e. 610 m) to put both measurements on an equivalent altitude basis.*
- *Three datasets have been calculated for the partial column between the sites, using measurements from the radiosondes, GRUAN, and sun photometers (the difference between near-simultaneous SPM measurements at both sites). A new section discusses the partial column between the sites.*
- *The table and correlation plots summarizing comparison results have been split up into three: one for comparisons among Ridge Lab instruments, one for comparisons among 0PAL instruments, and one for comparisons between measurements of the partial column between 0PAL and the Ridge Lab.*

*Showing the partial column comparisons demonstrates the substantial impact of the varied terrain around Eureka and PEARL on the observed $H_2O$ column.*

*In regards to the microwave radiometer comparisons, we believe showing disagreement between measurements has value, particularly in the context of a study assessing and validating measurement results. The complete microwave radiometer dataset is available for scientific analysis. One aim of the study was to evaluate whether the techniques available perform well in a difficult high Arctic circumstance. The primary reason for the MWR not having regular calibration and maintenance is the difficulty of accessing the remote location. That said, the MWR comparison results now focus on the four-year period where there appears to be well-calibrated data, from Aug. 2006 to June, 2010. The full timeseries of comparison results with the radiosondes are included (in the new Fig. 15) to both inform the reader of the difficulties in using this dataset and to justify the decision to use only this four-year subset of the available data.*

In many places the sign of found difference is unclear. Referring to Eq. 1, it is not clear what measurement that has been taken as X and Y. Table 3 is one example. It says "comparison of 125HR", but does this mean that 125HR is X or Y? I get the impression that the authors tend to select the mentioned instrument as X, but I can not find any general statement on this point. I would also say that the standard approach is to put the identified reference as Y, and recommend to use this scheme. (Also note that all differences are positive in Table 3. Is that really correct?) Another example is that the text in Sec 3.3.1 does not fit with a positive difference (i.e. should it not be -4.6%). Please check carefully the complete manuscript to avoid all uncertainty about the sign of differences.

*We agree it is important to be clear about the sign of the differences, and the application of the equation to each individual case. The caption of the tables and figures have been revised to make it clear which instrument is X and which is Y.*

Another problem can be illustrated by line 24 on page 15: "the mean percent difference is 11.4+-13.2%". As written like this, the value 13.2% must be taken as the uncertainty of estimated mean difference. That is, no significant difference has been found in this case! However, I guess that 13.2% refers to something else. Based on Table 3 I assume that this is the standard deviation of the sample variation (that is of interest if the measurement precision are investigated, but not for the mean difference). Further, are 1 or 2 standard variations given? Please revise the manuscript on these points.

*We agree this can be misleading. In your example, 13.2% was the one-sigma to the mean. This is commonly done in similar studies, such as Schneider et al. (2010) and Buehler et al. (2012).*

*The text has been revised. Where we describe the agreement as A ± B, B is now the standard error in the mean rather than the standard deviation. This has been mentioned explicitly in the text in Section 3.1 The standard deviation continues to be used in figures because it quantifies the spread in the difference values.*

The manuscript contains a high number of figures. Several figures can be removed without any important loss of information, such as Fig. 4, 7a, 11 and 15. The results displayed are either not important for the discussion, or are already covered by values in tables.

*While it is true that some figures in the paper represent values in the tables, they are offered to aid the reader so that they may more easily visualize the agreement and also illustrate how the agreement changes over the timeseries.*

*The following figures have been removed:*

*Figure 4. Its intention had been to illustrate the variability of the water vapour column from year to year, which is substantial and not otherwise shown in the manuscript.*

*Figure 7 (a) has been removed, as it duplicates the information in Fig. 7 (b).*

*Figure 11 has been kept because it provides a useful reference for the validation of the new AERI $H_2O$ retrieval product.*

*Figure 12, which had shown correlation plots of AERI vs. co-located instruments, has been removed. The new Fig. 11 and 12, which replace the original Fig. 13 (large set of all correlation plots) show all correlation plots from the Ridge Lab and 0PAL, respectively.*

*Figures 14, 15, and 16 have been removed because the profile comparisons have been removed to satisfy the comments of reviewer 3.*

*Figure 17, showing the comparison of SPM at Ridge Lab and 0PAL have been removed, as comparisons between instruments located at different altitudes are no longer presented.*

*Figure 18, showing the timeseries of differences between the current and previous MUSICA retrieval versions.*

*In addition, Figure 20 (b) has been removed, as it offers similar information as Fig. 20 (a).*

Use the term bias more carefully. For example, the only possible interpretation of the sentence in Sec 4.1.6 is that GRUAN has a dry bias, i.e. that it its mean is too low! If this is what is actually meant, this results should be stressed (and e.g. be part of the abstract). However, I assume it is an example of poor use of the term bias. Note that the correct meaning of bias is a deviation from the true mean, but it is frequently used incorrectly in comparisons like this. As it has become standard, I can't refuse the authors to use bias also in the later manner, but the text must then be very precise to avoid all risk for misinterpretation.

*The manuscript has been revised to use "systematic difference" in most cases where the difference is discussed between two instruments. When interpreting the overall meaning of the comparison results as a whole, the term "bias" is used because we intend to say the observations indicate a departure from the true mean.*

*The radiosondes, for example, appear to have a (dry) bias because we trust the GRUAN processing to represent the truth, and this result is consistent across the various other comparisons. We believe sun photometers appear to have a (dry) bias because co-located instruments at the Ridge Lab and 0PAL support that conclusion, which is also found in the cited literature. Lastly, the MUSICA data product appears to have a (wet) bias because all comparisons with co-located instruments show a systematic difference that is similar in magnitude when considering their own errors and biases.*

Paragraphs consisting of a single sentence should be avoided.

*While this is a common stylistic rule-of-thumb, single sentence paragraphs are sometimes succinct and grammatically acceptable. Cases of single sentence paragraphs have been reviewed and often modified.*

Some comments on a more detailed level:

The abstract should be more specific, i.e. more hard facts, and less vague comments such as "Good agreement is observed" and "A variety of biases and calibration issues are revealed". The fact that the MUSICA retrievals have deteriorated at the site shall be clearly expressed, even if this a negative result.

*The abstract has been revised to focus on the comparison results key to conclusions involving the new datasets presented, the MUSICA product, AERI retrieval, and GRUAN processed radiosondes.*

Page 7, line 22: What signifies "statistical errors"?

*Statistical errors involve measurement noise, as well as contributions from errors associated with the line of sight and temperature profile. See Schneider et al. 2012, Table 2 for a detailed breakdown.*

Page 7, line 25: N is not defined in the text.

*A definition for N (i.e. the number of measurements) has been added to Section 3.1.*

The difference between standard and extended MUSICA should be expressed more clearly. The text indicates that it is only the SZA threshold differs, but this is not consistent that are differences between the two versions for SZA < 78.5.

*The text has been revised to make sure there is clarity on this point. The MUSICA quality control filtering involves a few different aspects. For example, $CO_2$ is also retrieved and is used as a check on the retrieval. The MUSICA SZA limit of 78.5° has been lifted for this study in order to expand the time period of data available. The high latitude of the site would otherwise limit measurements to April to August.*

Page 8, line 5: How is this sensitivity defined?

*Sensitivity is the sum of the row kernels. This definition has been added to the text describing the MUSICA retrieval in section 2.1.*

Page 8, line 8: What is XCO2?

*$XCO_2$ is the column-averaged dry air mole fraction of $CO_2$. This definition has been added to the text in Section 2.1.*

Page 8, line 19-20: PCA in itself does not remove any noise! It is just a transformation the measurement. The text implies that a filtering is done after PCA, and it is then critical to describe how many components that are left. Another example on that the AERI algorithm is not described properly.

*We have modified the text to clarify and provided an additional reference that includes more detail specific to application to AERI instruments. We have also specified the number of principal components left after filtering. Because the noise reduction using PCA is described in detail in the references and is standard for AERIs, we hope that a more lengthy description that repeats the literature is not needed.*

*Page 8, lines 19-20: changed from*

*"principal component analysis was used to reduce noise (Antonelli et al., 2004)."*

*to*

*"Instrument noise in the P-AERI was reduced via a method that employs principal component analysis followed by noise filtering (Antonelli et al., 2004, Turner et al. 2006). This method is described in detail by Antonelli et al. (2004) and Turner et al. (2006), and thus we only mention specifics related to the P-AERI here. There are ~3000 spectral elements (e.g., for ~400 to 1900 $cm^{-1}$ at 0.5 $cm^{-1}$ spacing). To ensure that the number of samples, or downwelling radiance spectra, is sufficiently greater than the number of spectral elements, we process 10,000 samples at a time. The number of principal components retained is determined according to Turner et al. (2006), and is typically between 85 and 240."*

*We appreciate this reviewer's careful review and believe it has greatly strengthened the description of the AERI retrieval.*

Page 13, line 4-5: "two satellite instruments" is left from an earlier version.

*Thank you for catching that oversight. That text has been removed.*

Page 14, line 13-17: This test, does it apply closest or all version of measurement matching?

*This was done with the closest matched pairs. Text has been added to ensure this point is clear.*

Page 14, line 14: Add [] around mm, to be consistent with the next unit.

*Yes. Thank you for catching that.*

Sec 3.4: With satellite part removed, why having Sec 3.4.1 and 3.5, and not just 3.4?

*Agreed. This change has been made.*

For these profile comparisons you must also discuss the vertical resolution and the measurement response of the MUSICA retrievals. I assume that the good values at 15 km just reflects a low variability around a priori.

*It is true that the MUSICA retrievals have limited sensitivity at 15 km. We discussed the vertical resolution of the MUSICA retrievals when introducing the dataset in Section 2.1, with an example averaging kernel and sensitivity plot given in Figure 5.*

*To satisfy comments from another reviewer, the profile comparisons have been removed.*

Page 17, line 14: 2ppm/6% implies a mean VMR at 15 km of 30 ppm. Is that correct?

*Yes.*

Page 18, line 4-5: Don't understand what this sentence tries to stress.

*The sentence aimed to reiterate the challenging reality of comparing water vapour profiles, given the high variability. This section has been removed to satisfy the comments of another reviewer.*

Page 18, line 23-24: How can the agreement be close, when there is a consistent wet bias? That is, statements such as good and close agreement are objective and should be avoided. Just present the numbers.

*This sentence, and similar content for other instruments, have been revised to focus on the numbers.*

Page 19, line 9: "showed" refers to what? Earlier in the text or another article?

*This study was originally done using the MUSICA v2012 retrieval product. Prior to submission, it was revised to use the MUSICA v2015 retrieval product. The difference in results was interesting and became a conclusion for this paper. This sentence has been revised, and a table summarizing instrument comparison results for MUSICA v2012 has been added.*

Page 23, line 4: The statement is correct, but it does not justify to operate multiple remote sensing instruments. My point here is that a microwave radiometer gives highest possible temporal coverage. If the microwave radiometer can not be used (due to precipitation), there is no other measurement technique to use instead. Other techniques can have lower random or systematic error, but that is another question. That is, the logic of the paragraph does not hold.

*There are advantages to having multiple instruments measuring water vapour. The MWR's ability to capture water vapour's short term variability is excellent. The AERI instruments, sun photometers, and 125HR Fourier transform spectrometer (FTS) can (and often do) take measurements as frequently as every 5 minutes. While the microwave radiometer measures more frequently than this, investigating many atmospheric processes do not require sampling more often than 10 minutes. Thus, many of the instruments installed at PEARL offer high temporal-resolution measurements of water vapour.*

*Radiosonde measurements of atmospheric water vapour lack the temporal coverage of the microwave radiometer, and many other instruments. However, they offer the best available vertical resolution. Thus, they are complementary to the more frequent measurements taken by instruments such as the 125HR and microwave radiometer.*

*Atmospheric measurements taken with an FTS instrument such as the 125HR, while limited by a need for sunlight and limited in frequency to every five minutes (but often measurements once per hour because FTS measurements taken in the NDACC framework at PEARL typically rotate between six filters – only one of which is used for MUSICA retrievals), offer the ability to simultaneously measure many atmospheric gases. FTS $H_2O$ measurements thus offer complementary information that the MWR, for example, does not. Moreover, the FTS measurements also yields simultaneous accurate $\delta D$ information (Schneider et al., 2016) that can be paired with the $H_2O$ measurements to gain insights into transport processes (e.g. Noone et al. 2012). This breadth of atmospheric context is not available using most other instruments, and justifies the use of an FTS even with the presence of an MWR.*

*Operating and maintaining instruments in a remote and harsh environment poses significant challenges. The presence of multiple instruments enables continuous validation of instrument measurements to ensure any problems are identified and new techniques such as those presented in this study, are validated. This reality is shown clearly by the Figure 3 timeseries, which shows measurement gaps in all instruments for a variety of reasons (e.g. the E-AERI required an extended downtime for repairs). Reliance on a single instrument, however virtuous, is a vulnerability from a practical point of view.*

Finally, I must be bold and ask if really all authors have actively contributed to the manuscript? I ask this as the short summary just uses "I" and "my", indicating that just the first author has been active. My standpoint is that an active participation is required for co-authorship (i.e. persons should not be added just for politeness).

*The lead author sincerely regrets the mistaken use of "I" and "my" in the summary. This will be corrected to "we". This study involved contributions from all authors listed. These contributions are summarized in the added section for author contributions.*