Comments from the author provided in blue.

The retrieval of water vapour in polar regions is on the one hand highly challenging and on the other hand of high interest for various reasons, among others, due to the highly sensitive response of at least the Arctic to climate change. Thus, the overall topic and objectives of the paper are highly relevant. Retrieval improvements and enhanced applicability are presented and evaluated. Thus, the paper fits into the scope of AMT. However, in my view point the paper requires a few substantial improvements.

From my view perspective the following major points need to be addressed:

1) Section 2.6 introduces one of the new features of the retrieval. Unfortunately references are not properly linked or missing. But even when available it is likely not possible for me to understand the filtering method. I think this method needs to be explained in more detail. I further propose to show spatial maps which display the impact of the mask for one or two days in ,e.g., January and/or July. The objective is to showcase the impact of the screening in product space, other approaches for this are welcome.

Now Section 2.7: Added Figure 12 in Section 4 (Evaluation of changes/improvements in the retrieval) that shows the impacts of the mask for some selected areas through evenly spaced days - each three months approximately - through 2008.

It would be very helpful to explicitly mention the conditions (TCWV threshold, surface type) when the retrieval can be applied in, e.g., section 2.5 (i.e., extend the last sentence of section 2.3).

Section 2.5: Reformulated section 2.5 to better describe the different retrieval regimes and their working conditions, added Table 3 (P30) as summary.

Section 2 introduces a switch between retrievals at 7 kg/m2 and an upper application limit of 15 kg/m2. However, figs 1 and 8 exhibit features at 6 kg/m2 and figs 1 and 3 values above 15 kg/m2. Figure 1 also shows that the majority of values in summer are around and above 15 kg/m2. Please clarify this (seeming) contradiction.

Clarified in Section 2.5 that the switch between retrievals is done in the brightness temperature space, which doesn't correspond to a specific TWV value. We provide approximate limits and switch, but they are not 'hard limits', hence the values above 15kg/m2 in Figures 1 and 13.

2) It is clear that the availability of ground truth data hampers the evaluation of TWV in the Arctic. However, three GRUAN, a few more GUAN and maybe other WMO stations are within the area of interest (and were partly used for retrieval development). I can imagine that data from some stations might exhibit too large values even in winter. Nevertheless, I propose to assess the utilisation of radiosonde data from these sources for evaluation of MHS and AMSU-B over a common period and given sufficient collocations use it in addition to N-ICE data. The joint evaluation of AMSU-B and MHS and the application of the new ice cloud masking using ground-based or in-situ data is currently lacking but would strongly support one of the main objectives of the paper.

Added new subsection on paper (3.2) with a comparison of AMSU-B and MHS derived TWV with GPS and radiosonde data during the period 2008-2009.

3) To me, the current presentation of evaluation results requires rephrasing: I am not a validation expert for the polar regions. However, the interpretation of evaluation results/performances as "good" seems to go too far. I propose to not interpret the results this way and just summarise the results (which might be termed as indicative of successful application to MHS and improvements).

P8 L5-7: Changed most qualifiers from description to fit this comment, and removed 9 outliers from MHS TWV comparison with N-ICE TWV, which improves the performance significantly.

Alternatively, a brief summary of existing results from other evaluation efforts in the Arctic can be provided. Given superior quality such statements might be adequate.

P2 L26-30: Provided the following brief summary from the evaluation papers mentioned previously: In Rinke et al 2009, a comparison with the HIRHAM model showed realistic patterns and maximum root-mean-square

differences for monthly data in summer of 1-2.5 kg/m2. For the comparison with Ny Alesund radiosondes in Palm et al. 2010, the correlation coefficient was 0.86 and the slope 0.8 \pm 0.04 kg/m2. And lastly, in Buehler et al. 2012, ASMU-B TWV are compared to GPS data from Kiruna, with standard deviations of 1kg/m2 and a correlation coefficient of 0.86.

Depending on results from 1) a successful application to MHS and an improvement via ice cloud masking might have been proven as well.

While Figure 13 deals with AMSU-B data masking, the same mask can be applied successfully to MHS, as shown in the already masked maps from Figure 11.

I don't think that the terminology "benchmark" is adequate for a satellite based TWV product in the Arctic. Please speak of "comparisons" instead.

P9,L2 Description of AMSR-E in section 4; P10,l8 Conclusions — > Both mentions rephrased

4) The paper requires careful cross-reading for various reasons. Among them are: partly units are not provided, a few references are not properly linked or missing and various formulations don't seem to be correct. Some of the latter are mentioned below. I am not a native speaker and propose that a native speaker is cross-reading the paper.

Carefully proofread paper.

In addition I have the following minor comments, partly linked to the comments above:

#) The evaluation results exhibit features caused by changes between retrieval algorithms. A brief discussion on expected impacts of reprocessed products would be adequate. E.g., the application of thresholds can easily lead to temporal and spatial inhomogeneities. I propose that the team can find a more physical solution than the one mentioned in the conclusions to overcome this issue. I recommend to reformulate such potential future plans.

We think there is no easy "physical" solution for the discontinuities at the boundaries between the different sub-algorithms. The sub-algorithms, just like many retrieval algorithms that do not use inverse methods, are based on regression analyses which result in "calibration parameters". Thus, in the end, the average state of all atmospheric parameters except water vapour (e.g., the temperature profile) is contained in the calibration parameters.

It is known that the error of the retrieval for each subalgorithm increases strongly when approaching its upper (saturation) limit (see, Melsheimer & Heygster, 2008, Appendix IV) -- Therefore, an appropriate weighted average of the retrieval results of two sub-algorithms in their overlap range (one algorithm nearing its upper limit with increasing errors, the next one still being in it low range with small error) appears meaningful to us.

#) Page 1, line 17: Overall TWV increases due to increases in temperature. I propose to rephrase Accordingly.

P1L18 Added mention to TWV increase due to temperature

#) p 2, l4: Usually a frozen retrieval is applied consistently. However, various other factors are important in this context as well. Please rephrase, i.e., delete "analysis method" and add others, e.g., instrument degradation.

P2L8 Rephrased as suggested

#) p2, l23: Please delete "successfully" and briefly mention the results (i.e., quality indicators as used in this paper).

P2, L27-31 Added short summary of results from the three papers mentioned

#) p2, l28-29: Please cross-read.

P3, L1-4 Reformulated, ordered sections, added description of new analysis with radiosondes

#) p3, l14, l16, 17: Please delete "will" and remove definition of abbreviation here. Please mention Metop-A and Metop-B explicitly.

P3, L33 Solved

#) Section 2.2: Please mention briefly how the two retrievals are defined (or give reference to Appendix).

It is not clear to us what is meant by the “defining” the “two retrievals” - there is the retrieval according to the cited work by Miao (2001), briefly described in section 2.3 (formerly 2.2) and the extended version according to the cited work by Melsheimer and Heygster (2008), briefly described in section 2.4 (formerly 2.3). The calibration parameters used by the different algorithms are discussed in detail section 2.6 (formerly 2.5) and in the appendix. We feel this and the cited references sufficiently define or describe the algorithms and sub-algorithms.

#) Section 2.5: A brief discussion of where - in TWV space - these transitions occur in a climatological sense would be helpful.

Clarified in Section 2.5 that the switch between retrievals is done in the brightness temperature space, which doesn’t correspond to a specific TWV value. We provide approximate limits and switch, but they are not ‘hard limits’, hence the values above 15kg/m2 in Figures 1 and 13.

#) Section 2.6: Various references are missing (i.e., appear as “?”). Please provide them.

Section 2.7: References provided (Gonzalez and Woods, 2007), van der Walt et al. (2014))

#) I propose to change the order of sections 2.5 and 2.4.

Changed order of Sections: Former Section 2.4 is now 2.6, due to changed number of sections.

#) p6, l26: “unexpectedly small” – other months have only half of the amount of data. Maybe the feature has other reasons. Please explain.

P7L12-13 We wanted to refer to the fact that, for a winter month, this amount of data is small (comparing the 7723324 points in December with the 10691385 in January or 9858305 in February). Clarified this in the text

#) p6, l29, l30: Please provide unit for RMSD and delete “really”.

P7, L15-17. Solved

#) Section 3, last paragraph: In addition to 3) please mention a systematic high bias plus fairly large outliers.

P8, L15-17. This corresponds to the middle of Section 3.2 now. Mentioned high bias, and large outliers on the context of outlier removal.