REPLY TO REVIEWER #1

The authors highly appreciate the constructive comments. They are very useful contributions that will certainly help to improve the revised manuscript. In the following, the authors reply point by point to all Reviewer comments, which are written in italic while our replies are in standard font. Within the manuscript all changes from the submitted version are highlighted in red.

MAJOR COMMENTS OF REVIEWER #1:

General remark:
Different communities use different notations, which is also true for the lidar and microwave community. This paper should serve both communities and therefore, in order to keep things straight forward, we decided to stick as close as possible to the notation of Rodgers (2000) – the most important textbook for optimal estimation.

1. The discussion of the averaging kernel in lines 219 to 228 is wrong. Quoting p. 47 of Rogers (2000), "rows of A are generally peaked functions . . . with a half- width which is a measure of the spatial resolution of the observing system, thus providing a simple characterisation of the relationship between the retrieval and the true state."
On line 220, the averaging kernel actually describes the final values' dependency on their true magnitude, in this case indicating the smearing of information across multiple levels. The subspace of state space in which the retrieval must lie is constrained by the a priori covariance matrix.
I have never encountered the use of degrees of freedom to measure vertical resolution, as in Eq. (6). After conferring with colleagues that work more closely with OE of thermal profiles, we consider it to be at a confusing and poor choice of metric, if not intrinsically wrong. A much more common and robust metric would be the width of the averaging kernels (e.g. the full-width at half maximum of a row of A).

The authors consider correct what is stated in the manuscript in lines 219 to 228. To corroborate this we cite Rodgers (2000), pages 52, 53 and 54:

"*Resolution*, like *information*, is a word with a multiplicity of meanings, and tends to be used differently in different contexts. […] Possibilities include characteristics of the averaging kernel or state resolution matrix, such as the width of the averaging kernel or the point spread function, where *width* has many possible interpretations, the response of the retrieval to sine wave perturbations in the state, and **the range of heights covered divided by number of independent quantities measured**. […]
Possible characterisations of resolution are […]
(iv) the degrees of freedom for signal, d_s is the trace of the averaging kernel matrix. **Consequently the diagonal of A may be thought of as a measure of the number of levels per degree of freedom, and thus a measure of resolution.** "

Which is the definition of equation (6). Moreover, other authors have previously used this definition for vertical resolution, e.g. Liu (2014).

2. You still misuse the term 'error' in Section 5.3. In layman's terms, the error is how wrong your measurement was and the uncertainty is how wrong you think it might be. Robust definitions can be found at http://www.iso.org/sites/JCGM/ GUM-introduction.htm. The error can be approximated by considering the difference between the retrieved value and a more accurate reference measurement, as you do in Fig. 5. What OEM estimates is the uncertainty on the retrieved

value, which describes the range of errors you would expect to see if you infinitely repeated the observation.
I also understand why you add the word 'theoretical', but it isn't necessary as the uncertainty is a prediction of a probability distribution.

We agree with the reviewer, but as explained in the general remark, we prefer to consequently adapt the notation given by Rodgers (2000). According to this book, $\hat{S}$ (or S_{op} in our manuscript) is the total retrieval **error** covariance matrix (see page 58). In addition, with the word *theoretical* we want to emphasize that it is an estimate, not a direct difference to the true state.

Further, note that the term *theoretical error* is widely used in literature, e.g. Healy (2000), Phalippou (1996), Puliafito (1995), etc.

A clarification note has been included in the manuscript (former line 214), which now reads:

"From S_{op}, the theoretical error (in kg/m^3) associated to each altitude of the retrieved profile x_{op} is calculated as the square root of the main diagonal elements in S_{op}. The word *theoretical* emphasizes that it is an a posteriori estimate, and not a direct difference to a given reference".

MINOR COMMENTS:

TB is a non-standard (and frankly annoying) abbreviation for brightness temperature. I would suggest BT or T

We understand the argument but there are some reasons for the use of TB for *brightness temperature*. First, we do not want confusion with the physical temperature that's why the do not use T. Second, microwave radiometry developed from radioastronomy, where the notation of brightness temperatures as TB stems from.

TB is used throughout many different applications of microwave radiometry, being one of the most important textbooks Janssen, 1993. To our best knowledge the majority of publications in the ground-based microwave community uses TB while for space borne applications BT is preferred.

L145 The word 'drift' doesn't appear in that Whiteman paper. I think you mean the thermal sensitivity of the filters.

The reviewer is right. We indeed meant the "thermal sensitivity of the filters". Specifically, we refer here to the fact that the interference filter transmittance spectrum is slightly temperature dependent. As temperature increases, the thicknesses of all dielectric layers increase and all layer indices change. These effects combine in a way that the transmittance spectrum shifts to slightly longer wavelengths with increasing temperature, with the thermal coefficient being a function of wavelength. This is now more clearly stated in the text and the sentence now reads:

"For example, an additional uncertainty (<1%) may be considered related to the use of narrowband filters, the temperature dependence of H_2O Raman scattering and the thermal sensitivity of the filters (Whiteman, 2003) "

L150 There is a subtle point here that, though you don't need to mention it in the paper, you may wish to consider. Poisson statistics state that the variance of a measurement sample is equal to its mean. The lidar community uses this to assume that the value of a measurement is equal to its uncertainty squared. However, that measurement is only one sample from the distribution and is therefore an imperfect estimate of the mean; it's simply the best estimate available. This isn't usually important but in a statistical analysis, such as OEM, this approximation implicitly states that smaller values are more accurate as they have smaller uncertainty (e.g. the OEM will fit 1 ± 1 more closely than 100 ± 10). Hence, you may wish to investigate if your analysis is biased towards small

values in the presence of exceptionally negative noise (i.e. data noticeably smaller than that around it).

Poisson statistics have been applied to the lidar signals used in this paper after a careful verification of its validity and applicability. In this respect, an analysis was carried out to determine total variance profiles from high resolution water vapour mixing ratio profile measurements (temporal resolution of 10 sec and a vertical resolution of 90 m) for lidar data from the same field campaign. The autocovariance method defined by Lenschow et al. (2000) was then applied to effectively separate atmospheric variance from the noise variance in the total measured variance. This method is based on the consideration that atmospheric fluctuations are correlated in time, while instrumental noise fluctuations are uncorrelated. Profiles of the total noise error (determined as the squared root of the noise variance) affecting water vapour mixing ratio measurements obtained through this method have been compared with estimates of the water vapour mixing ratio measurement uncertainty due to shot noise derived with Poisson statistics. The shot-noise error is typically a predominant part (around 80-85 %) of the total statistical error, while the reminder part is to be attributed to other statistical error sources (among others, the detectors' dark noise error). The comparison we performed for this data set confirms that the major contribution to the total noise error originates from photon shot noise. But error estimates derived with Poisson statistics are at any altitude proportional to the noise error determined through the autocovariance method. This result confirms the correct applicability of Poisson statistics in estimating measurement error and consequently the possibility to apply results obtained from Poisson statistics in the observation uncertainty covariance matrix.

L208 The test this sentence describes doesn't match the condition given in Eq. (3). Which one do you actually use?

That is right. The sentence has been corrected in the manuscript as follows:

"i.e. the difference between the forward model applied to the atmospheric state at iterations n and n+1…"

L 234 If you're using the radiosonde data both to determine the lidar calibration factor and as an a priori, why don't you put the calibration factor in the state vector and constrain it (and it's uncertainty) with the a priori? For example, the difference between the blue and black curves in Fig. 2 is about 5%, which would be accounted for by the uncertainty in the calibration factor.

The a priori information is calculated from the complete set of radiosondes launched during HOPE (217 launches in total, at least two per day). The mean calibration coefficient for the Raman lidar was estimated comparing Raman lidar and radiosonde data, but only considering clear sky radiosonde launches when the Raman lidar was operational (60 launches in total). So, the dataset used for the a priori information is numerically different from the dataset used for the calibration of the Raman lidar. Additionally, for the purpose of the Raman lidar calibration, lidar and radiosonde data are compared in an altitude region with an extent of 1 km above the boundary layer (to minimize air mass differences associated with the distance between the lidar station and the radiosonde launching facility, approx. 4 km), while the complete radiosonde profiles are used as a priori information. So, the dataset used for the a priori information is also different in terms of vertical extent from the dataset used for the calibration of the Raman lidar. Nevertheless, while in principle it would be possible to put the calibration factor in the state vector and constrain it with the a priori, potential systematic effects associated with this approach could occur.

L245 This reviewer is pleased to see correlation matrices rather than covariances.

For your information, the covariance matrix is presented below. It is naturally dominated by the highest occurrence of water vapor in the boundary layer.

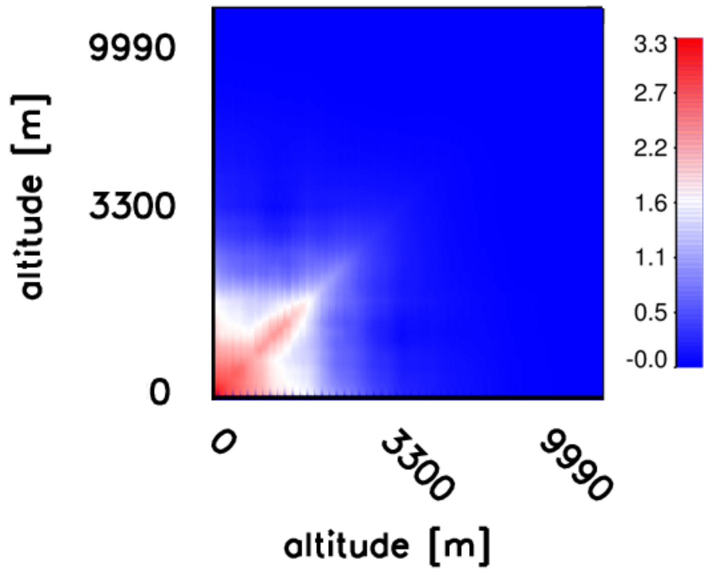

Figure R1.1 Covariance matrix derived from 217 radiosondes launched during HOPE. Covariance is shown for absolute humidity as a function of the altitude (from 0 to 10 km above the ground) in g/m3.

Sec. 3.4. I remain disappointed that you do not consider a more detailed forward model for the lidar. Could this be mentioned as possible future work, in an attempt to inspire other researchers?

Indeed, a more complex forward model can be mentioned as a possible future evolution for the considered approach. A variety of lidar forward simulators have been developed in the lidar community. The following clarification has been included:

"Therefore, the lidar FM for water vapor simply performs the conversion from absolute humidity to mixing ratio. However, the implementation of a more complex lidar forward model, e.g. the approach implemented by Sica (2016), could be considered in future studies."

Fig. 2 The error bars don't cover the discrepancy between the retrieval and the radiosonde. Does this mean that your uncertainty estimate is too small or is the uncertainty on the radiosonde data large enough for the two profiles to be consistent?

There are three sources contributing to the discrepancy between retrieval and radiosonde:

i) the uncertainty of the a priori and of the lidar and MWR measurements (which are combined into the retrieval uncertainty),
ii) the uncertainty of the radiosonde measurement (5% RH as reported by the manufacturer) and
iii) the difference in the atmospheric volume measured by the instruments (lidar, MWR and radiosonde). The magnitude of the latter is difficult to assess, as no truth about the 3dimensional distribution of the water vapor field is available. As reported in the manuscript the radiosonde was launched 4 km apart from the lidar location and drifts significantly during its ascent. This source of uncertainty can be high especially in the boundary layer (see Steinke et al., 2015) and is not represented in the error bars associated to the retrieval uncertainty.

The underestimation of the retrieval uncertainty is therefore most likely due to contributions from ii) and iii), which are not taken into account in the error bars from figure 2.

L335 The value at 5 km is consistent with those at 3 and 4 km, so you can't necessarily call that an increase. Your argument is strongest when pointing out that the joint technique gets 3 km closest to the radiosonde.

Indeed the sentence: "Only the combined retrieval can detect the drop in humidity at 3 km and the increase at 5 km" was not really correct as in fact the absolute humidity content keeps almost constant for the combined retrieval. As suggested by the reviewer we have changed the sentence pointing to the result that at 3 km the combined retrieval gets values closer to the radiosonde. The sentence has been changed accordingly and now reads:

"Only the combined retrieval reveals absolute humidity values in agreement with the radiosonde at 3 km".

TECHNICAL POINTS:

1. ***I do not know if this journal prefers 'ground-based' to be hyphenated; you use both so please pick one.***
2. ***Units are frequently italicised (presumably because they have been included within a $$ environment). Please consistently use plain font.***
3. ***Many of the references list both a DOI and a URL; the URL is redundant.***
4. ***L2 Nowadays there are a wide***
5. ***L31 which is difficult to capture with one instrument***
6. ***L41 Perhaps use 'have become' rather than 'became' and 'over recent years' rather than 'during the last years'.***
7. ***L45 You use 'day time' here and 'daytime' on line 49. Please pick one.***
8. ***L59 Are you sure you mean 'features'? I thought the measurements of a MWR would be better described as 'levels'.***
9. ***L83 'to incorporate' could be removed without changing the meaning of this sentence. If you prefer to keep it, 'one' needs to precede it.***
10. ***L85 You don't need to pluralise 'month' when used as an adjective. L111 Raman scattering of the 355 nm beam***
11. ***L133 During HOPE, BASIL was calibrated***
12. ***L134 calibration coefficient was estimated by comparing***
13. ***L144 Use H$_2$O rather than H_2O. Repeated on L165 and L167.***
14. ***L163 of the K-band contain***
15. ***L166 liquid water increases with***
16. ***L195 This equation doesn't conform to the journal's style guide.***
17. ***L197 the MWR and the profile of the mixing ratio***
18. ***L203 when a perturbation is added to the atmospheric state vector***
19. ***L222 degree of freedom and can be interpreted as***
20. ***L215 represent the number of independent***
21. ***L295 divides the atmosphere into layers***
22. ***L298 To typeset the second exponential, I would recommend***
    \exp\left(-\int_0^s \alpha(s') \mathrm{d}s' \right)
23. ***L317 a complete profile from the ground up***
24. ***L324 a dominant role in defining the vertical***
25. ***L329 The uncertainty is small in the region***
26. ***L350 the vertical resolution for only-RL becomes infinite.***
27. ***L381 during HOPE, and therefore this period***
28. ***Fig. 5 Invert the order in which the three plots of 5(b) are described to mimic the left-to-right manner in which they are presented.***
29. ***L462 regions (see section 5.2)***
30. ***L510 The average total number of DOF***

31. ***L511 increasing by almost 2 DOF***
32. ***L542 The magnitude of the increase in RL measurement uncertainty is based on the L545 error. Therefore, we have***
33. ***L548 The new averaged errors are very similar***
34. ***L569 different sensors has come more and more into focus***
35. ***L616 The page number for Delanoe and Hogan (2008) is D07204.***
36. ***L619 The page number of Di Girolamo et al. (2004) is L01106.***
37. ***L656 The page number of Löhnert et al. (2007) is D04205.***
38. ***L662 The page numbers of Löhnert et al. (2014) are 1157–1174.***
39. ***L680 An extraneous BibTeX field appears to have been printed between the DOI and year.***

Comments from 1 to 39 have been addressed in the manuscript.

Extra references:

Liu, D., Lv, C., Liu, K., Xie, Y., & Miao, J. (2014). Retrieval Analysis of Atmospheric Water Vapor for K-Band Ground-Based Hyperspectral Microwave Radiometer. *Geoscience and Remote Sensing Letters, IEEE*, *11*(10), 1835-1839.

Healy, S. B., & Eyre, J. R. (2000). Retrieving temperature, water vapour and surface pressure information from refractive-index profiles derived by radio occultation: A simulation study. *Quarterly Journal of the Royal Meteorological Society*, *126*(566), 1661-1683.

Phalippou, L. (1996). Variational retrieval of humidity profile, wind speed and cloud liquid-water path with the SSM/I: Potential for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, *122*(530), 327-355.

Puliafito, E., Bevilacqua, R., Olivero, J., & Degenhardt, W. (1995). Retrieval error comparison for several inversion techniques used in limb-scanning millimeter-wave spectroscopy. *Journal of Geophysical Research: Atmospheres*, *100*(D7), 14257-14267.

Janssen, M. A. Atmospheric Remote sensing by Microwave Radiometry, John Wiley and Sons, Inc., 1993.

Sica, R. J. and Haefele, A., Retrieval of water vapor mixing ratio from a multiple channel Raman-scatter lidar using an optimal estimation method, Appl. Opt. 55, 763-777 (2016)