

Dear D. N. Whiteman,

We have appreciated your valuable corrections and we have carefully considered your detailed comments on our manuscript. We would like to thank you for taking the time to provide such an insightful guidance. Herein, we explain how we revise the paper based on your comments and recommendations. We respond specifically to each suggestion below. Please find attached the revised version of the paper.

Review of “A Raman lidar at Maïdo Observatory (Reunion Island) to measure water vapor in the troposphere and lower stratosphere: calibration and validation” Vérèmes et al.

Submitted by David N. Whiteman

General comments:

It is exciting and very impressive to have a Raman lidar in the Southern Hemisphere dedicated to research on stratosphere-troposphere exchange processes and the long-term monitoring of water vapor trends in the upper troposphere and lower stratosphere. The commitment of the French and European research institutions to these tasks should be greatly applauded and these efforts need to be documented in the refereed literature. I find, however, that in its current form this manuscript needs too much revision to be publishable. I recommend that the authors carefully consider the comments below and revise the paper accordingly and resubmit. General comments are given first then more specific detailed comments follow.

We carefully considered the comments of both reviewers and we revised the paper accordingly. The new version of the paper includes :

- 1) A description of GRUAN requirements regarding the measurements and the uncertainties for addressing the monitoring of the water vapor in the UT/LS and studies of atmospheric processes (Sect. 1 Introduction)**
- 2) A correction of the saturation of the signal is applied to the whole dataset (new Sect. 2.2.1) and the figures (as their analysis) are updated accordingly**
- 3) A correction and an improvement of the calculation of the total uncertainty:**
 - Uncertainties of the external source of measurement for the calibration and of the temperature-dependence are taken into account (Sect. 2.3.6).
 - Uncertainties on the overlap function, on the fluorescence and on the correction of the saturation are discussed (new Sect. 2.2.2).
- 4) A refinement of the methodology of calibration (Sect. 2.3): the determination of the outliers is optimized in order to decrease the impact of the uncertainty of the GNSS measurements.**
- 5) A significant improvement of the radiosonde-based comparative calibration methodology (Sect. 2.3.4)**

(1) The goal of this paper is to demonstrate that the Maïdo water vapor lidar is ready to supply quality data for various scientific investigations such as process studies and trend detection. The data quality needed to address these two types of studies, however, is very different. A general discussion is needed that deals with the lidar measurement uncertainty and measurement and calibration requirements for addressing different scientific studies. Then the authors can assess how well this new instrument meets these measurement needs. Documents that pertain to this type of discussion include:

(1.1) For discussions of measurement requirements for addressing particular scientific studies

1. GCOS-134. Appendix 1 gives water vapor calibration stability requirement of 0.3% per decade for revealing water vapor trends.
2. GCOS-171 The GCOS Reference Upper Air Network Guide (2013) in particular see sections 4, 7

3. Whiteman, D. N., K. C. Vermeesch, L. D. Oman, and E. C. Weatherhead (2011), *The relative importance of random error and observation frequency in detecting trends in upper tropospheric water vapor*, *J. Geophys. Res.*, 116, D21118, doi:10.1029/2011JD016610.

We add a paragraph to the introduction (L61 to L83) to detail the requirements on the measurement uncertainties and on the calibration stability for the following scientific objectives regarding the water vapor: detection of trends in the Upper Troposphere, detecting of trends in the Lower Stratosphere, and process studies. We take care in separating UT from LS. Our statements are now based on the requirements of the GCOS reports that you mentioned, thanking you for this valuable reference. This approach is all the more important that Reunion Island is preparing a candidature to become a GRUAN site. We followed your advice: we conclude on how well (or not) the Lidar1200 meets the GCOS requirements (L819 to L830). We also mention what (and how) should be improved to reach the requirements that are not achieved (L849 to L859).

(1.2) For discussions of measurement uncertainty and calculating the total uncertainty budget of the lidar water vapor mixing ratio data product, I recommend starting with a fully detailed version of the lidar equation so that a correlation between the equation and the uncertainty terms evaluated can be described. Also, authors are advised to consult the following for assessing total uncertainty budget for the water vapor mixing ratio calculation

1. Immler *et al.*, *Atmos. Meas. Tech.*, 3, 1217-1231, 2010 <http://www.atmos-meas-tech.net/3/1217/2010/>doi:10.5194/amt-3-1217-2010
2. GCOS-171 Section 3.
3. Whiteman, D. N., Cadetola, M., Venable, D., Calhoun, M., Miloshevich, L., Vermeesch, K., Twigg, L., Dirisu, A., Hurst, D., Hall, E., Jordan, A., and Vömel, H.: *Correction technique for Raman water vapor lidar signal-dependent bias and suitability for water vapor trend monitoring in the upper troposphere*, *Atmos. Meas. Tech.*, 5, 2893-2916, doi:10.5194/amt-5-2893-2012, 2012. See Appendix A3.

We might have neglected several of these uncertainties based on other published studies that once neglect the overlap correction (for us : motivated by the geometry of our lidar), another time they neglect the effect of the aerosols (because above Reunion Island there are very few background aerosols excepting for specific situations like volcanic activities or biomass burnings) or even the temperature-dependence. To quote Whiteman *et al.* (2012) : « To our knowledge, however, there have been no previous efforts to quantify the total uncertainty of Raman lidar water vapor mixing ratio measurements including the effects of the random error and all significant sources of systematic errors. ». As you suggest us, we start freshly with the Lidar equation and then develop it (L192 to L212). We took care about the list of uncertainties in Whiteman *et al.* (2012). Each term of potential systematic uncertainty is described and calculated (L215 to L223 and L230 to L242). If they are neglected, it is now explicitly justified (L218 to L223). Sect. 2.2 is significantly improved, please see L187 to L262. Note that the left part of Figure 7 (P26) move to this section, it is the new Figure 1.

Regarding the calibration uncertainty, it is true that we did not formally take the uncertainty of the external source of measurement into account in our calculation. Looking at the paper of Turner *et al.* (2002) when the error on the MRW is between 4 and 10%, the standard deviation of about 4%; the total uncertainty on the calibration is of « roughly 5% ». We might have been misled on how the uncertainty of the external source of measurement was integrated to the total uncertainty calculation. It is corrected and both the uncertainty of the GNSS measurement and the uncertainty in the transfer of the calibration are taken into account, please see L450 to L487.

(2) After establishing the data requirements and lidar total uncertainty as requested above, the authors need to carefully justify that calibrating the Raman water vapor lidar data product with respect to GPS IWV is sufficient to meet the measurement requirements. As mentioned in the specific comments, it appears that the uncertainty in the GPS measurements themselves (before considering the uncertainty in transferring the GPS calibration to the lidar) may be 10-20% for as much as half the measurement periods studied. The authors plan to use lamp measurements to carry the calibration through periods of higher uncertainty in the GPS data so that would imply using the lamp measurements to carry calibrations forward for perhaps half the year. But the lamp measurements do not seem to be sensitive to some of the large changes in calibration that occur (Fig 2). These details need to be carefully considered.

It is true that in the literature, especially for the Raman lidar calibration, the uncertainty of the GNSS

measurement is around 7%. The experiments led during different campaigns conclude that the use of the GNSS or the microwave radiometer (with an uncertainty lower than 10%) is suitable for the calibration of water vapor (Bock et al., 2013; Leblanc et al., 2012; Whiteman et al., 2006, 2012). Turner et al. (2002) estimates the « total uncertainty on the calibration coefficient of roughly 5 % » with an uncertainty of the MRW between 4 and 10%.

There is a high number of weak IWV above Reunion Island, that seems lower than those of the other papers on Raman lidar calibration. It seems to be a regional specificity. It leads to a mean uncertainty of the GNSS measurement of 18%, related to the 10-20% for half of the measurement periods mentioned. Thus, it is a real challenge to calibrate the water vapor in Reunion Island because of the extreme water vapor contents. This raises a major interest in studying geophysical issues with this dataset. To minimize the effect of the high uncertainties of the GNSS measurement in case of low water vapor content, we refined the calibration methodology introduced in the first version of our paper by applying a threshold on the IWV used for the calculation of the calibration coefficients. The choice of the outliers has been rethought based on the remarks of both reviewers. The mean uncertainty is lowered to 7.5-11 % regarding the chosen threshold (please see Sect. 2.3).

It should be noted that despite this uncertainty on the calibration, the individual comparisons between lidar and CFH, and RS92 (not discussed in this paper) profiles don't show any systematic offset that might be associated to a large uncertainty on the calibration coefficient. What's more, the total uncertainty of the GNSS measurement is a mixture of systematic and random uncertainties. Thus, the real systematic uncertainty associated with the GNSS IWV should be about half weaker than the stated 7.5-11 %.

Regarding the potential limits of the methodology, the lamp measurement won't be used for the calculation of the calibration when it would be necessary. The combination of temporal series of the hourly calibration coefficient, the lamp measurements and the logbook will allow us to identify if there is any change in the coefficient. Then, if the sampling of IWV > 10 mm wouldn't be sufficient, sondes would be launched to get a calibration coefficient to compare with or simply to apply to the data. This should happen rarely. It might happen during one period in the year, in winter and not for half of the measurements. The paper is revised accordingly to all these points, please see our responses to the specific comments for more details.

(3) For easier editing in the future, I suggest using line numbers that carry through the entire manuscript (as opposed to starting at 1 on each page) for easier editing.

We share your opinion, it is easier to discuss about the paper by keeping the line numbering through the entire paper. It is changed. In this report, the references to the first version of the article are written this way: P2-l10 (page number-line number), and to the new version, specified this way: L65 (line number).

Specific comments:

INTRODUCTION

(1) I find this first important paragraph to be rather disjointed in its logical flow due to the large number of disparate topics that the authors attempt to cover in one paragraph. The range of topics introduced in this one paragraph would demand several paragraphs to smooth the discussion for the reader. Also some statements are unclear. I suggest a very significant re-write of this material.

We worked on the improvement of the introduction according to the suggestions below. The structure is changed, especially by structuring the different topics in more paragraphs. The statements that were unclear are rephrased.

Here are examples

PAGE 1

(1.1) *the second sentence* discusses “long term series” while *the third sentence* talks about selection criteria (not yet discussed) being important for international networks (not yet discussed).

(1.2) *The fifth sentence* introduces the need for meta data which is a fully different topic than what

precedes it.

(1.3) Lines 38-39 introduce uncertainty, algorithms and calibration which are large topics unto themselves.

The first paragraph of the introduction is split into different paragraphs. The structure is: Importance of the water vapor and variability, Long-term series & Reference observations, Networks & Scientific objectives. Regarding (1.1) - (1.2) - (1.3), they are still included in a same paragraph around the concept of « reference observations » (Immler et al., 2010). It deals with the different criteria to respect when supplying the data to the networks in order to establish reliable long-term series: regularity and longevity of the measurement, transparency in the data processing, identification and quantification of the sources of uncertainties, methodology of calibration and metadata. We add a paragraph (L61) to discuss about the measurement requirements for addressing particular scientific studies (1.1).

(1.4) Line 39 states “This rigor...” but what does “This” refer to? There is no rigor that is previously described that “This” refers to.

It refers to “The last exercise” (P1-L38) that has to be rigorous, we clarify this idea by adding an adjective : “This last exercise is carried out through a strict examination of the data processing algorithms and calibration methodologies. This rigor...” (L52-L53).

PAGE 2

(1.5) Line 1. The sentence “One of the challenging ECV to measure is water vapor mainly in the upper troposphere and lower stratosphere (GCOS, 2003)” could be more clearly stated as “Water vapor is a challenging ECV to measure in the upper troposphere and lower stratosphere (GCOS)”.

The sentence is changed following your suggestion, please find it from L45-L46.

(1.6) Line 2. The sentence “Water vapor is the main greenhouse gas” really does not fit in logically at this point. Such an assertion needs to come much earlier in the discussion as support for why the current effort is being undertaken.

This sentence is removed. The characterization of the water vapor as « a greenhouse gas » is now included in the second sentence of the introduction: “The factors influencing the spatiotemporal variability of this greenhouse gas are various: ...” (L39-L40).

(1.7) Line 4. please provide references also for the “transport and dynamical processes from eddies to synoptic scale events” portion of this sentence.

We add the study of Vogelmann et al. (2015) (L41) (that deals with this topic) to the text.

(1.8) Lines 5-6 discuss spatial variability of water vapor without regard to whether they are referring to upper troposphere or the lower stratosphere. The statements they make are much more appropriate for the troposphere than the lower stratosphere. This is an important distinction for the authors to make as they present different temporal and spatial averaging schemes for the lidar data processing later in the paper.

It is true that the sentence was more appropriate for the troposphere, now it is specified (L42). We also add a sentence to make a point on the difference of the water vapor variability between the UT and the LS (L44-L45). This piece of information is valuable considering that we will differentiate the establishment of the trends in the UT than in the LS later in the introduction, see response to (1.9).

(1.9) Line 7 introduces the concept of measuring trends in water vapor and what is needed to do so. But it is already published that monitoring trends of water vapor greatly depends on whether you are referring to upper troposphere or lower stratospheric trends. That distinction is not discussed here or elsewhere and surely needs to be. See reference below

1. Whiteman, D. N., K. C. Vermeesch, L. D. Oman, and E. C. Weatherhead (2011), The relative importance of random error and observation frequency in detecting trends in upper tropospheric water vapor, J. Geophys. Res., 116, D21118, doi:10.1029/2011JD016610.

A whole discussion of the measurement requirements to detect and quantify trends in the UT/LS is added L61 to L83. We clearly write it: « The concept of monitoring trends in the upper troposphere is different from doing it in the lower stratosphere » and the requirements are specified whether we are referring to UT or LS trends. The reference to Whiteman et al. 2011 is added (as 2011a).

(1.10) The remaining sentences in the paragraph introduce NDACC and GRUAN but make no mention of trend detection in reference to either of these networks. Instead, the discussion shifts to UT/LS exchange (for NDACC) and “characterization” for GRUAN.

This is true that both networks aim at monitoring trends, we forgot to write it clearly. It's done.

(2) *Line 19.* “Among the radiosondes, the hygrometers are the most efficient”. This sentence refers to a mixture of technologies and makes an inaccurate statement. Radiosondes measure temperature, pressure, RH and winds typically while hygrometers measure water vapor alone so hygrometers should not be considered “among the radiosondes”. And I suspect that the authors may be assuming that “hygrometer” refers to those instruments that measure water vapor using the chilled mirror technique, but that is not correct. Hygrometer is a more generic term referring to any instrument that measures the water vapor content. Also, what does it mean for a hygrometer to be the “most efficient”? I suspect the authors mean to refer to accuracy in some way instead.

We were actually referring to the frost point hygrometer as the most accurate (that we previously characterized of the most “efficient”) radiosonde's sensor in measuring humidity. These misuses is corrected (L89).

(3) *Line 20* contains a misstatement about CFH claimed accuracy from the 2007 Voemel paper. The 4% accuracy figure relates to the tropical lower troposphere not the tropical lower stratosphere.

It is true, we have made a mistake, it is now corrected (L90).

(4) *Line 21* “They are shown to be good in the UT/LS ” is not a very scientific statement. What does “good” mean? How is it quantified?

We used « good » instead of “provides high quality profiles in the UT/LS”. It was quantified with the previous statement (“o 9% around the tropopause and about 10% at 28 km (Vömel et al., 2007).” P2-121) There was no need to repeat it, they have already been described as “most accurate” L89. So, the sentence is removed.

(5) *Line 22.* Statement is made “Nevertheless, the CFH are rarely launched on a routine basis at these stations mainly because of their cost.” What stations are referred to here? There are no stations discussed elsewhere in the paragraph.

This was referring to “atmospheric observatories” (but was no mentioning earlier) in general. It is rephrased (L91-L92).

(6) *Line 24* states that “the MLS (Microwave Limb Sounder) is very accurate (Read et al., 2007) and even the most efficient in the lower stratosphere”. Again I ask the question of what efficient refers to here since this claim is not referenced. But also, are the authors aware of the recent divergence in the lower stratosphere between MLS and frostpoint hygrometer that Hurst et al. have documented?

1. Atmos. Meas. Tech., 9, 4447–4457, 2016 www.atmos-meas-tech.net/9/4447/2016/doi:10.5194/amt-9-4447-2016

Regarding the use of the adjective “efficient” (referring to “provides high quality profiles in the UT/LS / most accurate”) P2-124, we didn't find intercomparison studies in the literature concluding that MLS was the most efficient in the lower stratosphere comparing to other satellites. So, we delete this part of the sentence. Then, we were not aware of the publication of Hurst et al. (2016). We add a sentence about it for the readers : “It is noteworthy, however, that Hurst et al. (2016) found some divergence between MLS and balloon-borne frost point hygrometers exceeding the accuracy of both instruments.” (L99 to L100).

(7) *Line 41.* Authors reference Sherlock et al. to support a statement of the importance of calibration stability. But the Sherlock paper presented an independent calibration technique which did not refer to another measurement of water vapor to calibrate the Raman lidar. Following line 41, the authors only discuss dependent calibration activities, i.e. ones where the Raman lidar calibration is derived from another

measurement of water vapor. Authors should include a discussion of both dependent and independent calibration techniques, such as documented in Venable et al which presented an alternative independent calibration technique which needs to be referenced in the discussion.

1. DD Venable et al. Appl Opt 50 (23), 4622-4632. 2011 Aug 10

The Sherlock paper introduces an independent calibration that is compared with the use of Vaisala sondes, which corresponds to an external water vapor measurement. There are statements referring to both kinds of method that deal with the stability of the calibration. That's why we have chosen this paper. We add a discussion on both dependent and independent calibration techniques and use the reference of Venable et al. (L115 to L117).

PAGE 3

(8) Line 1. Statement is made that all the Raman lidars of NDACC use Vaisala sondes to calibrate their database. This is not the case as the independent technique described in Venable et al. has been implemented and used in the ALVICE Raman lidar, a member of NDACC.

You are right. We specify that all the Raman lidars of the NDACC with a fixed location (L119) used Vaisala to calibrate their database. Mobile lidars provides different possibilities for the calibration, we mentioned the calibration of the ALVICE Raman lidar (L119-L121).

(9) Line 14. Statement is made “Some critical points have been addressed in the upgrade, including fluorescence, power and parallax effects, in order to optimize the configuration of the system (Hoareau et al., 2012; Sherlock et al., 1999b)”. I suggest some additional text to explain to the reader at least something about what these critical upgrades are. Since they are so important, the reader should have more explanation about them here without having to read the referenced papers.

Some explanations of these critical points are detailed in the Section 2.1 (P4-I4 to P4-I17). To avoid the reader to go straightly to the reference, we move the statement in the next section (L167) as an introduction of those explanations.

(10) Line 31. Statements are made “It is noteworthy that this Maïdo Raman water vapor lidar (called hereafter "Lidar1200") was recently provisionally affiliated within the NDACC. The conclusive affiliation occurs when absence of fluorescence and a stable calibration method are both demonstrated using validation campaigns involving frost-point hygrometer measurements.” It would be very informative to discuss what the calibration stability requirements are to meet the NDACC goals of process studies and trend monitoring. These two types of studies have very different calibration stability requirements and those should be detailed.

To our knowledge, there's no NDACC official requirement in term of stability for the calibration of water vapor measurements, it has to be formalized. It is true that knowing the scientific objectives of the NDACC, we can assess them. For the water vapor, they will be the same than those of GRUAN that are based on the GCOS report, we already develop them in the introduction (L61 to L83), see response to (1.1).

PAGE 4

(11) Line 13. Statement is made “A spectrometer is used directly after this telescope to separate the Raman and Rayleigh signals.” Traditionally the term spectrometer is used to refer to a grating spectrometer, which can be used to perform the task at hand. Here the authors are using a standard combination of beamsplitters and interference filters, which only becomes apparent later in the discussion. I suggest using a different term here such as “wavelength separation package” or some such term and quickly state that it consists of beamsplitters and interference filters.

You are right. We change the word « spectrometer » into « optical box unit » (L178) and « separation unit » (L180 and L184).

(12) Line 16. Statement is made “The overlap factor is identical for both channels”. Perhaps the two functions are quite similar, but surely they are not identical. Some quantification of how similar they are and

what the authors did to quantify it is needed here. Also realize that what we call the channel overlap function contains any position dependent optical efficiency variations in the beam splitters, interference filters and pmts. So it is highly likely that the overlap functions will have at least some small differences from one channel to the next.

Considering the geometry of the system, we expect the two functions to be quite similar. The individual comparisons between calibrated lidar profiles and RS92, M10 and CFH radiosounding profiles did not show any difference in the 2.2 – 8 km asl range that would look like being associated with a large uncertainty due to the overlap factor. Nevertheless, it is true that we didn't perform a test that undeniably concluded to a perfect similarity (please see changed L182). Thus, for the moment we arbitrary fix the uncertainty due to the overlap factor to 4% at the ground, decreasing up to the maximum signal recovery altitude. This uncertainty is added to the list of uncertainties (L241-L242). Some further tests will be performed in the near future.

(13) Line 24, section 2.2.1. The authors present Eq 1 as the “total absolute error” of the water vapor measurement, yet the formula presented is not appropriate to account for systematic uncertainties of which there are certainly several in the water vapor calibration. The authors are referred to an earlier publication which attempted to present the full uncertainty budget of the water vapor mixing ratio calculated by a Raman lidar involved in the MOHAVE 2009 campaign. See Appendix A3. It's particularly important to note that both the uncertainty of the calibration source and the uncertainty in transferring that calibration to the Raman lidar mixing ratio measurement are separate and important sources of uncertainty, both of which are systematic and not random. There are other uncertainties not considered by the authors as well.

1. Whiteman, D. N., Cadirola, M., Venable, D., Calhoun, M., Miloshevich, L., Vermeesch, K., Twigg, L., Dirisu, A., Hurst, D., Hall, E., Jordan, A., and Vömel, H.: Correction technique for Raman water vapor lidar signal-dependent bias and suitability for water vapor trend monitoring in the upper troposphere, Atmos. Meas. Tech., 5, 2893-2916, doi:10.5194/amt-5-2893-2012, 2012.

Please see response to (1.2) for our improvement of the new Section 2.2.2 (L187 to L257). We recall that we start freshly with the Lidar equation and then develop it (L192 to L211). Each term of potential systematic uncertainty is described, calculated and/or estimated (L215 to L242). When they are neglected, as for the aerosols, it is justified (L220 to L223). A figure is added to illustrate this section (the former left part of Figure 7 (P26) that has been upgraded): new Figure 1.

(14) In the discussion of uncertainty, the terms “uncertainty” and “error” are both used. It would be useful to clarify what the difference is that the authors are making. More traditional might be to use the terms “random uncertainty” and “systematic uncertainty” and not use the term “error”.

We turn the term “error” into “uncertainty” through the entire paper; in the text, the tables, the figures and in the captions. The term “error” is only used in new Table 1 because it is based on the GCOS report's table using “error”.

PAGE 5

(15) Line 3. The statement is made “Indeed, the effect of aerosols on Raman channels in the UV is low.” It's not fully clear what the authors mean by this statement but aerosol attenuation of the UV Raman signals under discussion is quite significant under turbid conditions. Errors in quantifying the aerosol extinction profile result in systematic uncertainties in the transmission profile. See the following publications:

- 1. Whiteman, David N., Examination of the traditional Raman lidar technique. II. Evaluating the ratios for water vapor and aerosols, Applied Optics, 42, No. 15, 2593-2608 (2003). See Fig 8.*
- 2. Veselovskii et al., Atmos. Meas. Tech., 8, 4111–4122, 2015 www.atmos-meas-tech.net/8/4111/2015/ doi:10.5194/amt-8-4111-2015 Figs 1, 2 etc.*

We were implicitly referring to the comparison between UV and visible but it is not the main point. The sky above the Maïdo Observatory (located above the boundary layer) is very clear, there are very few background aerosols and a few turbid conditions. We assume that the differential transmission due to aerosols is much smaller than the differential transmission due to molecules. The text is revised accordingly (L220 to L223). The atmosphere above Reunion Island might be occasionally characterized by larger concentrations of aerosols due to major volcanic eruptions in the Southern

Hemisphere or biomass burnings. A multiwavelength lidar system, recently set up at the Maïdo Observatory, should allow to estimate the aerosol content and thus the induced uncertainty in such one-time conditions.

(16) Line 11. It is notable and somewhat confusing in a calibration paper, such as this, which claims to be presenting the “absolute error” of the Raman lidar water vapor mixing ratio data product that, referring to the reference GPS IWV product, “The complete evaluation of the uncertainty will be further detailed in a future publication.” Knowledge of the total uncertainty of the GPS calibration source is needed to quantify the total uncertainty of the lidar calibration which depends on the GPS. Authors should consider this point and revise their discussion appropriately.

You are true. What’s more the future publication will be focused on the variability of water vapor more than being a technical description of the measurement. The sentence is removed. By checking the data, the uncertainty is more about 1 mm. The calculation of the uncertainty is consistent with Ning et al. (2013). What is also interesting to notice is the fact that, for the GNSS measurements, not all the sources of uncertainty are systematic. If on the three main uncertainties, i.e. ZHD, conversion factor Q and ZTD, the two first ones are systematic, the last one (and the main one) is a mixture of statistical and systematic uncertainties. It is now explained from L288 to L292 and the references to Ning et al. (2013) and (2016) is added.

(17) Line 14. Statement is made “Data are smoothed with a filter using the Blackman coefficients:...”. This statement sounds like it is meant to follow earlier ones that have introduced what kind of filter is used and other details about it. So without that earlier material, this sentence is confusing to the reader. Please detail what filter is used and what the Blackman coefficients are.

We misused the term “Blackman coefficients”, we were talking about the Blackman window. The window is then applied to a typical low-pass filter. The explanation is corrected and refined (L249).

(18) Line 30. “the signal traveling between a GPS satellite (altitude of 20,200 km) and a ground-based receiver is delayed by atmospheric constituents (dry air, and water vapor) ...” Other atmospheric constituents such as clouds, hydrometeors, aerosols also influence the propagation of the microwave signals associated with GPS. The temperature profile of the atmosphere is also important for determining the signal delay. Please revise text.

The text has is revised (accordingly to your comment): “is delayed by atmospheric constituents (dry air, water vapor, clouds, hydrometeors and aerosols) and is affected by the temperature.” (L267-L268).

PAGE 7

(19) Line 3. “If no instrumental change occurs, the calibration coefficient is supposed to be almost constant.” The variability of the calibration constant is a subject of this investigation. The calibration value has certain variability which the authors are in the process of quantifying. I suggest removing statements like these and replace them with statements that provide quantities with uncertainties that specify variability based on their data analysis.

As suggested, the sentence is deleted. However, we remind the main reason of the empirical change of the calibration coefficient in order to explain why we use lamp measurements : “For the near-real time treatment, it remains necessary to check if there are instrumental changes because they might alter the calibration coefficient” (L328-L329). The variability of the calibration coefficient is further investigated in this Sect. 2.3.3 and especially L373 to L383 and L420 to L423.

(20) Paragraph starting line 5. This paragraph reads as if the lidar is calibrated each night using the nightly calibration coefficient. Figure 2 seems to contradict that concept so the discussion is confusing. There is a significant detail that bares mentioning in how frequently the lidar is re-calibrated. As mentioned earlier, the uncertainty due to transferring the calibration of the GPS to the Raman lidar entails a systematic uncertainty. Previous field campaign research shows that this systematic uncertainty is typically 2-5% depending on the particular experiment. Given that time series of lower stratospheric water vapor are desired to be stable at better than 0.1 ppm level per year (~2%) (GCOS requirement), the systematic uncertainty associated with the transfer of a calibration from another instrument to the Raman lidar water vapor mixing ratio data product,

by itself without considering any other sources of uncertainty, can be sufficient to make the time series of greatly reduced value or even useless for lower stratospheric trend detection. This statement is true assuming, as is usually the case, that the calibration coefficient is determined only infrequently (every several months or perhaps once per year) based on ensemble comparisons with radiosondes and then carried forward using lamp-based measurements. The way to address this weakness of the dependent calibration technique is to perform the dependent calibration frequently enough such that the uncertainty associated with the calibration transfer process becomes part of the random uncertainty budget instead of being part of the systematic uncertainty budget. This is the technique that the ARM Raman lidar has followed since its inception by using the technique of a running calibration with respect to microwave radiometer and one that the authors might be able to use here if quality GPS calibrations are available on a daily basis. While lower stratospheric water vapor measurements do not benefit from an increase in random uncertainty, at least random uncertainty can be reduced by making additional measurements. Introducing 2-5% systematic uncertainties in a time series by, for example annual changes in the calibration coefficient would prevent trends at the 1% per year (which is a common estimate of the magnitude of the water vapor trend in the UT/LS) to be determined. The authors are referred to the following publication which states the importance of converting sources of systematic uncertainty to random uncertainty when possible. The point should be made in this paragraph that the daily determination of the calibration coefficient by comparison with GPS turns the systematic uncertainty associated with the transfer of calibration from GPS into a component of the random uncertainty budget.

I. Whiteman, D. N., K. C. Vermeesch, L. D. Oman, and E. C. Weatherhead (2011), The relative importance of random error and observation frequency in detecting trends in upper tropospheric water vapor, J. Geophys. Res., 116, D21118, doi:10.1029/2011JD016610.

The lidar is not re-calibrated each night. In fact, we use the temporal series of the hourly coefficients in order to visualize the instrumental quasi-stationary periods (IQSPs) of the calibration coefficient. In our case, we have identified instrumental changes (P7-I32) that can alter the coefficient, by crossing the temporal series of the hourly coefficient in which IQSPs appeared visually, with the logbook detailing all the technical operations that has been made on the system. There is a single calibration coefficient calculated for all the measurements of a same IQSP. Our description of the methodology appeared to lack of clarity. We have restructured and rewritten it (L324 to L353). Even if we knew it and that it belongs to the main reason why we have chosen the GNSS to calibrate the lidar, it is true that we didn't formally explain that the use of the hourly GNSS IWV database has been chosen in order to randomize the process. We thank the reviewer and we add his valuable explanation (with the reference) on the importance of turning the systematic uncertainty associated with the transfer of the calibration into a statistical component in the Sect. 2.3.6 (Total uncertainty of the lidar calibration) (L450 to L479). This is also included in the introduction (when discussing about the measurement requirements regarding the detection of trends) based on the reference and on the GCOS report, see response to (1.1).

(21) Lines 7-8. reference is made to 1-hr time windows for lidar integration but the example times cover 55 minutes. Please reconcile.

You are right, it is a mistake. The lidar data are integrated between 17:31 and 18:30 LT, it is corrected L328.

(22) Figure 2. There are several very significant changes in calibration based on the GPS measurements that seem not to be identified by the lamp measurements. Is this the case? Why don't the lamp measurements identify these large changes in calibration coefficient? What is the explanation for this? Authors should be aware of several "failure modes" of the lamp-based technique that are described in the reference below. Is one of those failure modes in play here?

I. Whiteman, D. N., D. Venable, E. Landulfo, Comments on "Accuracy of Raman lidar water vapor calibration and its applicability to long-term measurements", Applied Optics Vol. 50, Iss. 15, pp. 2170–2176 (2011)

As we had indicated P7-I18, the lamp measurement is restricted to the identification of changes in the reception part of the system. We forgot to quote the reference of Whiteman et al. (2011) that introduces more precisely the reasons for using other techniques to detect alteration of the calibration coefficient. The reference is added (L337). In our case, we have identified the major instrumental changes (P7-I32) that can alter the coefficient. It has been made by crossing the temporal series of the hourly

coefficients (in which IQSPs appeared visually) with the logbook detailing all the technical operations that has been made on the system.

(23) Table 1 presents the calibrations used during the quasi-stationary periods. Please define more clearly the difference between “absolute” and “relative” error and, again, the term “uncertainty” is preferred over “error”. If the relative error that the authors refer to is the uncertainty of the transfer of the calibration from the GPS to the lidar, then this needs to be acknowledged as a systematic uncertainty in the time series that is introduced each time the re-calibration is done.

« Absolute error » was referring to σ and « relative error » to the standard error i.e. $\frac{\sigma}{\sqrt{n}}$. The terms were not appropriate. We change the categories of this table into : IQSP number; Dates; Calibration coefficient (i.e. C_{mean}); σ ; standard uncertainty (i.e. $\frac{\sigma \cdot 100}{\sqrt{n} \cdot C_{\text{mean}}}$)(%). Please see the discussion about the uncertainty in the transfer of the calibration from the GNSS to the lidar in response to (20). Table 1 (P35) is updated after the correction of the saturation on the whole dataset and according to the new categories (new Table 2).

PAGE 8

Regarding the unconvincing result in Sect. 2.3.4 which aims at intercomparing the GNSS and other devices to calibrate the profiles, we have decided to improve our methodology of calibration of the Raman lidar by the sondes. In the first version of our article, we took a single altitude range (between 3 and 4 km asl) and then we make the ratio between CFH/RS92 or M10 profiles with the uncalibrated lidar data to get the calibration coefficient. We started again by using the “radiosonde-based calibration technique” in Whiteman et al. (2012) (appendix A2). This methodology (and especially the slight differences with our adaptation) is described L400 to L409.

(24) Lines 10-11. Please provide the standard deviation of the derived calibration coefficients. Statement is made that the Vaisala sondes have a known dry bias and the authors use this statement in reference to the RS41 in addition to the RS92. The RS41 was not studied in the Miloshevich and Bock works cited, however, so it is not a proper reference. The apparent good agreement shown by the authors between RS41 and RS92 in the 3-4 km range and the recent measurement campaigns of the RS41 showing very good performance of the new instrument would tend to indicate that both RS92 and RS41 sensors were performing very well during the campaign in the required 3-4 km range. This contrasts with the authors claim of dry bias.

The standard deviation of the derived calibration coefficients is added to the mean values L414-L415. Actually, the reference was limited to the RS92. We spoke about the dry bias because it was an hypothesis to explain why the mean calibration coefficient was lower (and not higher) than the ones based on GNSS IWV and on the CFH data. To respond to your statement about the RS91: we are not aware of the results of comparisons between the RS41 and other reference instruments for water vapor measurements; we didn't have the papers. We read the paper of Jensen et al. (2016) but its experiment have been led under “typical midlatitude continental summertime conditions”. Regarding sondes in general, it has been previously shown (in different experiments) that their accuracy can be less important in tropical areas comparing to temperate conditions. The paper of Kawai et al. (2017) offers comparisons in the tropics but it is limited to RS41 vs RS92 comparisons, they didn't extend the experiment to other instruments or satellite that might be considered as (an) other(s) reference instrument(s). We couldn't conclude (or deny) on the fact that the RS92 and RS41 sensors were performing very well in the required 3-4 km based on these results. Nevertheless, considering that with the new methodology, the coefficient given by the RS92 is higher, the potential dry bias mentioned in the literature could not explain the difference. We delete the RS41 results because the aim of the paper is not discussing of those interesting new results between the two types of Vaisala sondes that should be further studied in a dedicated investigation.

1. Jensen et al. (2016) “Comparison of Vaisala radiosondes RS41 and RS92 at the ARM Southern Great Plains site,” *Atmos. Meas. Tech.*, vol. 9, no. 7, 3115–3129.
2. Kawai et al. (2017) “Comparison of Vaisala radiosondes RS41 and RS92 launched over the oceans from the Arctic to the tropics,” *Atmos. Meas. Tech.*, vol. 10, no. 7, 2485–2498.

(25) Line 13. Lidar1200 calibration using the “routine method of calibration” is given as 155+/-32. What is the “routine method” and how do the authors reconcile a 21% calibration uncertainty figure given here with the much smaller calibration uncertainty values shown in Table 1?

The term “routine methodology” is not really well chosen, it was referring to the methodology described in Section 2.3.3 (and Table 1 - P35), it corresponds to the calibration coefficient associated with the IQSP including the MORGANE campaign (cf Table 1 – P35 / new Table 2). Regarding the uncertainty, it was still about the calculation of the uncertainty divided by the square of the number of hourly coefficients in the QS period discussed later in this report (29). That number « 32 » corresponded to σ , i.e. the variability from one day to another inside the period n°9/11 (Table 1 - P35/ see new Table 2). With the newly corrected database, σ is now of 15. It is updated L414.

(26) Line 17. “Thus, it is confirmed here that the GNSS technique is as suitable as radiosoundings for the calibration of the water vapor profiles of the Lidar1200. “ None of the calibrations shown here seems to offer as stable a calibration value as shown in Table 1. Also, were the lamp measurements useful in quantify the same large changes in calibration coefficient shown in Figure 3?

In hindsight, we agree that the firsts results were not convincing. That's why we have decided to refine our methodology of calibration based on the sondes. You will see that now the methods are in agreement between 0.1 and 5% according to the type of sondes (L416-L417, new Fig. 4). The lamp measurement shows a small instability during the MORGANE campaign. The variability of the calibration coefficient is discussed, there are three main hypotheses: 1) the high variability of the water vapor and 2) small instrumental instability as mentioned by David et al. (2017) and 3) influence of the uncertainties of the different measurement techniques. It is true that using the methodology of Whiteman et al. (2012) should prevent from the influence of the atmospheric variability when comparing both techniques to calibrate the lidar. Even if there is a small instrumental instabilities, it might not be seen at a so short time scale. This is discussed L373 to L383 and L420 to L423.

(27) Section 2.3.5. Authors state that IWV comparisons with CFH show an uncertainty in the GPS IWV measurements of 1-2 mm. From Figure 4 it appears that as much as 50% of the time, the IWV at the site is 10 mm or less. Therefore, from the authors estimates, the GPS IWV uncertainty would seem to be 10-20% during approximately half of the measurement periods. Is this calibration uncertainty acceptable for use as the calibration source for measurements to be used within NDACC?

It is true that, regarding Fig. 4 - P23, more than 50% of the time, the IWV above Reunion Island is less than 10 mm. Thus, the uncertainty of the GNSS is high. That's why the methodology to choose the outliers for the calculation of the calibration coefficient of each IQSP has been refined, please see response to (20). When the IWV < 7 mm are not taken into account, the mean uncertainty on the GNSS measurement is around 9.5%. We would like to improve the temporal resolution of the ZTD that we are using to calculate the GNSS IWV. There should be enough IWV data to set the threshold to 10 mm which would correspond to an uncertainty on the GNSS measurement of about 7.5%. What's more, this value includes both systematic and random errors. The systematic error on the calibration associated with the GNSS measurement would probably be less than 7.5%. According Ning et al. (2013, 2016), the systematic part shouldn't represent more than 50% of this uncertainty. Thus, as an external source of measurement, the GNSS would imply a systematic uncertainty less than 7% and the statistical part should be minimized by the use of large samples.

The authors state that the lamp measurements can be used to carry the calibration forward during these dry periods, but the lamp measurement results shown in Figure 2 do not seem to show sensitivity to some of the large calibration variations that occur so can the authors really rely on the lamp to carry the calibration forward? And getting back to one of the main questions, authors need to discuss what the calibration accuracy and stability requirements are for water vapor data to be useful for both process studies and trend detection. The former has considerably more stringent accuracy and stability requirements than the latter and that would be useful for the authors to detail here since, given the above considerations, there could be broad skepticism among readers about these measurements being suitable for trend studies.

It might happens that there wouldn't be a sufficient number of GNSS IWV > 5mm. In such conditions, especially in winter, the GNSS can fail in calibrating our measurements. Thus, we have to be aware of the instrumental changes that might alter the calibration coefficient. We use the lamp measurement

but not only. As described in Sect. 2.3.3, we also use our list of identified changes of the system that might alter the calibration coefficient. We just forgot to repeat it in this section. It is now done L443-L444. They can both warn us on the interest of launching a radiosonde to be sure that we would be able to recalculate the calibration coefficient. For the discussion on the calibration accuracy and stability requirement for the different mentioned scientific objectives, please see response to (20).

(28) Section 2.3.6. The authors discuss here what I have referred to as the uncertainty in the transfer of the calibration coefficient from the external measurement of water vapor to the Raman lidar water vapor profile. The authors have correctly identified, by the title of this section, that this is a source of systematic uncertainty in the total error budget and so cannot be propagated as they have shown in Eq. 1. Please reconcile.

We had not separated the two uncertainties: the transfer of the calibration from the GNSS to the lidar (the one we were discussing in Sect. 2.3.6 and taking into account in our calculation of the total uncertainty) and the uncertainty of the source of the external water vapor measurement, i.e. GNSS. This latter uncertainty was discussed but not included in our calculation, we had wrongly considered the uncertainty in the transfer of the calibration as the total uncertainty of the calibration. We correct that, it is explained in Sect. 2.3.6 (L456 to L458). We also differentiate the systematic and statistical components of each source of the total uncertainty on the calibration, please see response to (20). We have checked the propagation of the uncertainty based on Leblanc et al. (2006), our propagation was correct but only until the filtering. After the filtering, we were not separating the correlated from the uncorrelated information. It is true that considering the nature of the uncertainties for the calibration, they should be propagated differently, we reconcile it. Eq. 1 is developed as requested in (1.2).

(29) Line 34 “If the calibration is considered as stationary and only due to random fluctuations, the uncertainty on the calibration coefficient of each period is mainly due to the term corresponding to the standard deviation divided by the square of the number of nightly calibration coefficients.” The authors need to be clear that the value they are considering here is the uncertainty in the transfer of the calibration from GPS to lidar and does not consider the uncertainty in the GPS calibration itself. The variability in this calibration transfer coefficient is surely influenced by, perhaps dominated by, atmospheric variability since the GPS is sampling a large volume whereas the lidar is sampling just the atmosphere directly overhead. Also, the atmospheric conditions are different each time the calibration transfer is done. So the assumption that the variation in the calibration transfer coefficient is only due to random fluctuations is not in general satisfied. Thus it is not correct to divide by the square root (not square as stated by authors) number of samples in calculating the uncertainty in this transfer of calibration. The more conservative way to perform this calculation, so as to specify an upper bound to the uncertainty in the transfer of the calibration, is to simply use the standard deviation (and not standard error) as the uncertainty for the transfer of calibration and not divide by the $\text{Sqrt}[N]$ term. The real uncertainty is likely somewhere in between the standard deviation and the standard error, but is surely larger than the standard error that the authors have used.

1. It should be noted that the uncertainty of the calibration of the external IWV measurement itself needs also to be factored into the total uncertainty budget of the calibration. That also is a source of systematic and not random uncertainty.

We suppose here that the fluctuations of the calibration coefficient are independent of the variability of the atmosphere. The calculation of the calibration coefficients is based on different periods with a large sampling: the uncertainty corresponds to type A (JCGM, 2008) and is calculated as the standard deviation divided by the square of the number of hourly coefficients. Regarding note 1, please see our response to (28).

1. JCGM 100: Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement. Technical Report, JCGM, 2008.

PAGE 9

(30) Line 25. “With regard to relative humidity, it is recognized by many that the CFH sondes are among the most accurate especially in the UT/LS (Vömel et al., 2007).” There are two dominant cryogenic frostpoint hygrometer instruments in the world currently, the CFH and the NOAA FPH. They are considered

comparable in performance. So the statement cited is not correct.

1. Atmos. Meas. Tech., 9, 4295–4310, 2016 www.atmos-meas-tech.net/9/4295/2016/doi:10.5194/amt-9-4295-2016

You are right. We change the statement to explain they are both considered comparable in performance (L512 to-L514), and the reference is added (L514).

(31) *Section 3.3.* Given the highly variable signal to noise of the lidar versus the relatively constant one of frostpoint, the technique described for comparing the two instruments is quite reasonable. One thing that is puzzling, though, is why such a high power, large aperture lidar system requires such long averaging time (48 hours) to produce a quality profile extending beyond 20km.

The 48-hours averaging is in agreement with Leblanc et al. (2012) : “With the assumption that lidar data contains random noise following a Poisson distribution, using the one to two CFH profiles yields a precision in the UTLS equivalent to integrating the lidar measurements for 5 to 8 full nights (i.e., 40 to 70h).”. We add this reference to the paper (L664).

1. Leblanc, T., McDermid, I. S., and Walsh, T. D.: Ground-based water vapor raman lidar measurements up to the upper troposphere and lower stratosphere for long-term monitoring, Atmos. Meas. Tech., vol. 5, no. 1, 17–36, 2012.

Previous research shows Raman water vapor lidar measurements extending to these altitudes with lower laser power (16 vs 24W), smaller telescope (0.6 vs 1.2m) and shorter averaging time (9 vs 48 hr). In considering the relative signal to noise of these two measurement examples, the difference must be in the noise term instead of the signal term. For the higher performance measurements, the noise was reduced by use of a 0.25 mrad field of view, 0.25 nm interference filter and thermo-electrically cooled water vapor PMT. The authors may want to consider whether further optimizations of the noise term would reduce the averaging time required to probe the lower stratosphere. It should be noted that in the MOHAVE measurements cited below no indication of fluorescence was present and agreement with CFH and climatology was very good. See Fig 13 from the reference below.

1. Whiteman, David N., Kurt Rush, Scott Rabenhorst, Wayne Welch, Martin Cadirola, Gerry McIntire, Felicita Russo, Mariana Adam, Demetrius Venable and Rasheen Connell, Igor Veselovskii, Ricardo Forno, Bernd Mielke and Bernhard Stein, Thierry Leblanc and Stuart McDermid, Holger Vömel, Airborne and Ground-based measurements using a High-Performance Raman Lidar, doi:10.1175/2010JTECHA1391.1 (2010).

Your questions are legitimate. Nevertheless, regarding the measurement in Whiteman et al. (2010), the results are based on a single profile for a measurement made in optimal conditions (new moon), no systematic uncertainty is calculated and it seems that it is not a standard calibration. No relative difference with the CFH is shown, it is just indicated that it is “generally with +-10 % through the extent of the CFH profile”, whereas the balloon burst at 19 km. As it is well indicated in the paper “A single profile comparison is not sufficient for “validation” of any instrument”. With 9 hours, the Lidar1200 measurements can also reach UT/LS and even the lower stratosphere with a maximum vertical resolution of 1.3 km but obviously the total uncertainty is very large for the Lidar1200 (see Section 4.2) as it might be for the ALVICE in Whiteman et al. (2010). More broadly, the local context of (sub)tropical environment of the Lidar1200 has to be taken into account in comparing the performance of different lidars, considering the challenge of calibrating profiles with very small water vapor contents.

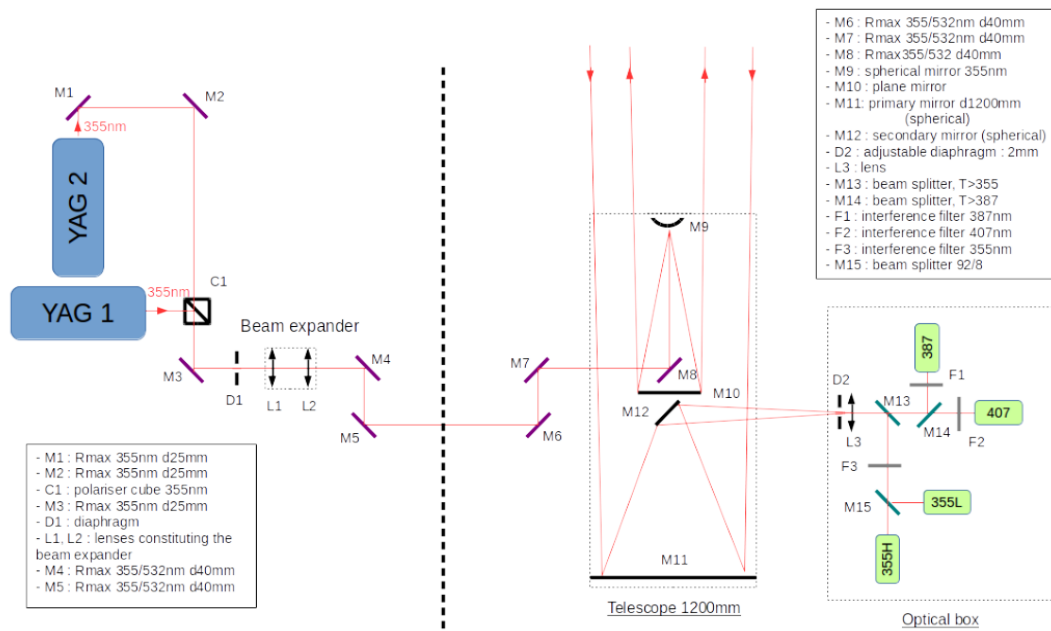
Our FOV of 2 mm was chosen because it was a good compromise between both temperature and water vapor measurements. It should be rethought in the future due to the recent instrumental upgrades of the lidar of the Maïdo Observatory. Indeed, a little telescope and a new channel with its own FOV was recently added for the lowest layers. The half-height width of the effective interference filter is 0.3 nm, reducing it might increase the temperature-dependence to height. Regarding the cooling of the PMT, a chiller was bought, we plan to test it to evaluate its efficiency. We recently identified two potential upgrades of the Lidar1200:

- 1) Change of the 407 nm detector that is still the one from the old version of the instrument. The technological improvement since then should give it a higher efficiency.**
- 2) Change of one of the mirror, the number 10 (M10 on the scheme below). It has been coated for both UV and visible, we would change into a specific coating to UV. The gain on the signal is**

estimated to 20%.

The further optimizations of the system are now detailed in the perspectives in Sect. 6 (L849 to L859).

LIDAR 1200 OPTICAL SCHEME



(32) Line 22. Referring to the comparisons shown in Figure 6, statement is made “No positive or negative bias appears.” By contrast the figures seem to indicate clear biases between lidar and CFH in the 3-5 km range that sometimes exceed 20%. Please reconcile. The later statements in the paragraph relating to the case of 19 May may offer an explanation for the biases, but this can be checked by performing the comparisons excluding the 19 May case. But certainly the quoted statement is incorrect and needs to be changed.

It is true that the biases could not be fully explained by the 19 May, this sentence is deleted. We further investigated it: we found in a first time that we had forgotten a correction in the data processing. Our initial hypothesis that the N₂ channel was not saturated in the lowest layer was false. A correction of the saturation is applied to the whole dataset (L198 to L205). Nevertheless, there still are a difference in all the individual comparisons: a negative one at 5 km and a positive one at 3 km, this is mentioned L554. We can see both on each individual comparisons during MORGANE. The peak at 5 km corresponds to a significant gradient of water vapor around this altitude that is detected slightly differently in altitude regarding both instruments. The main reason is the filtering of the lidar data. The explanation is added to the paper L557 to L562.

(33) Line 38. “There is no obvious reason to explain this bias”. Agreed. It could be an indication of PMT signal induced noise with just the right decay constant, but that is just speculation.

We agree, it is one of the hypotheses. We will try to further investigate it in the future with more intercomparison experiments. This bias might be explained by the large spatiotemporal variability of the water vapor which can be more important with the altitude. Regarding the variability of the water vapor in the UT, the integration of the lidar on several hours and the spatiotemporal difference in the measurement, the comparison is difficult. Even if there is a bias, it doesn’t question the quality of the data, it might be a warning on the intercomparison methodology. More comparisons should bring new explanations on the peaks at 3 km and the bias in the UT.

PAGE 11

(34) Line 2. “To conclude, the Lidar1200 and the CFH profiles are in a good agreement in the whole region of the troposphere sampled by Lidar1200, and the MORGANE campaign profiles have been validated by the CFH sondes up to 22 km asl.” This statement is not consistent with the significant biases in the altitude

ranges of 3-5 and 14-16 km. Please reconcile.

The saturation is corrected on the whole dataset and the methodology of calibration refined. It induces slight changes on the results of the comparison. The mean relative difference for the 2,2-14, 14-17 and 17-22 km asl are 8.8, 11.4 and 4.5%, respectively. Considering both instrument uncertainties, we can say that both techniques are in good agreement. It is true that there is a disagreement at two altitudes (found on individual comparisons): around 3 and 5 km. In the lowest layers, the presence of these peaks might be explained by potential persistent geophysical structures seen differently by both instruments because of the filtering of the data. The peak at 3 km asl needs to be further investigated. It seems to have a bias of around 10% on the whole 14-17km partial column.

Regarding the 14-16 km, there is a bias but it is included in the sum of the uncertainties of both instruments. The standard deviation of the CFH in the right panel of figures doesn't correspond to the uncertainty of the CFH but to the variability of the 5 CFH profiles. Thus, we change the right panel of Figure 6, the standard deviation is substituted by the CFH uncertainty (black dotted line), the lidar uncertainty (blue dotted line) and the sum of the CFH and lidar uncertainties (pink dotted line). The results are updated and the presence of biases is discussed (L554 to L583). The conclusion is also refined on this specific point (L812-819).

(35) *Figure 7.* See earlier discussion about errors and uncertainties. There are contributions not considered here and I do not believe that the authors have properly calculated the total calibration uncertainty.

Please see response to (1.2) and (13). Please note that Figure 7 - P26 is split into two parts: new Fig. 1 and new Fig. 7.

(36) *Section 4.1.* There is much discussion here of how long an averaging time is required to reach what altitude. But such discussion needs to be based on what the measurement requirements are for certain types of studies and, as mentioned before, that material is lacking. Authors need to add an early important section about what random and total uncertainty are acceptable for the types of analyses (e.g. process studies, trend detection) they want to do with the data from this instrument.

Please see response to (1.1), the paragraph is added to the introduction (L61 to L83) to detail the requirements on the measurement uncertainties and on the calibration stability for addressing different scientific objectives. It is also discussed in the conclusion L819 to L830.

(37) *Section 4.2* Some introduction to this section is needed that explains what the authors mean by optimal performance. As it stands it is not really clear what the intent of this section is.

We intend to determine the ability of the lidar to reach the lower stratosphere, i.e. the maximum range of the measurement limited to one night in a first time and, then, to several nights. The title of the section is turned into: Maximum altitude range (L632), and an introductory sentence added L633-L634.

(38) *Figure 8.* I am not sure I see the value of including this figure. I think the authors can discuss these data without showing this figure.

We agree with you, this figure is deleted.

PAGE 12

(39) *Line 17.* "The profile of 24 September 2015 integrated on the all-night measurements reaches approximately 18 to 19 km altitude with an uncertainty of the order of magnitude of 35% (4-5 ppmv, Fig. 8)." I doubt that 35% uncertainty corresponds to 4-5 ppm in the lower stratosphere based on climatology or MLS. A value closer to 2 ppmv seems more reasonable. It might help to show other measurements/climatology for comparison.

With this specific measurement, the calculation of the uncertainty in our previous version led to 3.7 ppmv (35%) and 4.7 ppmv (37%) at 18 and 19 km asl, respectively. The calculation of the total uncertainty budget (as the analyses) is updated. Now, the uncertainty is of 4 ppmv (39%) and 5 ppmv (38%), respectively. Nevertheless, it might indicate that the profile range is lower. At 16 km asl, the

uncertainty is of 2.5 ppmv (33%), which is more in agreement with your comment. Comparisons with satellites is a work in progress. In this paper, the CFH sondes endorse the role of the reference instrument.

(40) *Line 27.* “The upper limit is the altitude where the lidar uncertainty corresponds to twice the variability of the water vapor in the lower stratosphere.” Different values of “variability” are considered but they do not seem to be the result of an analysis. So variability needs to be defined here and how it is calculated must be presented.

The choice of the different values of variability are explained (P12-L27 to L31; L649 to L652). If some variability of 1 ppmv can be found in the literature, the weekly and even the monthly variability of the water vapor in the lower stratosphere is still a topic of investigation, that's why two values had been chosen empirically, we restrict this value to 1 ppmv.

(41) *Figure 9.* The first three plots show unreasonable values in the upper several km. The authors are advised to overlay MLS (even with the recent divergence issue discussed in Hurst et al. It will be useful since the divergence is measured in tenths of ppmv) or climatological data for illustration of this. Only the 4 th plot seems to have a reasonable behavior throughout its range. An overlay of other LS data onto these plots would make this point.

We agree, it is fully detailed in Sect. 4.2 – P12-P13. For the use of the MLS data, please see response to (39).

(42) *Figure 10.* Please see earlier discussions concerning sources of uncertainty and their calculation.

The figure 10 (now Figure 9) is updated according to the new total uncertainty calculation. At those altitudes, it especially takes into account the temperature-dependence uncertainty. In order to avoid any misunderstanding (referring to (23)), the terms “absolute” and “relative” are removed from the figure and the caption.

(43) *Section 4.3.* Authors discuss the influence of varying averaging times on revealing fine structures in the atmosphere. These are good illustrations of the need for algorithms that automatically make best use of the information content of the data. Other methods of processing the water vapor data such as adaptive or optimal estimation algorithms have been presented before that permit these kind of structures to be optimally revealed in an operational way. See relevant publications below:

1. See Section A.4 of previously mentioned Whiteman, D. N. et al. : *Correction technique for Raman water vapor lidar signal-dependent bias and suitability for water vapor trend monitoring in the upper troposphere*, *Atmos. Meas. Tech.*, 5, 2893-2916, doi:10.5194/amt-5-2893-2012, 2012.
2. Sica RJ, Haeferle A., *Retrieval of water vapor mixing ratio from a multiple channel Raman-scatter lidar using an optimal estimation method.*, *Appl Opt.* 2016 Feb 1;55(4):763-77. doi: 10.1364/AO.55.000763.

You are right, we will work on it, we add a sentence in the conclusion to explain it: “...adaptive or Optimal Estimation Method (OEM) algorithms could be used in an operational way in order to reveal fine-scale structures (Sica and Haeferle, 2016; Whiteman et al., 2012)...” (L855 to L857).

(44) *Section 5.2.* How does the seasonal cycle captured by the lidar measurements compare with space-borne measurements (e.g. MLS) or climatology? Please discuss.

The lidar temporal series of monthly profiles was already compared with the climatology of Hoareau et al. (2012) in Sect. 5.2 – P14 – P15. In response to your comment, we refined the comparative study (L746-L747): in the first version of our paper, we were not using the definition of Hoareau et al. (2012) of austral summer and winter, it is done in the new version. The results are updated (L744 to L749).

(45) *Section 6.* “The spatio-temporal variability of the water vapor is not well documented through direct observations.” The near global measurements of space-borne sensors such as MLS and others contradict this statement. Please reconcile.

We clarify our statement and add the references motivating it as follows: “The spatiotemporal

variability of the water vapor in the UT/LS is not well established through direct observations (Kunz et al., 2013; Müller et al., 2016). To fully understand the water vapor variability in the atmosphere, it is necessary... (L758-L759).

ACKNOWLEDGEMENTS

(46) “The lidar data used in this publication were obtained as part of the Network for the Detection of Atmospheric Composition Change (NDACC) and level 2 product as daily vertical water vapor profiles will be publicly available through the NDACC portal (<http://www.ndacc.org>) and the French atmospheric data portal (<http://www.pole-ether.fr/>).” Authors provide 2 sites where water vapor data may be downloaded, but neither seems to have data from this system. How can one access the measurements?

As indicated in the acknowledgments, the data « will be » soon available on line. For the NDACC, the free access to the Raman lidar measurements is subject to the establishment of a consensus on the format of the data processing and the data themselves. We are not aware of any deadline for that, a NDACC working group is surely working on it. It's quite the same situation for Ether. Nevertheless, the data are available (after validation) on the local website, the url of the website has been added: “The raw data and other products of the Lidar1200 water vapor data are available at the OPAR web portal (<https://opar.univ-reunion.fr/>)” (L869), the login and the password to download the data are available on demand by contacting us, osureunion-informatique@univ-reunion.fr or franck.gabarrot@univ-reunion.fr.