Response to David Whiteman

1. L1 – statement is made that "Lidars are well-suited for trend measurements in the upper troposphere and lower stratosphere, particularly for species such as water vapor."

1. The measurement requirements for detection of trends in water vapor differ dramatically between the UT and the LS. Paragraph 16 in Whiteman et al, 2011b and the first several paragraphs of the discussion section of Whiteman et al, 2012 detail the argument that Raman water lidar is much better suited to trend detection in the upper troposphere than the lower stratosphere. Also, I might suggest that instead of just saying "Lidars" here, to specify "Carefully calibrated and quality-controlled Raman Lidars..."

Thank you for pointing out the results from these two papers. We have changed the first sentence to include your suggested text.

Regarding UT vs LS, this work is targeted at both current and future lidar systems, which we anticipate/hope will have the capability to measure routinely in the LS. However, we agree that reaching the UT is more feasible than reaching the LS.

2. L4 and beyond – the current technique is improved with respect to the traditional technique but no comparisons are done with respect to other "improved" techniques. It is my hope that we can address that in follow-up research.

We will be comparing this method with another calibration technique in a follow-up paper which discusses the processing of the RALMO data for trend analysis and combining the radiosonde technique with another independent long-term calibration method. We felt that it was not appropriate to put this comparison in this paper since the trajectory calibration can stand alone and we did not want to move focus away from the GRUAN radiosonde calibrations.

3. L5 – Whiteman et al (2006) is cited for a "track-sonde" technique that was used. It is worth noting, however, that the track-sonde technique as used in 2006 did not perform as well as the more simple variable temporal-spatial smoothing routine described in that same publication. More importantly, a significantly more sophisticated technique for performing radiosonde calibration was presented in Whiteman et al, 2012. It does not explicitly track the sonde but the geometrical similarity requirements imposed in that routine, I expect, achieve some of the same collocation benefit that is discussed in the authors' technique. These details should be mentioned.

We agree that using the method in 2012 should have a similar effect, however, we have separated the 2012 method into a separate category which is discussed in the new traditional section along with Dionisi et al. 2010. We have mentioned your 2012 paper and its goals in the new Section 3.2 . We have added the points you made about the tracking technique in the 2006 paper being superseded with the 2012 study to the discussion.

4. L29- "paralyzation" → "paralysis"

Thank you for catching this.

Introduction Comments:

5. Statement is made that "instruments with high spatial-temporal resolutions, such as lidars, are uniquely suited to long-term stratospheric and tropospheric water vapour studies". For lidars to provide a good signal-to-noise measurement in the UTLS requires significant temporal and spatial smoothing. So I do not agree that high spatial-temporal resolution measurements make lidars uniquely suited to long-term UTLS studies of water vapor since the temporal and spatial resolution must be degraded to achieve an acceptable S/N in the UTLS.

We agree that the sentence is a little misleading and to prevent confusion, we think it's best to just delete the sentence. Therefore we have removed it from the paper.

6. L6 - "of" → "from"

We have changed this as suggested.

7. Lines 7-9. Statement is made "Lidar measurements are particularly useful for creating statistically significant water vapor trends of the UT and LS region … " and Weatherhead, 1998 and Whiteman 2011b are used to support the claim. I don't believe that Weatherhead et al makes any statement about the suitability of lidar for this task. Also, as stated above, Whiteman et al 2011b expresses doubt that Raman lidar would be suitable for LS trend detection; a claim that is amplified in Whiteman et al, 2012. So I would suggest a statement such as "Carefully processed, stably calibrated Lidar measurements can be particularly useful for creating statistically significant water vapor trends in the UT region ..."

We had included Weatherhead 1998 because it discusses the uncertainty thresholds necessary to obtain statistically significant trends, which we thought was relevant here. However, it does not discuss lidars specifically. We have removed the reference.
We have incorporated your suggested comment in to the paper but we have not removed the LS from the paper for the reasons we have discussed previously.

8. Paragraph starting "Internal calibration techniques …

1. reference is made to Venable et al, 2011 as an example of the white light technique, which is correct. The next sentences, however, refer to the limitation of using a single lamp and the need for multiple lamps or a scanning technique. This is confusing since Venable et al showed the utility of the scanned lamp technique so that work does not suffer from the limitations of the single lamp technique as implied by the current discussion. I suggest revising the paragraph so that the first reference cited is one that makes use of a single lamp.

Thank you for pointing this out. Venable et al.2011 was indeed not the appropriate first reference. In fact, the first reference should have been Leblanc et al. 2008. We have made the change and do not think the paragraph needs to be changed as it now matches the reference.

2. Later in the same paragraph it is stated that the uncertainty in the knowledge of the ratio of the Raman cross sections is 10% from Penny and Lapp, 1976. The work of Avila et al, 2004 and Venable, 2011 however point toward an uncertainty of this cross section ratio closer to 5%. To support this, Fernandez-Sanchez (the lead of the group in which Avila did his work) has privately communicated with me that 5% is his assessment of the absolute accuracy of their water vapor cross sections and given that the nitrogen cross section uncertainty is in the range of 1-2%, this is consistent with a claim of ~5% uncertainty in the cross section ratio. Venable et al has some text concerning this. So I believe that an assignment of 5% to the uncertainty of the Raman water vapor/nitrogen cross section ratio is justifiable. But at least this more recent work makes the 10% Penny and Lapp uncertainty from 1976 no longer appropriate.

Thank you for noticing that this was missed in the original literature review. We will change the numbers accordingly and cite Avila 2004 and Venable 2011.

The new sentence is as follows:

"The limiting factor in the white lamp calibration technique is the degree to which we know the molecular cross-sections which are known to have uncertainties on the order of 5% (Avila et al. 2004, Venable et al. 2011)"

9. P4, L19. Immler et al, 2010 is used as a reference for the GRUAN RS92 correction technique. Immler et al discusses error characterization in general but does not present the RS92 correction technique. The Dirksen et al, 2014 reference is more appropriate.

We agree with you and will remove the Immler reference.

10. P9, Lines 8-9. Statement is made that "we do not correct for aerosols as they are considered to have a very small contribution to the overall mixing ratio". I take this to mean that the differential transmission due to aerosols is not accounted for. In the 1992 reference that is cited to support the authors' statement, it is shown that with aerosol optical thickness at 355nm of 1.0 the calculated mixing ratio would change by ~4% as compared to a pure Rayleigh atmosphere. Indeed, AOT of 1.0 is quite a turbid atmosphere but this result also implies that AOT of 0.25 would yield a 1% change in mixing ratio. One's first impression might be that 1% uncertainties are small enough to neglect (I do not agree). But neglecting aerosol differential transmission does not introduce a random uncertainty but rather a systematic one. And surely in a paper that has long term trend detection as a stated goal, elimination of systematic uncertainties that can be up to 4% must really be done. So I strongly encourage that the authors address this deficiency. Note that it is not necessary to calculate aerosol extinction directly from the lidar data to adequately make this correction. One can instead use collocated aerosol optical thickness measurements along with a reasonable estimate of the height of the boundary layer to

develop a simple model for calculating the aerosol differential transmission such that the residual uncertainty in the aerosol differential transmission correction is well below 1% even under turbid conditions. This is the technique that we generally use to handle this tricky part of the Raman water vapor lidar analysis.

We agree that not including aerosols does induce a systematic bias and should be taken care of. We had not done this originally because we do not currently have a lidar extinction product for RALMO, nor was there a co-located instrument capable of measuring AOD during nighttime at that time.We do, however, have an aerosol scattering ratio product and therefore we did not think it was appropriate to use a daytime AOD measurement due to the variability of aerosols over the course of a night. RALMO's total backscatter ratio product is calculated by taking the ratio of the elastic and the sum of the pure rotational raman signals at 355 nm. Therefore, we have used this product and assumed lidar ratios and an angstrom exponent using an angstrom exponent time series in order to estimate aerosol extinction profiles.
Similarly to the method followed in Sica and Haefele et al. 2016, we calculate the extinction profile using the following equation:

$$\alpha_{aer}(z) = LR(z) * (\beta_{mol}(z) * (BSR(z) - 1))$$

Where $\alpha_{aer}(z)$ is the extinction profile which changes with altitude *z*, *LR(z)* is a lidar ratio profile, $\beta_{mol}(z)$ is the molecular backscatter profile taken from the NCEP model, and the *BSR(z)* is the total backscatter ratio profile. The lidar ratio profile is a step function with a constant value in the boundary layer and another constant value for the free troposphere. The height of the boundary layer is estimated using the backscatter ratio profile. We have assumed lidar ratios of 20 for the free troposphere and 50 for the boundary layer using climatological values from the Payerne station. The transmissions for each channel are calculated using the equation below:

$$T_{aer,x}(z) = exp\{ -(\lambda_x/\lambda_0)^A * \int \alpha_{aer}(z)dz \}$$

Where $T_{aer,x}(z)$ is the aerosol transmission profile for a given molecule *x* (e.g. N2 or H20), λ_x is the wavelength for a particular channel, λ_0 is the reference extinction profile which in this case is 354.7 nm for the elastic channel, and *A* is the angstrom exponent which is assumed constant with altitude. The Ångstrom exponent, *A*, is assumed constant with altitude. The Ångstrom exponent is measured during the daytime using the co-located Precision Filter Radiometer (PFR). We have no measurements of nighttime angstrom exponents, therefore we are forced to use the daytime values, unlike the case of the optical depth calculation where we could use the lidar backscatter ratio. The Ångstrom exponent is not measured daily as it requires stable, cloud-free conditions to get an accurate calculation. Since it is not always available, we fit the sum of a 6 and 12 month sinusoid to the angstrom exponent time series over measurements from 01 January 2012 until 31 December 2015, with 2014 removed due to a faulty sensor. The fitted sinusoid was then used as the values for the angstrom exponents. The standard deviation of the residuals was ± 0.34 and was used as the uncertainty for the angstrom exponents

We have calculated the uncertainty induced by our assumptions by using the uncertainty equation introduced in the paper. We assume the worst case scenario of 100% uncertainty in the extinction profile calculation and the standard deviation of the Angstrom exponent residuals for the uncertainty calculations.

We found that the uncertainty in the calibration constant due to the uncertainty in the extinction profile was much less than 0.01% for all cases. The uncertainty in the calibration constant due to the uncertainty in the angstrom exponent was only an order of magnitude higher and on average of 0.4% +/- 0.5%. Therefore, the radiosonde remains the largest calibration source.

While the uncertainty in the calibration constant is low due to our assumptions, the calibration constants did change by an average of 2% when adding in the differential aerosol transmission to the calibration. On nights with a strong boundary layer (2 cases), we did see a change in 6 - 8% in the calibration constant. The results we get seem to be consistent with the results of Whiteman 2003.

11. P10, Lines 16-17. "we require … to be correlated to greater than 90%." I assume that by this the authors mean than the correlation coefficient of the linear regression is 0.9 or greater. If so, please restate in terms of correlation coefficient to avoid confusion.

You assume correctly, we will change the wording as you suggest so that it is clearer.
We have changed the sentence as follows:
" To ensure that the …. We require the resulting uncalibrated lidar and radiosonde mixing ratio profiles to have a correlation coefficient which minimizes the variance of the fit's residuals and must be above 0.75". We have changed the calculation of the slope to one similar to what you discuss in Whiteman 2012 and have therefore changed the sentence accordingly.
12. P10, last line. I do not see that the results of Aug 8, 2012 showing a 5% offset. Is this something that is apparent from the Table? If not, please clarify that this information cannot be gleaned from the Table.

Thank you for pointing out that this should not be referring to Table 1 and is in fact missing a reference to Figure 6. We have changed the wording in this section of the paper to no longer discuss 5% offset and instead discuss the differences in the features measured by the lidar and the radiosonde. We believe it is more appropriate to discuss the features individually instead of discussing overall biases. We have not included the text here since the entire section has been reorganized and we would refer you to the re-worked Section 4 in the revised paper.

13. P15, lines 5-10. A qualitative comparison of results of homogeneous and heterogeneous cases is made but the actual standard deviations, for example, are not given. From Table 1, it seems that for the homogeneous cases, the standard deviation of the calibration constants derived using the traditional technique is less than that of the trajectory technique. For the heterogeneous cases, the trajectory technique gives slightly smaller standard deviation as

stated. I do suggest giving the actual standard deviation values in the table and discussing the significance of these standard deviation differences since they seem to be rather small.

You are correct that the standard deviations are not given here, they were originally put in the summary which was not the appropriate place. The discussion of the standard deviations has been moved to where we introduce the Table and is much clearer. It is true, that when considering the entire population of both the standard deviations are small and therefore the differences between the two would not be statistically significant. However, when the two extraneous cases were removed from the calculation then the differences between homogeneous and heterogeneous nights become more apparent. Please refer to the new text in Section 4.

14. P 17. "Lidar Calibration Uncertainties for Trajectory and Traditional Methods". Three sources of uncertainty are listed: lidar statistical uncertainty, GRUAN radiosonde uncertainty, dead time uncertainty. The "usual" way that radiosondes have been used in a Raman lidar calibration effort (e.g the MOHAVE, AWEX, IHOP, PECAN field campaigns) is to assume that the lidar calibration value has been constant over the duration of a field campaign and that differences in calculated calibration constants relate to statistical uncertainties, collocation uncertainty, etc. Following this procedure, a single calibration constant is determined from all the radiosonde comparisons in a field campaign and that calibration value is used for some period of time until another large intercomparison effort with multiple radiosondes is performed. This is the technique outlined in discussions of the hybrid technique, for example, and in such cases there is another very significant component of the calibration uncertainty, which I call the calibration transfer uncertainty, that is not listed here by the authors since it does not pertain to what they are doing (but it is very significant in the overall discussion of lidar calibration). This systematic uncertainty can be taken to be the standard deviation of the individual calibration constants used to determine the mean calibration constant that is finally used in a field campaign type of study. I understand that the authors are doing things differently and are re-calibrating the lidar with every available radiosonde. In fact, the authors approach is much preferred from the standpoint of developing a time series for trend detection because each time a different calibration constant is used for the lidar, a step-change systematic uncertainty is introduced into the time series. This is inevitable. So to decrease the influence of these systematic step-changes, frequent calibrations are needed so as to make these systematic uncertainties, in effect, components of the random uncertainty budget in the time series. The authors refer to some of this later in the paper but here is where it should be introduced. Thus to recalibrate the lidar as frequently as possible serves to transform a component of the systematic uncertainty budget (where it can really destroy a trend calculation) into a random uncertainty. Note that the DOE/ARM Raman lidar is recalibrated with respect to microwave radiometer every three days achieving this randomization of calibration constant. However, campaign mode calibration efforts as described in the MOHAVE papers do not achieve this. So … I suggest that the authors clarify this. There is discussion in Whiteman et al, 2011b about the need to randomize components of the systematic uncertainty budget to improve time series for trend detection.

This comment is very helpful, and while we tried to discuss exactly this concern towards the end of the uncertainty section, we did not stress how our method differs from typical practice enough or its implications for trend analysis. We agree that it would be better to discuss the typical methods of calculating calibration uncertainty at the beginning of the section instead of the end and would lead to a natural transition to our preferred method. Therefore, we will remove the first few sentences in that last paragraph and add the following text to the beginning of the uncertainty calculation section :

> The standard practice for determining the uncertainty of the calibration constant has been to conduct extensive calibration campaigns and assume that the calibration value does not change over the campaign period and then measure the variability of the constant \citep{Ferrare1995,Turner2002,Whiteman2006,Leblanc2008,David2017}. The variability of the constant is then assumed to be the uncertainty and the calibration constant is not changed until the next campaign when multiple radiosondes are available for calibration. The assumption that the calibration constant does not change over long periods of time introduces another source of uncertainty into water vapour measurements, which is often unknown until the next calibration period. Uncertainties calculated in this way vary between 4 and 5\% of the calibration constant during the calibration period, but do not account for the individual sources of contribution nor do they typically account for the variability in the calibration constant beyond the campaign period.
>
> Accounting for drift or changes in the calibration constant is extremely important for long term trend analyses, since such a drift/change could easily be larger than the uncertainty of the calculated trend and would render the analysis invalid if it was not considered \citep{Whiteman2011b}. Many systems have now taken this into account by conducting daily or semi-daily calibration measurements either using an internal, hybrid, or external calibration. Taking more frequent calibration measurements with uncertainties calculated for each calibration then turns a systematic uncertainty component of a trend analysis into a random uncertainty component, particularly if the uncertainty of the calibration constant is recalculated with each calibration.
>
> While it is possible to calculate the uncertainty budget of a calibration constant based on the lidar's measurements and components, often the largest unknown uncertainty is the uncertainty of the reference instrument \citep{Leblanc2008}. It was not until recently that such detailed information was available for radiosondes. The GRUAN radiosondes are the first radiosondes to have a published uncertainty budget as a function of altitude for each measurement \citep{Dirksen2014}. By using the GRUAN radiosondes, we are now able to calculate the uncertainty in the calibration constant due to the radiosonde's uncertainties.

14a. It's also in this section where the uniqueness of the use of GRUAN sondes for this calibration task should be highlighted. This is the first time, to my knowledge, that linear

regressions of radiosonde/lidar data have been performed with weights that make use of carefully characterized radiosonde uncertainties. This is significant.

You are certainly correct that these points have not been stated with enough clarity in the paper and need to be more heavily emphasized. We believe the last paragraph in the response to the previous comment makes the distinction more clearly.

15. P18, line 21. The term "scan" is used here and earlier but it is not clear what "scan" means. Please go back in the paper and define how you use this term the first time it appears.

Thank you for pointing this out - scan seems to be a colloquial term within this research group. A scan refers to a 1 minute or 1800 shot raw measurement profile. We have added this definition in the first instance where "scan" appears.

16. P19, line 11 … I chuckled when I read that eq 5 is a simplified version of eq 4. Upon inspection eq 5 is about twice as long as eq 4 so does not appear much simplified. You might just say "With these assumptions, eq 4 becomes ..."

Perhaps "reduced form" might have been the better wording. We will change it as you suggest.

17. P20, line 3. This is where it becomes clear that you are recalibrating the lidar with each radiosonde. You also make the point that this is different than for field campaigns as in the Leblanc and Dionisi references. Good. Now, as mentioned earlier, you can make the point that this approach helps to randomize a component of the systematic uncertainty making the resulting time series more appropriate for trend detection.

Thank you, we did try to make this distinction albeit not very well. We have added new text to the beginning of the uncertainty section which now explains the difference between our approach and the standard field campaign approach.

18. P20, lines 13-16. I've already commented that ignoring aerosol differential transmission neglects a systematic bias which is a strong concern and goes against the prescription of the BIPM/GUM where all known systematic biases should be corrected (see quote in Whiteman et al, 2012 or go to the GUM itself). Also, though, the way that the sentence reads it is not clear what 5% refers to. Finally I would say that one should perform the calibration of the lidar data in the same way that it is analyzed for trend detection and one would not want to neglect aerosol differential transmission when trying to create trend-detection quality time series of water vapor measurements. So aerosols really do need to be accounted for in this analysis and in the full analysis of the lidar data.

We agree that aerosols should have been included. We have answered this concern after your previous comment and have tried, to the best of our ability, to include them and account for the uncertainty in our assumptions.

19. P23, lines 3-5. "frequent and accurate lidar calibrations are critical for detecting water vapor trends …" The earlier discussion of randomizing components of the systematic uncertainty budget is the main argument for why this statement is true so you should add a citation here. But I need to repeat that the measurement challenge in the LS is very different than in the UT so that your statement really only applies to the UT. BTW, these are the reasons why trend detection in the UT is so much easier with Raman lidar than in the LS:

1. The natural variability of water vapor in the UT is much higher than in the LS. So the relatively large random uncertainty of Raman water vapor lidar does not deteriorate the time to detect trend by a large fraction in the UT.

2. On the other hand, the natural variability of water vapor in the LS is very low and the random uncertainty of lidar measurements is much, much larger in the LS since it is farther away than the UT and water vapor concentrations are so small in the LS. So the random uncertainty of Raman lidar measurements in the LS typically will swamp the uncertainty budget and greatly extend the time to detect trend using the methodology of Weatherhead et al, 1998.

3. According to the modeling cited in Whiteman et al, 2011b the anticipated trends in LS water vapor are smaller than those in the UT making trends more difficult to quantify in the LS.

4. Because of much lower S/N lidar measurements in the LS, small sources of systematic bias in the lidar measurements can more easily corrupt the time series. The larger signals in the UT are more resistant to such unknown sources of bias.

We will add a citation here for Whiteman 2011 b.

We agree that calculating trends in the UT is undeniably *easier* than LS, but we would argue that for future lidars, or even for the latest improvements, that measurements in the LS and trends in the LS should be possible. As stated previously, we would prefer not to limit the discussion to only the UT since we hope that this paper will serve as a reference for future lidars which may be built specifically for the purpose of studying the LS and will have the ability to detect trends at those heights.

20. P23, last paragraph. At the end of the study a conclusion is that the trajectory method does not produce statistically different calibration values than the trajectory method. This does not argue strongly for the technique presented here. I would suggest looking for ways to decrease the standard deviation of the calculated calibration values. In Whiteman et al, 2012 we found that by using the adaptive technique described there we could reduce the variability of the calculated calibration values by requiring that the correlation coefficient between the lidar and radiosonde profile segments be higher. You might try adding that into your algorithm since, as I understand, you already require R2 > 0.9. The point here is that it should be a goal of this work to achieve a more stable calibration constant than that achieved with the traditional technique.

We agree that we haven't written a strong enough conclusion here and that the goal should be to achieve a stable calibration constant. The conclusion will be revised to make the following important points:

1. The trajectory method does improve the differences between the radiosonde and lidar, particularly on the heterogeneous nights.
2. This is the first paper to use the GRUAN sondes for a nightly calibration uncertainty analysis.
3. The height dependent uncertainties reported by GRUAN allow us to calculate the uncertainty of each calibration constant.
4. This method should allow one to do more frequent calibrations using radiosondes launched farther away from the observatory which in turn will help randomize the calibration uncertainty in any trend analysis.

We have done as you suggested and implemented the variable correlation method you used in Whiteman et al. 2012. With one difference where we do choose the calibration constant directly from the slope. This is because we propagate our uncertainties directly from the least squares fitting equation. We have changed the paper accordingly and updated it with the new calibration values. Implementing the variable correlation did not significantly change the final value of the calibration constant - at most there was a shift in 0.5% of the calibration value. However, adding in the aerosol component does seem to have decreased the variability of the constant.