Review of Hicks-Jalali et al, “Calibration of a Water Vapour Lidar using a Radiosonde Trajectory method” by David N. Whiteman

General Comments

The subject manuscript presents a new twist on the use of radiosonde profiles to determine the Raman water vapor lidar calibration constant. The method defines a vertical column surrounding the lidar and considers radiosonde measurements that fall within this vertical column for use in the calibration. Other criteria for selection of data are applied as well, but the requirement of physical co-location is the principal novelty mentioned by the authors. A concern is that the authors’ trajectory technique does not produce significantly different results from the “traditional” technique so at the end of the paper one is left wondering about the value of the approach.

But I think the authors are overlooking a significant novelty of this paper which, with a modest amount of new analysis, could make a significant contribution to the lidar calibration literature. Thanks to the very large efforts of GRUAN, radiosonde data with fully characterized uncertainties are now available at GRUAN launch sites such as Payerne. Thus, in the long history of water vapor lidar-radiosonde intercomparisons for the purposes of calibrating a water vapor Raman lidar (since Melfi et al (1969); so now extending for half a century!), to my knowledge, this is the first effort to use rigorously determined uncertainty values for radiosonde measurements in the process. This novelty permits the radiosonde data to be weighted properly in the calibration regressions. Thus, as mentioned, I suggest a shift of focus away from the radiosonde trajectory technique and toward quantifying how the use of GRUAN sondes for calibration differs from using ordinary sondes. The authors could quantify this effect by comparing the calibration results when using weights in the regression of radiosonde data (appropriate due to the availability of the GRUAN data product) versus when not using weights in the regression. I would expect that this would result in different calibration regions being chosen at least. But there could be other effects as well that are worth documenting. I don’t believe that this has been done before and would constitute a valuable addition to the literature. It might be appropriate to consider a title change to something like “Use of rigorously characterized radiosonde data for calibrating a Raman water vapor lidar”.

Also, as mentioned in the specific comments, I believe that it is necessary for the authors to account for the influence of aerosols on the differential transmission correction.

Specific Comments

Abstract

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...”
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.

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.
4. L29- “paralyzation” → “paralysis”

Introduction

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.
6. L6 - “of” → “from”
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 ...”
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.
    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.
9. P 4, 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.

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.
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.
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.
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.
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.

1. 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.

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.
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 ..."
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.
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.
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.

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 $R^2 > 0.9$. 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.