A General Response to the Editor and to the Reviewers:

Thank you very much for your detailed comments and the time you took to consider this paper. We have responded to the general comments below. Some of the general comments are repeated in the "specific comments" section, therefore, we have responded to both the general and specific comments but only provided the changed or relevant text in the specific comments section. We have also included a marked-up version of the manuscript showing what sections are removed and what has been added.

We believe that there may have been some confusion as to the goal of this paper and we would like to clarify this for the reviewers and the editor. The goal of this paper was to create a general method to calibrate raman water vapour lidars using a radiosonde's trajectory. This method may be applied to any lidar system which uses radiosondes, provided the radiosondes record wind measurements (speed and direction). Otherwise, there is no other limiting factor in the method's design.

In the conclusions, we state that this method will be used to calibrate the RALMO water vapour measurements for an upcoming trend analysis. We would like to stress that this is future work, and the trend analysis is not the subject of this paper. Therefore, we will not go into detail on the trend analysis here. We believe discussions regarding the calibrations for the trend analysis should be reserved for the following papers. We have other calibration techniques which will be used for trend analysis, however, this paper is not the place to include that discussion. We have answered the reviewer's comments regarding the trend analysis accordingly.

Thank you again for considering this paper and we hope you find the answers to your concerns satisfactory.

Response to Reviewer #1

Response to General Comments:

Comment: The structure of the paper needs to be reworked to improve clarity and precision. I suggest to include section 3.2 into current section 2.2. I suggest to improve internal structure of Sect. 4 (see major comments n°4), to remove the Summary section (you should share its content into sections 4, 5, the conclusion and avoid repetitions) and to split the current section 7 into a Discussion section (that you need to develop) and a Conclusions section.

We have reworked several of the sections as per your suggestion, particularly sections 4 and the discussions and conclusions. Please see the specific comments for the specific changes.

Comment:

The « traditional » method is not described simply at the beginning of the article, there is just a list of bibliographical references that the reader must read without knowing which one is used precisely. Explanation loops are given as the article progresses but they arrive too late. A short or detailed description of the so-called « traditional » method should be added to the introduction and/or methodological part.

We agree that the definition of the traditional method was not clear. We had hoped that Figure 1 would be sufficient to explain the traditional method, but we do agree that some text is necessary.

We will address this problem in two ways:

1) A new section will be added to discuss the traditional method before the trajectory section. We will include a short introduction to the traditional method in the introductory paragraph to this section but refrain from a full literature review as we believe it has been well discussed in several papers and we do assume the reader is somewhat familiar with the technique if they are interested in reading the paper.
2) The previous sections 3.2 and 3.3 have now been added to the traditional method procedure. The procedure discussed in these two sections is used for both the trajectory and the traditional, therefore, we think it is appropriate to be introduced here. This way the readers have a clear view on how the traditional method is calculated in our study.

Comment: In the description of the lidar measurement and the description of the « improved » methodology, a lot of choice in filtering the data are made based on high or low SNR but it is never quantified. Please be more specific on this point.

We agree that the SNR values should be quantified and have clarified or removed the few cases which were ambiguous. Thank you for pointing this out. We have answered all of the specific comments regarding the SNR values individually where they occur, as this was more efficient.

In general our cutoff SNR value was 2 as lower SNRs have too much of a noise contribution.

Comment:
Sect. 4 presents many issues:
4.1 The structure needs to be reworked so that the reader can have the following elements:
◦ Presentation of data
◦ Methodology to differentiate between nights when water vapour is homogeneous and nights when water vapour is heterogeneous
◦ Presentation of Table 1
◦ General comment (the current last paragraph)
◦ Illustration of the different characteristics with Fig. 5 and 7
◦ Conclusion with Figure 8

We agree that this section does not flow appropriately and have reorganized it as follows:

- Introduction of the calibration constant comparison, Table 1, and definitions of homogeneous and heterogeneous nights.
- Discussion of Homogeneous nights
    - Why are the calibration constants different? Referencing Figure 6 (now Figure 5).
    - Discussion of Figure 5 (now Figure 6) and the example homogeneous nights
    - Analysis of the homogeneous calibration constants
- Discussion of Heterogeneous nights
    - Discussion of Figure 7 and the example heterogeneous nights
    - Analysis of the calibration constants
- Conclusions with Figure 8

The new section includes changes over several pages of text; please see the included mark-up version of the revised manuscript.

4.2 Figures:
◦ they are under used, even not used for specific subplots. Maybe there is too many Figures, the legends need to be shortened (some analyses of the figures are made in the legend whereas it should be done in the text),

We agree that the discussion of the Figures should be expanded and will do so for each of the Figures not discussed in enough detail. Comments regarding specific figures are discussed below. We would prefer not to shorten the captions of the figures because it is often the case where figures are "borrowed" from papers for presentations or discussions and can be often taken without the proper context if it is not clear in the caption. Therefore, we would like to keep some analysis and context in the caption to account for these situations.

4.3 The quantification of biases of « 0 % » in average. I suggest taking absolute values and indicating the sign of the bias.

Thank you for pointing this out. We have changed the way in which we discuss the comparison for the two methods, and believe that this removes the problem of biases around 0. Please refer to the new Section 5 for the changed text.

It is repeated over and over in the summary and discussion/conclusions that the method is « more accurate », in other words but the results on uncertainty does not quantify this improvement.

Perhaps it would be better to be more specific about the use of the term "more accurate" and we should specify in exactly which scenarios we believe this is the case. Here, we use "accuracy" to mean "closer to the true state". If we assume the radiosonde is the true measurement, then

by "more accurate" we are saying that we produce a profile which is closer to the radiosonde measurements (or the percent difference between them is closer to 0).
In the case of the homogeneous nights, the trajectory method does just as well as the traditional in terms of matching the radiosonde profile and the uncertainties. However, in the heterogeneous nights, the method does improve the differences between the lidar and radiosonde and does more accurately reproduce the radiosonde measurement by up to 20% (Figure 7). In this case, we do not use the term "accurate" to refer to the uncertainty of the method, but with respect to the shape of the radiosonde profile.
With respect to the propagated uncertainties, it is true that the traditional and trajectory methods do produce similar uncertainties, therefore, it would not be correct for us to use the word "more" when referring to the uncertainties of both methods. We will re-examine the wording in these sections to make sure that the discussion is fair to the traditional technique.

The discussion/conclusion about the advantages of the method, it must be thorough:
5.1 Almost no discussion about the limitations of the method: 32 % of night are calibrated (what about the others?) on 6 years (the first 3 years not being calibratable also)
5.2 The benefits presented are based on the theoretical expectations that motivated the implementation of this methodology. The uncertainty is presumably better but this is not reflected in the budget calculation.

We agree that there should be a clear discussion of the limitations of the method. However, we would suggest that removing nights due to bad weather is not a limitation that is inherent of the method, but is due to the signal attenuation of the lidar measurements on cloudy nights which, depending on the height of the cloud, could remove a large amount of the altitudes available for calibration. It is not preferable to calibrated during cloudy conditions because we do not measure the lidar ratios inside clouds, additionally clouds are not uniform and signal level would not be consistent over all scans.

We agree that there should be a clearer discussion of the traditional uncertainties and will include the appropriate references inside the uncertainties discussion, such as Wandinger 2005 and uncertainty discussions from Leblanc et al. 2012. We have answered the specific comments regarding the uncertainties individually in the "Specific Comments" section.

5.3 Discussion should be pushed before opening the perspective of using this dataset to « UTLS climatology over 10 years »: how to do that with 24 nights above 2008-2016, how to calibrate night where this methodology could not be applied? It should have thought because the authors wants to use the whole dataset for trends study.

We agree that the discussion should be expanded to include the limitations of the method. However, we would prefer to refrain from discussing the details and the methodology of the trend analysis in this paper (particularly in the conclusions/future work section) since the subject of this paper is not the trend analysis, and is the calibration technique itself.

RALMO has bi-weekly RS92 radiosondes launches with which this method can be used. We did not include them in this study because these radiosondes do not report uncertainties as a function of altitude and thus, should be considered secondary, not primary calibration opportunities, like the coincident GRUAN sondes. The non- GRAUN sondes will be used along with the GRUAN sondes and other calibration techniques when forming a climatology of the RALMO measurements.

Specific comments:

Comment: *Title*
I strongly suggest to refer to RALMO in the title or to find a way to indicate that it is a test of this new methodology of calibration done on one lidar which could potentially be applied to others.

We disagree with this suggestion to add RALMO to the title. There is nothing in this manuscript specific to RALMO. The limiting factor in this method is in fact only due to the radiosonde and whether or not the final data reports wind speed and direction. We will add a sentence in this regard to the Abstract and the Conclusion.

Abstract sentence: This trajectory technique is a general technique which may be used for any lidar, and only requires that the radiosonde report wind speed and direction.

Comment:
Abstract – page 1
Lines 1-4: I would keep theses sentences for an introduction because it is too general. At least,please shorten this part.
The authors disagree. It is necessary for some introduction to understand the context of the study, even for experts in the field.

Lines 5 & 6: Reference citations should not be included in this section unless they are essential. Using radiosondes is the most used technique for calibrating so please select a maximum of one reference, I would suggest Whiteman et al. (2006) which is the closest from the one you will use?. Maybe, the method is even better summarized in Whiteman et al. (2012). -

While we agree that references should be kept to a minimum, all three papers used similar techniques and we feel that it is appropriate to acknowledge their involvement and to make it clear that the work is built off of three different papers. It would not be fair to the other authors to not include them as well.

Line 7 « movement of radiosonde »: I suggest replacing it with « movement of air masses »
We have changed the sentence to: "However, they did not consider the movement of the radiosonde relative to the air mass and frontal boundaries"

Line 12: Precise on which period the calibration has been performed (i.e. 2011-2016). **-**
We have added "from 2011 - 2016" to the sentence on Line 12.

Lines 14-15: The authors use « more accurately » but there is no conclusion in the article that quantifies that the uncertainty associated with the new technique is better than the « traditional » one . I suggest replacing « reproduces more accurately » with « reproduces accurately ».
It is true that the method does more accurately reproduce the radiosonde profile and decreases the differences in the profiles by up to 20%. However, to avoid confusion over the word "accurately", we have changed the sentence to say:

" We show that the trajectory method reduces differences between the radiosonde and lidar by up to 20% when the water vapour field is not homogeneous over a 30 min calibration period."

Lines 16-21: The summary associated with the uncertainty budget is too detailed. Please replace this part by one value (or range of value) quantifying the total uncertainty associated with the calibration.
We have removed the sentence regarding the deadtime uncertainty to shorten the section and only include the most important result, which is that the uncertainty in the calibration due to the uncertainty in the radiosonde is 4% of the calibration value.

Page 2

Line 2: Please replace « the primary contributor» with « one of the main contributors ».
We have changed this as requested.

Line 3: Add reference to « ...high temporal and spatial variability »
The authors believe that this a common knowledge statement and is generally stated throughout the water vapour community. However, we can add references to a few papers (Trenberth et al. 2005, Ross and Elliot 1996, Kampfer 2013).

Line 4: I suggest deleting « uniquely ».
We have replaced "uniquely" with "well".

Line 6: I suggest deleting « more » or please add a reference.
We have removed the word "more".

Line 8: Replace « take » with « make ».
We have changed this as requested.

Line 9: Delete « also ».
We have changed this as requested.

Line 10 « Several Raman... external methods »: I would place this sentence in the following

Paragraph.
We would prefer to keep that sentence where it is because it is necessary to explain why a calibration constant is necessary for water vapour lidars and its importance in a trend analysis.

Lines 17 to 22: It's too detailed whereas it is not the main subject of the article. I suggest deleting this part.
We have removed two sentences from this section and condensed them into one sentence.

Line 29: Please add the use of GNSS as an external instrument to calibrate Raman lidars and a reference. I suggest: David, L., Bock, O., Thom, C., Bosser, P., and Pelon, J.: Study and mitigation of calibration factor instabilities in a water vapor Raman lidar, Atmos. Meas. Tech., 10, 2745-2758, https://doi.org/10.5194/amt-10-2745-2017, 2017.
We have added this paper and mentioned GPS satellites to the first sentence on the paragraph detailing standard external measurements.

Page 3 Comments:

Lines 1-2 « External...do not contribute »: I suggest deleting this sentence. -
This sentence is necessary for the reader to understand why some lidar stations prefer to calibrate externally rather than internally.

Line 7 « as RS92 radiosondes are the most frequently used calibration radiosondes Immler et al., 2010 ; Dirksen et al., 2014). »: I'm not convinced that the main objective of this article was to correct sondes for the calibration of Raman lidars. Please rephrase.

While the purpose of this article was not to correct sondes, this sentence is necessary to explain why we choose to use GRUAN sondes for calibration instead of the uncorrected RS92 sondes. Additionally, it shows that we are aware of the disadvantage of using radiosondes, but have tried to reduce that problem by using the most accurate product. Therefore, we believe this sentence should not be removed.

Line 8: Replace « Vaisala » with « not corrected »
We have changed this to "uncorrected RS92"

Line 9: Replace « errors » with « uncertainties »
We have changed this as requested.

Lines 9-10 « A portion of... uncertainty »: I suggest moving this sentence in the next paragraph, line 17 before « This paper attempts to... ».
We have moved this sentence as requested.

Lines 12-15: I suggest moving the whole paragraph line 27.

The paragraph starting on Line 28 directly follows from the previous paragraph and moving the paragraph on line 12 would break the flow of the section. Therefore, we would prefer to keep it where it is.

Line 19: Please explain in few sentences (2-3) what is the « traditional » method, references are not sufficient considering that the improvement of this technique is the main subject of this paper.
We agree that adding a few sentences discussing the traditional method would be helpful for the reader. As mentioned in the general comments, we believe that we have answered this sufficiently in two ways:

1) A new section for the Traditional Method has been added.
2) Text introducing the traditional method has been added and is shown below:
    a) New Text: “ The ``traditional'' method for calibration water vapour lidars is done by integrating a fixed number of lidar profiles as a function of height starting at a time which is coincident with the radiosonde launch and then calculating a linear weighted least-squares fit between the radiosonde and lidar measurements to determine the calibration constant \citep{Melfi1972, Whiteman1992}. The altitudes over which the fit is conducted are either fixed (e.g. always 1 - 5\,km), or the optimal altitude region may be determined by calculating the correlation between the radiosonde and the lidar measurements. For the purposes of this paper, we refer to the traditional method as using 30\,min of integration with a weighted least-squares fit over altitudes determined by maximum correlation. “
3) The previous sections 3.2 and 3.3 have also been added to this section in regards to another comment so that the entire traditional method is clearly discussed and the reader can then better understand our implementation of the method.

Line 20 «as the radiosonde takes approximately 30 min to reach the tropopause »: It depends on which latitude the sonde is launched, it will be larger near the equator. Please specify it or add a location.

We have added the clarifier “at mid-latitudes”.

Line 21: Your statement should be supported by a reference or some statistics from your database.

We agree that we should state where these numbers came from. We also thank you for drawing our attention to a wording mistake - it should not say “up to” as this implies a maximum. 4 km was in fact the minimum distance traveled. This statement is based on statistics from our study and will be rephrased as follows:

… during which time radiosondes in this study drifted a minimum of 4 km away from the lidar’s field-of-view ….

Line 28 « in order to ensure that the lidar and the radiosonde are measuring the same air »: It was also the goal of the « traditional » method. Please check.

Yes, this was also the goal of the traditional method. We will rephrase the sentence as follows:

Original Sentence:
"In order to ensure that the lidar and radiosonde are measuring the same air, we have developed an improved lidar-radiosonde calibration technique that utilizes the position.."

New Sentence:

As in any atmospheric calibration method, it is important that the instruments involve measure the same air mass. To improve the coincidence for periods where calibration is required but the atmospheric water vapour content is changing, we have developed an improved lidar-radiosonde calibration …"

Line 28: I would suggest deleting « improved ».

We believe that changing the sentence as above fixes this issue.

Line 29: Replace « of the radiosonde and the » with « , ».

This would not be correct since "wind" applies to both the speed and direction and therefore requires an "and" beforehand instead of a comma. We can change the sentence to:
"... we have developed an improved lidar-radiosonde calibration technique that utilizes the position of the radiosonde and the radiosonde wind measurements."

Page 4 Comments:

Line 7: Add « in Payerne » at the end of the sentence.
We have changed this as requested.

Line 9 « their respective uncertainties »: The full uncertainty budget for the calibration is not given for the « traditional » method. What about the « representation uncertainty » for example? I will come back to this in more detail in following comments.
We agree that we did not discuss this in enough detail for the traditional method. Indeed, we did calculate the uncertainties in the same way for the traditional method as we did for the trajectory, however, we did not make it clear enough for the reader. We will clarify this in the uncertainty section and in the summary flow chart. We have also added the uncertainties to Table 1 so that the reader can see that they are similar.

Line 13: Does the lidar always start working at 0:00 UTC? Does « bi-weekly » refer to twice a week or one every two weeks? Please specify it in the text.
Since this section is for the radiosonde, it would not be appropriate to put lidar details here. RALMO is an operational lidar and operates 24/7 so long as it is not raining or undergoing repairs/routine maintenance. We will add the sentence to Page 5, Line 3- "As an operational lidar, RALMO runs 24/7 except when the cloud cover is below 800 m". Bi-Weekly refers to every other week, which we have added to the text.

Line 14: Why is their only « a subset » of these radiosonde processed by GRUAN? Please explain it in the text.
Unfortunately, not every RS92 radiosonde launched at the Payerne station is GRUAN compliant and many are launched for only internal studies, which is why only a subset may be processed by GRUAN. We will change the sentence to "Only a subset of these radiosondes are processed by GRUAN because not every RS92 flight before 2019 was GRUAN-compliant."

Line 23: Please specify that the analysis was conducted on an initial set of 76 flights but in the end only 24 of them were used.

We have edited the last sentence of section 2.1 to say: "A total of 76 GRUAN RS92 nighttime flights were initially used to conduct this analysis, however, due to clouds and lack of coincident lidar data, only 24 flights were used in the end for the calibration."

Page 5 Comments:

Section 2.2: Precise that the RALMO is operating day and night. Give the effective measurement time (« 50% » in the conclusion, it should be precised earlier) and what explain that 50% of the time is not exploitable.

We have added this sentence to the end of the first paragraph in section 2.2 to make this more clear: " RALMO runs day and night with an average of 50% uptime from 2008 - 2017. RALMO downtime is due to the presence of clouds below 800 m, or repairs/routine maintenance."

Line 2: The authors explained that the instrument « is designed to be an operational lidar, and as such, needs to have high accuracy, temporal measurement stability, and minimal altitude-based corrections (Dinoev et al., 2013; brocard et al., 2013) ». The study of the instrument's performance made in the bibliographical references and years of operation should determine whether the instrument really has a high accuracy, temporal stability of measurements and profiles that start close to the ground. Please be more specific.

You are correct, this manuscript is part of a broader study of the instrument one of us (SHJ) is performing as part of her PhD thesis work. We have amended this sentence to: "RALMO was designed to be an operational lidar and therefore was designed to have high accuracy, temporal measurement stability, and minimal altitude-based corrections."

Lines 4-6 « RALMO operates...scattering channels »: Move this sentence to the previous paragraph.

We have moved this sentence as requested.

Line 9 « a sufficiently high SNR »: This is a major issue in this paper. The SNR is often used to select some data and reject others, but no threshold based on bibliographic references or empirical tests specific to this study is defined. You need to clarify this.

We agree that this is not well defined throughout the paper and we will fix this by providing specific SNR values, or remove the discussion of SNR where it is not appropriate. In this case, the second half of that sentence is not appropriate for this section and we will remove it. Only a background value threshold of .01 photon counts/bin/s is applied at this step in the process.

Lines 12 & 13 « After the filtering process...small features »: I wonder if this sentence should not be found in the description of the calibration methodology. Or at least move the interpolation part of the radiosonde profiles to section 2.1 and specify here that there is no interpolation for lidar profiles.

We have moved the sentence to the radiosonde section and have now specified that there is no interpolation of the lidar profiles at the end of section 2.2.

Line 15 « 3.1 Tracking air parcels »: I would move this title at the end of the page. Paragraph line 16 to 20 is more an introduction to sect. 3.

We have moved the first paragraph before section 3.1 as we agree that it is more of an introduction to Section 3.

Page 6 – Figure 1
I would use « homogeneous region » instead of « homogeneous cylinder » to be more consistent with the text. I would suggest adding letters or numbers to the different steps and refer to them in the text.

We will change this to agree with the rest of the text in the paper and use "homogeneous lidar region" instead.

Page 7 Comments

Line 7: As simple that it could be, you should provide mathematical explanation/an equation to illustrate your calculation.

This conversion is standard and can be easily found online as the local, flat earth approximation which is appropriate for distances smaller than 20 km. We have used the reference website used by NOAA for our calculations: http://www.edwilliams.org/avform.htm#flat. The theory behind the transformation is discussed in
Smart, W. M.: Textbook on Spherical Astronomy 6th Ed., edited by R. M. Green, Cambridge University Press, Cambridge., 1977.
The authors have also verified this through derivation, which if necessary can be added here. We would be happy to include the citation in the text to guide the reader to the appropriate material, however, we don't think it would be appropriate to add the conversions to the paper.

We will add to the text that the conversion can be assumed using the "local, flat Earth conversion for a spherical Earth" which should provide enough context for the reader to find the appropriate conversions.

Line 8: This assumption should be discussed a little more. You might provide a physical discussion about when (or if) this assumption is (would be) realist and its limitation, please cite references that could support this hypothesis or this discussion.

As this method is not used at all in this paper because all of the 2008 - 2011 radiosondes were removed due to the presence of clouds, we will remove this sentence from the paper. However, if you are curious about why this assumption is reasonable please see pg 47 of :
Daidzic, N. E. (2017). Long and short-range air navigation on spherical Earth. International Journal of Aviation, Aeronautics, and Aerospace, 4(1). https://doi.org/10.15394/ijaaa.2017.1160
Appendix D, demonstrates that assuming plane trigonometry is acceptable for short distances on Earth (20 km or less). The radiosondes in this study traveled between 4 and 18 km away from the lidar during a 30 minute calibration, therefore, the assumption is valid here. Additionally, since we are calculating the distances between each radiosonde measurement, this is certainly valid at distances on the order of meters.

Line 9: What do you mean by « We do not explicitly consider the vertical movement of the air parcel in this method »? From your description of the methodology, I understand that you do not consider at all the vertical movement. Please be more specific.
We agree that the word "explicitly" is confusing here. We have removed it because we do not consider the vertical movement.

Line 14: For my point of view, the « lidar region » refers more to the ~1 m diameter at 5 km that the radius of 3 km choosen after for the sensitivity test. You might use a name more related with the assumed homogeneity of the water vapor like or its use for calibration, « homogeneous lidar region » or « calibration region » as you call them after in the paper.

Thank you for catching this. That is in fact a typo and should be "homogeneous region". We have redefined this as the "homogeneous lidar region" as suggested.

Lines 14-15: What is the decisional parameter? Is this the SNR? Which threshold?
The decisional parameter in this case was the SNR of the water vapour channel. Radii smaller than 3 km resulted in SNRs smaller than 2 below 7 km and in some cases halved the SNR of the water vapour channel at altitudes below 5 km. Therefore, a radius of 3km was chosen to maximize the SNR in the lower altitudes without creating too much smoothing of features and to maintain SNRs of at least 2 above 7 km in the majority of cases.

We will reword the paragraph as follows for clarity:

"In order to maintain a SNR in the water vapour greater than 2 above 7 km altitude for the majority of the cases, we defined the homogeneous lidar region to be a circle of 3 km radius centered around the lidar. The size of the homogeneous region was chosen by varying the radius from a range of 1 - 25 km and finally increasing it to infinity. Radii below 3 km resulted in SNRs smaller than 2 below 7 km and in some cases halved the SNR of the water vapour channel at altitudes below 5 km, which decreased the altitude coverage for the calibration and increased the noise in the primary calibration region. While radii above 3 km resulted in SNRs larger than 2 above 7 km, the water vapour profiles started to exhibit biases due to using too long integration times at certain altitudes and losing small features which had previously been visible. The 3 km radius provided the most altitude coverage with profiles closest to the radiosonde measurements and was the best compromise."

Lines 16/17/18: You characterized some « very low SNR », a « large enough SNRs » and « the highest SNR ». Please quantify or explain why you define it this way.

The above paragraph is intended to fix both this comment and the previous comment.

Page 8 Comments

Figure 2 : Is this a conceptual scheme or a real example? Please precise it in the legend. If it is an example, give the date of the measurement. The second sentence of the legend is also explained in the text. Please replace this sentence with something like « The purple circle corresponds to the lidar region».
This is a conceptual example, which we will specify in the caption. We would prefer to keep that sentence there for readers who choose not to read the text and only go through the figures.

Lines 6-7 « The standard thirty minutes...tropopause »: It sounds quite general in your text but it corresponds to mid-latitudes. It would be a different duration for polar regions or in the tropics. Please be more specific by adding « at mid-latitudes » for example.
We have added the specifier " at mid-latitudes".

Lines 7-8 « Integrating...by radiosonde »: I suggest deleting this sentence.
We have removed this sentence as requested.

Lines 9-10: The analysis of Figure 3 consists in one sentence. Either your analysis is too short, either the figure is not necessary. Please reconcile.
We would be happy to increase the discussion of Figure 3, as we believe that it helps the reader understand how the integration time changes with altitude. We would add the following text.

"The decrease in integration time with altitude will also change depending on the rate at which the radiosonde moves away from the lidar. The majority of nights had integration times less than 5 min above 7 km. However, if the wind is strong at a particular altitude, sharp decreases in integration times may be seen as the radiosonde moves quickly away. It is also possible to see the integration times decrease and then increase again as the radiosonde drifts in and out of the homogeneous lidar region."

Page 9 Comments

Figure 3 Legend: « The integration time...analysis. »: Please remove these sentence from the legend. It is already explained in the text or is part of the analysis.

We would prefer to leave these sentences in the caption to aid readers who do not read the text. It is common for readers to use figures from a paper in presentations or discussions, and we would prefer to keep the analysis in the figure caption so that they have it readily available.

Figure 2 – Figure 3 – Figure 4

In this section, Figure 2 seems to be a conceptual method, Figure 3 refers to July 21 and Figure 4 to July 22. It is not explained why these 3 dates are choosen for each example. Considering that Section 3 details the method of calibration by sonde, you might choose the same date to illustrate the different aspects of it. If not, please justify why.

July 21st is a typo, thank you for catching that. Only one date has been used for these figures - July 22nd.

Section 3.2 - Pages 9 and 10
I suggest moving this section in or after Section 2.2 because it is more about the « Lidar measurement » than about the « Radiosonde Trajectory Method ».

We agree that technically this is more related to the lidar measurements section. However, to address your request for a better description of the traditional method, we have created a whole new section for the traditional method and put it there. This way the traditional method is clear and we only discuss the differences introduced by the trajectory method.

Page 9 Comments

Lines 10-11: I would rephrase this way: « The central wavelengths of the water vapour and nitrogen channels of the RALMO were choosen to minimize temperature dependence. »
We have changed this as requested.

Line 14: Please add a reference.
We have added the citation of Whiteman 1992 as requested.

Page 10 Comments

Lines 4-5 « In RALMO's case, the ratio...2014). »: Please move this sentence to Sect. 2.2. - We have moved this entire section to the new traditional method calculation Section 3.1.

Line 8 « the corrected signal »: Which correction? Please be more specific.

By "corrected" we mean "transmission-corrected and cloud-filtered lidar signals". However, I think it is not necessary to mention corrected here. To be clearer, the sentence has been changed to : " After calculating the un-calibrated mixing ratio profile from the ratio of the two lidar signals (Section 2.2), we use a ….".

Line 9: Does « the correlated and weighted least squares fitting » correspond to the « traditional » method? If so, why not call it this way in the abstract so that you don't list so many references? It could be included in the introduction in this way as well.

Yes and no … the correlated and weighted least squares fitting does not refer to the traditional method - although our implementation of the traditional method does use these techniques to fairly compare with the trajectory technique. The traditional method refers to the selection of scans used to calibrate the lidar.
We believe we have fixed this problem by including this section in the new Traditional method section.

Line 14 « low SNRs »: Please specify.
Water vapour scans integrated less than 5 minutes typically have SNRs less than 2.
We will rephrase the sentence as follows:
".... due to SNRs less than 2 …. "

Line 20: I suggest replacing « lidar region » with « calibration region ».

Thank you for your suggestion. We agree that the term needs to be consistent and appropriately descriptive for the reader. However, the term calibration region is not the best because it could refer to the altitudes at which we calibrate, which is not the case. Therefore, we have decided to use the term "Homogeneous lidar region" throughout the paper and will change all accordingly.

Section 3.3 – Pages 10 and 11
It seems to be the traditional method that is mainly described here except that the lidar data are selected as described in Section 3.1. Be careful to distinguish in the text of this part (and even in the entire article) between what is specific to your study and what is traditional. Perhaps the structure you have chosen is a little confusing on this point.

We have added a few sentences in Section 2.2 discussing the traditional method to make sure the differences are clearer to the reader. Thank you for your suggestion regarding this issue.

Page 11 Comments
Lines 2-10: I suggest moving this part to Sect. 2.1 or at least be more concise.

We have moved this section to the new traditional method section. Unfortunately, we do not see how it can be made more concise as it is necessary to explain the reason for the average pressure profile and why we believe the assumption is justified.

Line 12 « for July 22, 2015 »: Why did you choose this date instead of July 21 as in Figure 3? That would seem more consistent.

Thank you for catching this error. This is indeed the same date. The time formats were changed from local time to UT time in the paper at the last minute and we missed some of the changes.

Line 15: Please replace « error » with « uncertainty ». - In this case, standard error is a specific term. What to do here?

Some sources will use the term standard error, but since we are following Bevington's derivation, we will replace the term "standard error" and simply use the "uncertainty of the slope" to avoid confusion here and to keep terms consistent throughout.

Line 15: Please add a reference. What was the range of uncertainty found in the litterature for the « traditional » method?

Thank you for pointing this out. We have added three more paragraphs to the uncertainty section discussing the typical way that the uncertainties are calculated for the traditional method and have cited several papers as examples. We had discussed this very briefly towards the end of the section, but the other referee thought it would be better to make this distinction in the beginning as it was not clear to the reader. To answer your question, the typical traditional calibration uncertainty value is between 4 and 5%, but is calculated differently (the details of which are discussed in the new version of the paper).

Figure 4 – Page 11
The date is not specified, please do it. The last sentence is not necessary.

We will specify the date - it is 22 July 2015. The sentence is necessary for readers who may choose to not read the text and it is necessary to explain the uncertainty value shown in Figure 4 b.

Page 12 Comments

Line 3: What do you mean by « repairs »?
In this case "repairs" really means routine maintenance such as alignment, changing the flashlamps, etc. We have clarified this in the paper.

Line 4 « abnormaly high »: Please quantify or explain.
We have clarified this sentence to say "abnormally high background values above 0.01 counts/bin/min".

Line 6: I suggest: « and presence of clouds ».
We have changed this as requested.

Lines 6-7 « The filtering process...radiosonde launch »: This sentence seems to say that there were clouds every night of measurement for all the duration of the measurement. Is that what you mean? Yes, that is exactly what we mean.

Does this mean that there are 3 years of measurement that cannot be used because of systematic presence of clouds?
No, there are other radiosondes available with which we can calibrate during those 3 years. However, there are no RS92 radiosondes during that period that were launched on clear days.

Line 8: If I understood correctly, between 2008 and 2016, there are only 76 calibratable nights and they are all condensed over the period 2011 and 2016. What solution for 2008-2010? On the other hand, over the 76 nights, with the implementation of this methodology, only ~30% can be calibrated? What about the other nights of measurements? It is essential to discuss these aspects in Section 7.

We think there has been some confusion over the goal of this paper. This paper is meant to present a calibration technique, which we have implemented using GRUAN RS92 radiosondes. We cannot calibrate a water vapour lidar on cloudy days, therefore, the weather has removed the majority of the radiosonde flights from *this* study. This is not a problem with the method itself, but an unfortunate happenstance caused by nature. It is possible to calibrate with non-GRUAN RS92 radiosondes and there are also operational radiosondes which are launched from Payerne daily which would increase the number of calibration points, but as we stated earlier, these are secondary sources and not primary sources for calibration and therefore were not included.

Page 13 Comments

Table 1: Column « Difference »: Sign of the difference is missing or it should be specified that it is absolute value.
Thank you, we will specify that it is the absolute value.

Legend: I suggest deleting the word « calibration » in the second sentence. The text from « Two nights in the homogeneous » to « variability in the water vapour » should not appear in the legend.
We can remove the word “calibration” in the second sentence. However, we would like to keep those sentences in the Table captions because they are necessary to explain the comments in column 6.

Line 6 « show a good agreement »: Please quantify or explain.
This sentence is a hypothesis of what we should see, therefore we did not quantify “good agreement”. However, we did neglect to follow up with the results for the homogeneous section which were included in the Summary instead. We have moved the last paragraph in the section into this paragraph to help quantify the agreement between the traditional and trajectory methods on homogeneous nights. Hopefully reorganizing the paper in this way will make it more understandable to the reader.

Line 7: I suggest adding « water vapor conditions around the location of the lidar measurement ». We have changed this as suggested.

Line 8 « as « homogeneous » or « stable » nights in Table 1 »: Only « homogeneous » is used in Table 1 in comparison with « heteregenous » , « stable » is used in the « Comments column » for heterogeneous nights. Please reconcile. - We agree that this is not consistent. We will remove the word “stable” since this is confusing and only use “homogeneous”.

Line 9 « an average bias of 0 % »: Please used absolute value to calculte the average bias. Maybe you should use the standard deviation and specify if the bias is more positive or negative.

Perhaps the best way to discuss the differences between the trajectory method and the traditional method is not in terms of biases, but in terms how well it removes the large differences between the radiosonde and lidar. For example: In Figure 7, on 2012-07-27, the difference between the radiosonde and the lidar produces sharp differences on the order of +/- 15% - 20%. However, these differences are reduced by 10 - 15% in the trajectory method. By discussing the features individually, and discussing the changes in the profiles, we believe this will clarify the problem of the bias calculation in addition to increasing the discussion of each of the figures. Please refer to the revised manuscript for the added text as there is too much to include here.

Page 14 Comments

Line 1 « The bias on that night is reduced when using the trajectory method »: What do you mean? Please rephrase.

We believe our answer to the previous comment has also answered this comment.

Lines 2-3: Figure 5 is composed of 9 subplots and show results for 3 dates but there is only one sentence that refers to it in the text. You need to analyse your figures or do not put them in the paper.

We agree that we should discuss these and all figures in more detail. We have included 2 paragraphs of discussion for Figure 5 in the new manuscript and would refer you there.

Figure 5 – Page 14 Legend - 4th sentence: Why do you mention « White vertical regions » whereas there is none of them in the figure? Please delete this sentence.

Our apologies, that sentence should indeed be removed from the caption. Thank you for catching the error.

Legend - 7th sentence: You should use « magenta » instead of « pink » or change the color of the corresponding line in Figure 5.

We have changed this as requested.

Legend - 8 and 9th sentences: It is part of the analysis and should not appear in the legend. Please delete them.

As stated previously, we would prefer to keep these sentences inside the caption for readers who choose to skim the paper. We will make sure these sentences are also included in the text.

Figure 6 – Page 15
There is no 2012-07-27 measurement in Table 1. Is this the right date? Please check.

Thank you for noticing this discrepancy. It is the right date. The date formats in the paper were changed at the last minute and the dates in the table were not changed to UTC time. The dates in the table have now been updated to the correct format.

Legend – 1st sentence: I suggest deleting « launch at 0 min ».

We have changed this as requested.

Page 15 Comments

Lines 1-5: This part explains the method to characterize a night as homogeneous or heterogeneous, it should come earlier in this section; before discussing about the results. Please reconcile.

We agree that this should come earlier in the section and have changed this as requested. This section has been reorganized as per your general comment at the beginning.

Line 3: Figure 7 is quoted before Figure 6 (p16 - l.11) in the text. The order of the figures is not Respected.

Thank you for noticing this. We have rearranged the figures in this section such that Figure 6 now comes before both Figure 5 and Figure 7 as we feel that this is a clearer way of discussing the comparisons.

Line 3: As for other figures in the article, Figure 7 is commented in only one sentence.

We will increase the discussion of this figure in the paper, as you suggested.

Line 8: See my previous comment on the average bias.
Please see our previous answer. We will increase the discussion of Figure 7 to include how the trajectory method reduces the maximum differences between the radiosonde and the traditional method lidar profile.

Line 9: Please refer to a figure or a table.
We will refer to Figure 7.

Line 14: Why « 11 » whereas there are 12 heterogeneous nights in Table 1? - Thank you for seeing catching this error. That was a typo and should have said 12. Note that we have now amended one of the previous heterogeneous nights and moved it to the homogeneous section. Therefore, there are now 11 heterogeneous nights.

Page 16 Comments:

Figure 7 – Why did you choose these three nights? You must analyze all three of them, when you add a figure it should bring an supplementary information, otherwise it means that the figure is not necessary.
We agree that these nights could be analyzed in more detail. We chose these nights because they were prime examples of the heterogeneous nights. We have added another paragraph discussing this figure in more detail to the new manuscript.

Legend – 2nd to 6th sentences: This is already explain in legend of Figure 5, please just leave the part on the white vertical regions and refer to legend of Figure 5.

We can refer to Figure 5's caption. Thank you for the suggestion.

Legend – 7 and 8th sentences: Again, this does not have to appear in the legend.

As per our previous comments, we would prefer to keep these sentences inside the caption.

Page 16 Comments

Lines 1-2: Shouldn't these nights be described as homogeneous in this case? Does this not question the methodology of homogeneous versus heterogeneous characterization?

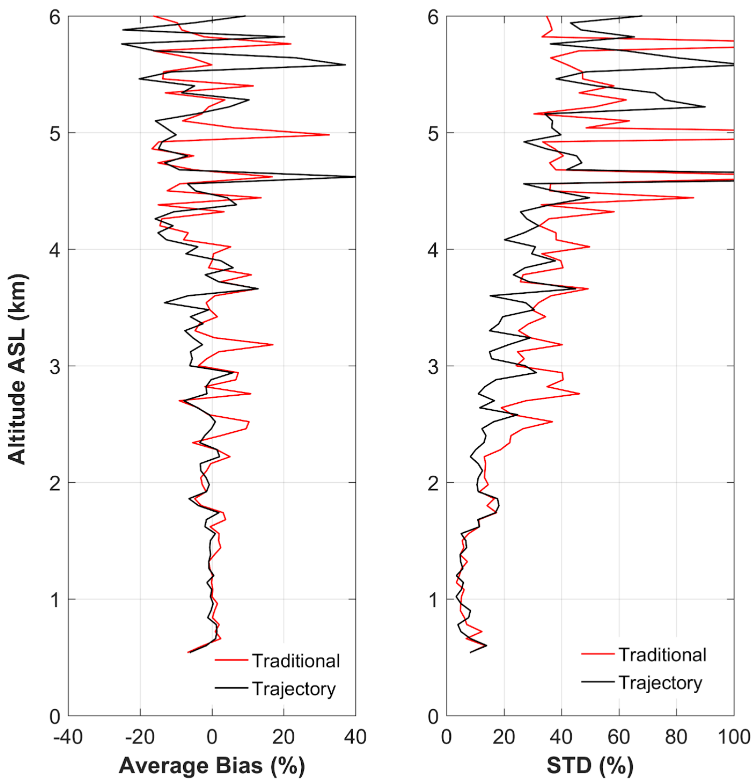
These nights were defined as heterogeneous nights because they had features that changed by over 50% at a given altitude over the 30 minutes of lidar measurements. Therefore they were defined as heterogeneous nights. However, the correlation algorithm determined that the regions which had over 90% correlation with the radiosonde where the regions which were homogeneous over 30 minutes. This is not always the case, therefore we would consider the heterogeneous vs. homogeneous distinction valid in general. What it does suggest, is that the regions which we choose for calibration are important.

Line 4: The average bias is around 1 % here whereas it was 0 % earlier (p15 - l. 8). Please reconcile.

In the previous case, the average bias was referring to the plots in Figure 7, not the average for the entire data set, which is the case here. We will clarify the language in both sentences so that it is clear which figures we are referring to.

Line 5: Please specify what threshold you chose to define that the variability increases above 5 km.

Thank you for catching this, it should in fact say 4.5 km and is referring to the increase in the standard deviation of the percent differences in Figure 8. However, it is not clear on the scale of this figure, therefore, we will increase the x axis to 100% STD for clarity.



Line 6: Here you precise the sign of the bias whereas you have not done it earlier, same for the significant number of digits. Please reconcile.

In this case we are discussing the average of the entire data set, where previously we were discussing the bias of a specific date. This is not clear in the paper and we will change the language such that the reader can easily determine which is which.

Line 7: Please rephrase the sentence.

Original:

> The standard deviation of all of the percent difference profiles shows that the trajectory method more accurately fits the radiosonde profile above 2 km on a profile-by-profile basis and will more consistently provide better fits.

Revised:

> The standard deviation of the ensemble of percent difference profiles between both calibration methods and the radiosonde shows that the trajectory method more accurately matches the radiosonde profile above 2 km.

Line 8 « better fits »: Please quantify.

The standard deviation of the trajectory method is on average 10% smaller than the traditional method for altitudes above 2 km and below 4.5 km. This implies that the profiles more accurately fit the radiosonde on a profile-by-profile basis. We believe that the new sentence above fixes this problem.

Line 11: Once again the commentary on the figure does not exceed one sentence. Please develop or remove the sentence.

We apologise that this was not clear, but the next few sentences also referred to Figure 6. We have rectified this in our restructuring of this section.

Page 17 Comments:

Figure 8 -I suggest choosing one reference (either the trajectory or the « traditional » method) to avoid the multiplication of the data and to put the average bias and the standard deviation on the same plot.

Thank you for your suggestion. We would prefer to keep both data sets on the plots so that the reader can easily see how they differ from each other.

Starting Page 16 -line 10 to Page 17 – line 7: This is a description of the methodology and should come earlier in the section. Please put this explanation after the introduction of Table 1.
We agree that this discussion should come sooner and have moved it as suggested.

Lines 3-4: What do you mean by « the majority »? How many nights? In Table 1, it seems that it is true for 6 nights on 12. This represents half of the data not the majority.
You are correct, the percentage should say 1% to encompass the majority (10 out of 12).

Lines 4-5: Please specify the dates
We have specified the dates as requested.

Page 18 Comments:

Lines 2-7: You repeat yourself. This was already mentioned earlier in the article.

We agree that this is a repetition, however, we feel that it is necessary to remind the reader of how these are calculated as they may not remember and it is not useful to refer back to a previous section.

Lines 17-18: Please precise how the fitting uncertainty is calculated.

This is a standard calculation which is easily referenceable. Therefore, we will reference the Bevington textbook so that the reader has a source for the equation. However, we don’t feel that it is necessary to include the equation in the paper.

Page 19 Comments:

Line 2: Which part is the systematic part versus the statistical part of the uncertainty? This is important to determine given that one of your objectives is to establish trends in the UT/LS.

The radiosonde uncertainties are the total uncertainty resulting from the combined Statistical and Systematic uncertainties for Relative Humidity, Pressure, and Temperature and are calculated by the GRUAN uncertainty algorithms (Dirksen et al. 2014). We will clarify this sentence and change it to say: “The uncertainty of the water vapour calibration constant due to the lidar’s random uncertainty and the radiosonde’s total uncertainty (both systematic and random) was determined ….”

Line 5: Please explain what do you mean by the « measurement vector » with regard to the lidar Measurement.

The measurement vector is a vector which includes all of the lidar and radiosonde measurements, e.g. - MV = [L1, L2, L3, …. R1, R2, R3…] where Li is a lidar measurement and Ri is a radiosonde measurement.

Line 12: Please correct the indices of the third term of Eq. (5)
We have changed the indices in the third term to R_j instead of R_i+1. Thank you for catching this discrepancy.

Lines 13-25: The comments of the equation is not clear enough. First you should describe each term,precise the source of each uncertainty (radiosonde measurements, lidar (including lidar’s photon counting and deadtime and covariance term), quantify the uncertainties before (« 8 % » coming to late in the next section, ?, 5 %, calculation?) and after propagation (values of Table 2) and you should use Table 2 to support this section.

1. We have described each term for this equation directly after the equation, and before the equation since it uses many of the same variables as Equations 3 and 4.
2. The sources of the lidar and radiosonde uncertainties were previously discussed in the manuscript in the first paragraph of Section 5 (now Section 6). , However, we have not discussed the quantities of each and can add them to the same paragraph.

Actually, Table 2 comes too late in the paper.

- We would prefer to leave Table 2 as a summary of the uncertainties, however, we will make sure to quantify the uncertainties used in the text more clearly.

You should also precise how is considered your uncertainty on the calibration constant due to the radiosonde: statistical or systematic?

- The uncertainties for the radiosonde are the combined statistical and systematic uncertainties as calculated in Dirksen et al. 2014 and is what is reported in the GRUAN corrected radiosonde files. We had simply said "total uncertainties" in the paper but we will specify that they are a combination of all uncertainties.

Then you can conclude on the total uncertainty of the calibration constant for which you are found after propagation: an average value of 4%.

Line 16: Why do you mentioned that the fitting uncertainty is the same that lidar's photon counting uncertainty. As you do it l.24 to compare it to the deadtime uncertainty? Is there a physical or metrological explanation? The fitting uncertainty seems to be due to the methodology and is different from these two uncertainties.

The fitting uncertainty is the uncertainty in the fit calculated from the variance of the residuals. The variance is largely dominated by the photon counting uncertainty from the lidar measurements, therefore, the fitting uncertainty and the uncertainty in the calibration constant due to the lidar statistical uncertainty should be the same or very close. Indeed, upon inspection, we found that the two agreed within a tenth of a percent in all cases (page 19, Line 18).

We have changed the wording in the text, starting on Page 18, Line 18 to hopefully clarify the our definition of the fitting uncertainty:

Using the variance of the residuals of the least-squares fit, one can calculate the uncertainty in the fit, or ``fitting uncertainty." This fitting uncertainty is the result of the amount of photon counting noise in the lidar measurements, and can be treated as the uncertainty in the calibration due to the lidar photon counting statistics. The average trajectory method fitting uncertainty is 0.4% of the average calibration constant. The average fitting uncertainty for the traditional method is 0.3% of the average calibration constant. The traditional method has smaller fitting or statistical uncertainties than the trajectory method due to the larger number of scans used per altitude, on average, compared to the trajectory method. The fitting uncertainty does not encompass the entire uncertainty of the calibration constant, since it is largely due only to the photon counting noise. It is also necessary to examine the uncertainty of the calibration constant due to the uncertainties in the radiosonde measurement as well as the deadtime.

Line 23: I suggest not making a line break here and including it in the previous paragraph. '
- We have changed this as suggested.

Line 23 « we assume a deadtime uncertainty () of 5% or .2 ns »: How is estimated this uncertainty: literature? Test?
The 5% came from testing these nights using the OEM retrieval by Sica and Haefele 2016. We found that the standard deviation of the deadtime retrieval value was 5%. The choice was also a result of discussions with Licel who suggested that 5% would be a reasonable value to assume

on a nightly basis. Sica and Haefele 2016 tested various *a priori* uncertainties for dead time and found that they all gave similar retrieved dead times.

Line 25: Please remove « including dead time effects ».
We have removed this as requested.

Page 20 Comments:

Line 7 « The calibration time series »: Do you mean the 24 nights? Why don't you show them on a figure?
Yes, we mean the 24 nights used in this study - we have added this clarification to the end of the sentence. We originally had the time series figure to the paper but we felt it did not add enough to the paper and that there were already too many figures. We also thought it would detract from the discussion of the calibration.

Line 8: What do you mean by « de-trended »? Please explain.
It means that a trend line was fit to the time series and then removed from the time series in order to calculate the standard deviation of the calibration factors. We have added this clarification to the paper and removed the word "de-trended".

Line 8 « over ten years »: You work on the 2011-2016 (i.e. 6 years), why do you speak about 10 Years?
You are correct, therefore we have changed this to say 6 years instead of 10.

Line 17 – Page 20 to Line 14 – Page 21: This « Summary » part should not be a part in itself. These information should be in sections 4 and 5 and in the conclusion. It is necessary to rethink the structure at the end of the article, which for the moment consists in two parts: 6 Summary and 7 Discussion and Conclusions (see comments below).
We respectfully disagree with this opinion. We believe that a summary is a necessary component of a paper. We recognize that some of this information should also be included in sections 4 and 5, and that the summary section should reiterate those points. We would prefer not to remove the summary section as we believe that it helps readers who quickly read through the paper. By keeping the summary separate from the conclusions, readers can choose to skip this section if they so choose and continue on.

Line 23 « due to the lidar profiles »: Please rephrase.
We are not sure how you would like this to be rephrased. We have changed the sentence to say " … due to the difference in 1-minute lidar scans selected".

Page 21 Comments:

Line 4: Please change « error » into « uncertainty ».
We have changed this as requested.

Line 4 « is negligible »: Do you neglect it in your calculation? This uncertainty does not appear in Table 2. Please specify it in the text.

We have stated that the fitting uncertainty is analogous to the photon counting uncertainty and we have included it in Table 2 as the photon counting uncertainty. We did not think it would be necessary to include both the fitting and the photon counting since they are equivalent to within a tenth of a percent.

The magnitude of the fitting uncertainty/photon counting uncertainty is a tenth of the total uncertainty which is dominated largely by the uncertainty of the radiosonde and has a magnitude of 4%. When considering the components of the uncertainty “budget” of the calibration, the fitting uncertainty is negligible when compared to the uncertainty of the radiosonde.

We have changed this sentence to:

“... although the fitting uncertainty is negligible relative to the uncertainty of the calibration constant due to the uncertainty of the radiosonde measurements (4% on average).

Line 5: This part is supposed to be a summary but this value « 8 % » appears for the first time. This should have been discussed earlier.

We agree that this should be discussed sooner. We will move it to Section 4.

All Sect. 6: This section contains many repetitions and brings many repetitions in relation to the whole article. There are key elements (such as the quantification of the uncertainty on the radiosonde of 8 % or Table 2) that should appear earlier in the article.

We agree that some of these sentences could be moved to the previous sections, however, as it is a summary section it is necessary to repeat the conclusions made in the previous sections. We would prefer to leave it in this format since it makes the paper more readable.

Page 22 Comments:

Line 1: This article lacks a real discussion part, so it is necessary to restructure it as follows:
6 Discussion
7 Conclusions
Line 2 « has several advantages »: What about the limitations? The following points should be discussed: no consideration of the vertical movement of air masses, only 6 years calibratable over 9 and only 32% of the 76 nights with exploitable radiosonde profiles according to the method's prerequisites. A real issue is: which method did you use for all other nights of measurements than the 24 selected? Because otherwise it means that over 9 years of data there are only 24 nights of calibrated and therefore usable data? We can't make trends over 24 nights.

Thank you for pointing out that it is definitely important to discuss the limitations of the method. We agree that a limitation of the method is it does not consider vertical movement. The fact that

only 24 dates out of the 76 original dates were useable was not due to the method itself but the poor weather that was coincident with the sonde launch.
Your concern for the trend analysis is correct, however, I do not believe it is a concern for this manuscript, which is describing a new calibration method which can be used independently of whether the lidar measurements will be used for trend studies. We do in fact have other radiosondes with which to calibrate in Payerne, but chose not to include them in this paper since we wanted to focus on using the GRUAN products which include detailed uncertainties for the radiosonde measurements. Calculating trends is important, and we are currently undertaking that study using the RALMO measurements.

Line 19 « an automatic [...] scheme »: What do you mean? Does this mean that this method has been implemented in a production chain for water vapor profiles of the RALMO lidar?
Yes, the method has been implemented for water vapour calibration for the RALMO lidar and can be done automatically.

Page 23 Comments:

Line 1: If we read the reasons for the exclusion of measurement nights for the calculation of the calibration coefficient on page 12, there does not seem to be more or less nights used with the new method rather than with the « traditional » method. Please reconcile.
We had meant this sentence to imply that if a sonde shows significant disagreement with a lidar measurement on the order of 1 g/kg then it may be the practice of some researchers to not consider that flight for calibration. In this scenario, one might be able to use this calibration technique to compensate for the difference between the two instruments and thereby gain more calibration nights.

Line 5: The uncertainty has not been quantified for the « traditional » method (at least not indicated in the article) or in this article. This is something that is difficult to quantify, but as a result it is difficult to conclude that it has improved it. Indeed from a theoretical point of view it is, but in practice (and in this discussion) you do not relate it to your results. This uncertainty appears in your introduction and in the conclusion part but no word in the section focusing on the uncertainties.

Thank you for pointing it out that this is not said explicitly in the discussion or clearly in the article. To fix this, we have added the uncertainties of both the Traditional and Trajectory calibration values in Table 1. We have also added a discussion at the beginning of the uncertainties section discussing the uncertainties of the traditional method and how others have calculated it in the past. We did discuss how Leblanc and others calculated the uncertainty by calculating the RMS of the calibration time series, but did not link it to the traditional method.

Line 10 « The method could... »: Yes but it should be presented as a limitation and not only as a perspective because it means that the « representation uncertainty » may not be as negligible as that.

We will discuss how it is a limitation to not include the vertical movement, as was mentioned in your comment for Page 22, Line 2. It is not obvious to us in general, outside of periods of intense convection, that typical vertical air motions over 30 min or less time scale would impact this analysis, especially for the trajectory method which is, by design, making sure the sonde and lidar sample the same region of air.

Line 12: I suggest starting here Sect. 7 Conclusions.
We would prefer to keep the discussion and conclusion together and not separate them.

Line 15: How do you assess that? Reference? Maybe Whiteman et al. (2011b)?
This sentence refers to Figure 8 in this manuscript which we neglected to include. We have now added the reference.

Line 15 « 50 % uptime over ten years »: Why is this only appearing now? 9 years of measurements -> 76 nights that can be calibrated -> 24 nights calibrated in practice, what about the other nights? Where is that 50%? Which calibration methodology for the entire database?
We did not include the operational components of the RALMO lidar because these have already been introduced in Dinoev et al. and Brocard et al.. These are being introduced now to discuss the future work that will be conducted using this study.

Line 24: To study trends in the UT/LS, please recall the total uncertainty associated with your profiles or refer to the article that assess the performances of the RALMO to measured water vapor in the UT/LS on a routine basis.
To date there are no studies assessing RALMO's measurements in the UTLS. RALMO was originally designed as a meteorological lidar with a focus on operational tropospheric measurements. The upcoming study in progress will discuss RALMO's performance in the UTLS, develop a climatology, and look for trends in that climatology. This work is beyond the scope of the current manuscript which is developing one of the calibration methods which will be used to develop a climatology.

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.

Calibration of a Water Vapour Raman Lidar using GRUAN-certified Radiosondes and a New Trajectory Method

Shannon Hicks-Jalali[1], Robert J. Sica[1,2], Alexander Haefele[2,1], and Giovanni Martucci[2]

[1]Department of Physics and Astronomy, The University of Western Ontario, London, Canada
[2]Federal Office of Meteorology and Climatology MeteoSwiss, Payerne, Switzerland

Correspondence to: Shannon Hicks-Jalali, hicks.shannonk@gmail.com

Abstract. NEW TEXT REMOVED TEXT

Carefully calibrated and quality-controlled Raman lidars are well-suited for trend measurements in the upper troposphere and lower stratosphere, particularly for species such as water vapour. Trend determinations require frequent, accurate, and well-characterized measurements. However, water vapour Raman lidars produce a relative measurement and require calibration in order to transform the measurement into mixing ratio, a conserved quantity when no sources or sinks for water vapour are present. Typically, the calibration is done using a reference instrument such as a radiosonde. We present an improved trajectory technique to calibrate water vapour Raman lidars based on the previous work of Whiteman et al. (2006), Leblanc and Mcdermid (2008), and Adam et al. (2010), and Herold et al. (2011), who used radiosondes as an external calibration source, and matched the lidar measurements to the corresponding radiosonde measurement. However, they did not consider the movement of the radiosonde. However, they did not consider the movement of the radiosonde relative to the air mass and frontal boundaries. Our trajectory method is a general technique which may be used for any lidar, and only requires that the radiosonde report wind speed and direction. As calibrations can be affected by a lack of co-location with the reference instrument, we have attempted to improve their technique by tracking the air parcels measured by the radiosonde relative to the field-of-view of the lidar. This study uses GCOS Reference Upper Air Network (GRUAN) Vaisala RS92 radiosonde measurements and lidar measurements taken by the MeteoSwiss RAman Lidar for Meteorological Observation (RALMO), located in Payerne, Switzerland from 2011-2016 to demonstrate this improved calibration technique. We compare this technique to the traditional radiosonde-lidar calibration technique which does not involve tracking the radiosonde and use the same integration time for all altitudes. Both traditional and our trajectory methods produce similar profiles when the water vapour field is homogeneous over the 30 min calibration period. We show that the trajectory method more accurately reproduces the radiosonde profile reduces differences between the radiosonde and lidar by up to 20% when the water vapour field is not homogeneous over a 30 min calibration period. We also calculate a calibration uncertainty budget that can be performed on a nightly basis. The calibration uncertainty budget includes the uncertainties due to phototube paralysis, aerosol extinctions, the assumption of the Ångstrom exponent, and the radiosonde. The study showed that the radiosonde was the major source of uncertainty in the calibration at 4% of the calibration value. We include the contribution of the radiosonde measurement uncertainties to the total calibration uncertainty, and show that on average the uncertainty contribution from the radiosonde is 4%. We also calculate the uncertainty in the calibration due to the uncertainty in the lidar's counting system, caused by phototube paralysis, and found it to be an

average of 0.3% for our system. This trajectory method allows a more accurate calibration of a lidar, even when non-co-located radiosondes are the only available calibration source, and also allows additional nights to be used for calibration that would otherwise be discarded due to variability in the water vapour profile.

1 Introduction

Water vapour is the primary one of the main contributors to the greenhouse effect due to its ability to absorb infrared radiation efficiently. Water vapour has high temporal and spatial variability, making it difficult to characterize its influence on the atmosphere (Ross and Elliott, 1996; Trenberth et al., 2005; Kämpfer, 2013). Instruments with high spatial-temporal resolutions, such as lidars, are uniquely well suited to long-term stratospheric and tropospheric water vapour studies. When conducting climatological studies, ground-based lidars have an advantage over satellite-borne instruments in that they have the ability to provide more frequent measurements of from the same location. Lidar measurements are particularly useful for creating statistically significant water vapour trends of the Upper Tropospheric and Lower Stratospheric (UTLS) region, as they are able to take make long term and frequent measurements (Weatherhead et al., 1998; Whiteman et al., 2011b) (Whiteman et al., 2011b). Minimizing the uncertainty in the measurements is also critical in order to establish a valid trend. A large component of a lidar measurement's uncertainty budget is its calibration constant. Water vapour lidars measure relative profiles, and therefore require a calibration to convert the measurements into physical units (here mixing ratio). Refining the calibration process is critical to detect the small changes anticipated in the trend analysis. Several Raman lidar calibration techniques have been developed over the years, including internal, external, and a hybrid of internal and external methods.

Internal calibration techniques require no external reference instrument. They can account for the entire optical path in the lidar system to find the water vapour calibration constant. In essence, all optical transmittance, quantum efficiencies of the detectors, Raman cross-sections, the geometric overlap, and their associated uncertainties must be quantified and accounted for. Some of these can be derived simultaneously using the white light calibration discussed in Venable et al. (2011) Leblanc and Mcdermid (2008). The white light technique is advantageous in that it can accurately track changes in the calibration constant. However, the calibration is incapable of detecting shifts in spectral separation units, and is not able to accurately detect the cause of calibration changes unless multiple lamps in different locations are used (Whiteman et al., 2011a). Venable et al. (2011) improved the technique by further expanded on the white light technique by using a scanning lamp instead of a stationary lamp. When the scanning method was compared to the external radiosonde method, it was found that both methods agreed with each other within their respective uncertainties. However, the internal white lamp method is dependent on the degree to which we know the filter bands, the lamp intensity function, and the molecular cross-sections. The limiting factor in the white lamp calibration technique is the degree to which we know the molecular cross-sections, which have uncertainties on the order of 5% (Avila et al., 2004; Venable et al., 2011).The last factor is the most limiting due to its current uncertainties on the order of 10% (Penney and Lapp, 1976). While internal calibration offers many advantages, it is impractical for many systems, such as lidars that use multiple mirrors (Dinoev et al., 2013; Godin-Beekmann et al., 2003) or large-aperture mirrors such as the rotating liquid mercury mirror of The University of Western Ontario's Purple Crow Lidar (Sica et al., 1995).

The standard external method involves comparing the lidar and a reference instrument; typically the reference instrument is a radiosonde (Melfi, 1972; Whiteman et al., 1992; Ferrare et al., 1995) but microwave radiometers or GPS satellites may also be used (Han et al., 1994; Hogg et al., 1983; Foth et al., 2015; David et al., 2017). External calibrations are often preferable because there is no need to characterize every system component and the uncertainties in the Raman cross sections do not contribute. However, the accuracy of the external calibration is dependent on the accuracy of the reference instrument. Radiosondes are widely used calibration instruments, as they have high spatial resolution, are routinely available, and widely available. Measurement uncertainties for relative humidity varies between 5 to 15% depending on the time of day (Miloshevich et al., 2009; Dirksen et al., 2014). To minimize the calibration uncertainties induced by biases in the radiosonde reference, the GCOS (Global Climate Observing System) Reference Upper-Air Network (GRUAN) has established a robust correction algorithm for the Vaisala RS92 radiosondes, as RS92 radiosondes are the most frequently used calibration radiosondes (Immler et al., 2010)(Dirksen et al., 2014). GRUAN RS92 relative humidity profiles have been shown to be 5% more moist than Vaisala uncorrected RS92 relative humidity profiles, while reducing the relative humidity uncertainties by up to 2% (Dirksen et al., 2014). A portion of the calibration uncertainty when using radiosondes can occur from the radiosonde's lack of co-location with the lidar, hereafter the “representation" uncertainty.

Hybrid internal-external methods, which also attempt to minimize variations in the sampled air mass due to the balloon's horizontal motion, have also been implemented by Leblanc and Mcdermid (2008) and Whiteman et al. (2011a). In these hybrid techniques, the white light calibration lamp is used to monitor the efficiency of the lidar optical paths, but is supplemented with radiosondes for the absolute calibration value. The hybrid technique will monitor relative changes in the calibration constant, but must be supplemented periodically with an external calibration (Leblanc and Mcdermid, 2008).

For any external calibration where the lidar and the calibration instrument do not share a common field-of-view, variations in water vapour cause an additional uncertainty in the calibration that is often not quantified in the uncertainty budget. A portion of the calibration uncertainty when using radiosondes can occur from the radiosonde's lack of co-location with the lidar, hereafter the “representation" uncertainty. This paper attempts to resolve the co-location problem and minimize the representation uncertainty by using a tracking technique that expands upon those discussed in Whiteman et al. (2006), Leblanc et al. (2012), and Adam et al. (2010). The co-location problem can be particularly acute for calibration via a radiosonde, as the radiosonde takes approximately 30 min to reach the tropopause at mid-latitudes, during which time the radiosondes in this study travelled a minimum of 4 km or morefrom the lidar’s field-of-view (assumed here to be the zenith, which is typically how water vapour lidars are operated). The distance traveled by the radiosonde has little effect on a calibration measurement if the air mass being sampled is horizontally homogeneous. However, this is not necessarily the case, and when we calibrate while on the edge of an airmass, or the air mass simply is not horizontally uniform, then the water vapour field may change dramatically over the distances the radiosonde travels. Lidar stations which have the resources to use daily radiosondes may not see this as much of a hindrance; however, if the station relies on infrequent calibration campaigns then the campaign calibration results are dependent on the weather at the time air masses which are sampled.

As in any atmospheric calibration method, it is important that the instruments involved measure the same air mass. To improve the coincidence for periods where calibration is required but the atmospheric water vapour content is changing,In order to ensure that the lidar and the radiosonde are measuring the same air, we have developed an improved lidar-radiosonde calibration technique that utilizes the position of the radiosonde and the wind speed and direction measured by the radiosonde. The wind speed and direction measurements allow us to track the air parcels as measured by the radiosonde with respect to the position relative to the lidar. If the air is within a 3 km radius around the lidar, we use the corresponding times and lidar scans for calibration. A lidar "scan" refers to a 1 minute (1800 shots) raw measurement profile. We have implemented the technique using 76 nighttime GRUAN RS92 radiosonde flights from 2011 to 2016. The GRUAN sondes represent the best characterized sonde measurements available in terms of calibration and uncertainty budget. Daytime calibrations were not tested due to the significantly reduced signal-to-noise (SNR) in daylight measurements and the inability to reach above 5 km effectively with the lidar. We will illustrate the method using measurements from the MeteoSwiss RAman Lidar for Meteorological Observing (RALMO) in Payerne, Switzerland (Dinoev et al., 2013; Brocard et al., 2013) on July 22nd, 2017 corresponding to the 00:00 UTC GRUAN RS92 radiosonde launch.

Section 2 will outline the measurements used in the study. Sections 3 and 4 discusses the methodologies of the traditional and trajectory methods. Section 4 will discuss the methodology. Sections 5 and 6 will compare the new trajectory method with the traditional calibration technique and their respective uncertainties. Sections 7 and 8 will summarize the results and discuss their implications and the next steps forward.

2 Calibration Measurements

2.1 Radiosonde Measurements

The MeteoSwiss Payerne research station launches Vaisala GRUAN RS92 radiosondes within 100 m of RALMO bi-weekly (every other week). A subset of these radiosondes are processed by GRUAN because not every RS92 flight before 2019 was GRUAN-compliant. GRUAN requires that radiosondes undergo several pre-flight checks and calibrations, which are detailed in Dirksen et al. (2014). These calibrations are needed to correct radiation and systematic relative humidity biases in the radiosonde temperature, pressure, and relative humidity profiles. This study uses the official GRUAN RS92 radiosonde product to minimize and accurately calculate the calibration uncertainty and the contribution from the radiosonde. All radiosonde measurements from 2011 to 2016 taken by the RS92 Vaisala sondes were processed by the GRUAN correction software (Dirksen et al., 2014). Radiosondes prior to October 2011 were RS92 radiosondes, but were not processed by GRUAN because they were not compatible with the GRUAN requirements listed in Dirksen et al. (2014). All radiosonde measurements were interpolated onto the lidar resolution grid (3.75 m).

The radiosonde water vapour mixing ratios are calculated using the GRUAN-corrected relative humidity profiles and the Hyland and Wexler 1983 formulae for the saturation vapour pressure (Hyland and Wexler, 1983). By convention, the relative humidity measurements are assumed to be over water for all altitudes. A total of 76 GRUAN RS92 nighttime flights were initially used to conduct this analysis , however, due to clouds and lack of coincident lidar measurements, only 24 flights were used for calibration.

2.2 Lidar Measurements

Lidar measurements in this study were made using RALMO. RALMO was built at the École Polytechnique Fédérale de Lausanne (EPFL) for operational meteorology, model validation, and climatological studies and is operated at the MeteoSwiss Station in Payerne, Switzerland (46.81° N, 6.94° E, 491 m a.s.l.). RALMO was designed to be an operational lidar , and as such, needs to have and therefore was designed to have high accuracy, temporal measurement stability, and minimal altitude-based corrections (Dinoev et al., 2013; Brocard et al., 2013). RALMO operates at 355 nm with a nominal pulse energy of 300 mJ and a repetition frequency of 30 Hz. Measurements are recorded for one minute (1800 laser shots) with a 3.75 m height resolution from both the nitrogen (407 nm) and water vapour (387 nm) Raman scattering channels. RALMO runs day and night with an average of 50% uptime from 2008 - 2017. RALMO downtime is due to the presence of fog,clouds below 800 m, or precipitation (40%) as well as repairs/routine maintenance (10%).

RALMO operates at 355 nm with a nominal pulse energy of 300 mJ and a repetition frequency of 30 Hz. Measurements are recorded for one minute (1800 laser shots) with a 3.75 m height resolution from both the nitrogen (407 nm) and water vapour (387 nm) Raman scattering channels. The lidar measurements are processed for calibration in several steps. First, we select ±2 hours of 1-minute lidar profiles around the launch time of the radiosonde. While two hours was chosen as an arbitrary time range to allow for scan selection, in practice the method rarely selects scans more than 30 min before or after the launch. The 1 min scans are filtered to remove scans with abnormally high backgrounds above 0.01 photon counts. and to ensure that we have a sufficiently high SNR in each scan. We assume clouds are present if the nitrogen SNR is less than 1 at 13 km. If a cloud is present, the scan is masked and removed from the calibration. After the filtering process, the radiosonde measurements are linearly interpolated onto the lidar altitude grid of 3.75 m. The calibration is conducted at the lidar's native altitude resolution in order to provide as many data points as possible and to avoid smoothing out small features.

3 The Traditional Method

The "traditional" method for calibration water vapour lidars is done by integrating a fixed number of lidar profiles as a function of height starting at a time which is coincident with the radiosonde launch and then calculating a linear weighted least-squares fit between the radiosonde and lidar measurements to determine the calibration constant (Melfi, 1972; Whiteman et al., 1992). The altitudes over which the fit is conducted are either fixed (e.g. always 1 - 5 km), or the optimal altitude region may be determined by calculating the correlation between the radiosonde and the lidar measurements (Dionisi et al., 2010; Whiteman et al., 2012). For the purposes of this paper, we refer to the traditional method as using 30 min of integration with a weighted least-squares fit over altitudes determined by the correlation coefficient which minimizes the variance of the fit's residuals.

3.1 Calculation of the Water Vapour Mixing Ratio for RALMO Measurements

The water vapour mixing ratio (w) for the RALMO is calculated from the background- and saturation-corrected lidar signals using the water vapour Raman lidar equation (Melfi, 1972; Whiteman et al., 1992; Whiteman, 2003):

$$w(z) = C_w \frac{N_{\mathrm{H_2O}}(z)}{N_{\mathrm{N_2}}(z)} \frac{\Gamma_{\mathrm{N_2}}(z)}{\Gamma_{\mathrm{H_2O}}(z)} \tag{1}$$

where $N_{\mathrm{H_2O,N_2}(z)}$ is the background- and saturation-corrected water vapour and nitrogen photon signals as a function of altitude (z) and $\Gamma_{\mathrm{H_2O,N_2}(z)}$ is the total Raman-backscatter transmissions for the water vapour and nitrogen channels, including molecular and particulate scattering. The molecular transmission values are calculated using the GRUAN-corrected temperature and pressure profiles from the corresponding radiosonde and the Rayleigh cross-sections are determined using the formulae from Nicolet (1984). We do not correct for aerosols as they are considered to have a very small contribution to the overall mixing ratio (Whiteman et al., 1992).

Whiteman (2003) discussed the necessity of accounting for aerosol transmissions, as the presence of aerosols can create uncertainties in the lidar profiles of up to 4%, depending on the aerosol load. Therefore, to minimize this effect on our calibration constants, we have calculated the aerosol extinctions using the RALMO backscatter ratio product which is calculated by taking the ratio of the elastic backscatter signal to the sum of the pure rotational Raman signals (Whiteman, 2003). Similarly to the method followed in Sica and Haefele (2016), we calculate the extinction profile ($\alpha_{aer}(z)$ using the following equation:

$$\alpha_{aer}(z) = LR(z)(\beta_{mol}(z)(BSR(z) - 1)), \tag{2}$$

where $LR(z)$ is the assumed lidar ratio profile, $\beta_{mol}(z)$ is the molecular backscatter profile taken from the NCEP model, and $BSR(z)$ is the backscatter ratio profile. The lidar ratio profile is a step function with a constant value of 50 in the boundary layer and 20 in the free troposphere. The height of the boundary layer is estimated using the backscatter ratio profile. The assumed lidar ratios are climatological values which have been based on the typical aerosols detected using the co-located Precision Filter Radiometer (PFR). The aerosol transmissions for water vapour and nitrogen were calculated using the following:

$$\Gamma_{aer,X} = e^{-\tau} = \exp\left[-\left(\frac{\lambda_X}{\lambda_0}\right)^A \int \alpha_{aer}(z)dz \right] \tag{3}$$

Where $\Gamma_{aer,X}(z)$ is the aerosol transmission profile for a given molecule X (e.g. N_2 or H_2O), and the optical depth is τ_X. The wavelength for a particular channel is λ_X, while λ_0 is the reference extinction profile, which is 354.7 nm for the elastic channel. The Ångstrom exponent, A, is assumed constant with altitude. The Ångstrom exponent is also measured during the daytime using the co-located PFR. However, it is not calculated 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 Ångstrom 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 Ångstrom exponents. The standard deviation of the residuals was ± 0.34 and was used as the uncertainty for the Ångstrom exponents. The uncertainty in the calibration due to our assumptions of the aerosol extinction and the Ångstrom exponent is small, and is included in the uncertainty budget for the calibration constant discussed in Sect. 6.

RALMO uses a polychromator with a bandpass of 0.3 nm (Simeonov et al., 2014). RALMO was designed to minimize temperature dependence and the central wavelengths of the water vapour and nitrogen channels were chosen accordingly. The central wavelengths of RALMO's water vapour and nitrogen channels were chosen to minimize temperature dependence of the Raman cross-section. Dinoev et al. (2013) showed that the nitrogen channel had a relative change in transmitted intensity of 0.4% per 100 K and the water vapour channel intensity changed by roughly 1% when varied between -60°C and $+40$°C.

The calibration constant C_w is defined as:

$$C_w = 0.781 \frac{M_{\mathrm{H_2O}}}{M_{Air}} \frac{\eta_{\mathrm{N_2}}}{\eta_{\mathrm{H_2O}}} \frac{O_{\mathrm{N_2}(z)}}{O_{\mathrm{H_2O}}(z)} \frac{\sigma_{\mathrm{N_2}}(T(z))}{\sigma_{\mathrm{H_2O}}(T(z))} \frac{F_{\mathrm{N_2}}(T(z))}{F_{\mathrm{H_2O}}(T(z))} \quad (4)$$

(Whiteman et al., 1992).

The calibration constant contains all scaling constants and unknown factors, such as the fraction of nitrogen molecules in air, 0.781 , the molecular weights of water and dry air ($M_{H_2O,Air}$), the system efficiency of the nitrogen and water vapour channels ($\eta_{\mathrm{N_2,H_2O}}$), the overlap function for both channels ($O_{\mathrm{N_2,H_2O}}(z)$),the Raman cross-section for each molecular species ($\sigma_{\mathrm{N_2,H_2O}}(T(z))$), and the temperature dependency of the Raman cross-section ($F_{\mathrm{N_2,H_2O}}(T(z))$). In RALMO's case, the differential overlap is designed to be unity (Dinoev et al., 2013; Simeonov et al., 2014).

3.2 Correlated and Weighted Least Squares Fitting of the Lidar Water Vapour Mixing Ratio to the Radiosonde Water Vapour Mixing Ratio

After calculating the ratio of the corrected lidar signals the uncalibrated mixing ratio profiles from the ratio of the two lidar signals (Section 2.2), we use a correlated and weighted least squares fit to normalize the lidar profile to the radiosonde and find the calibration constant (Dionisi et al., 2010; Whiteman et al., 2012). The radiosonde relative humidity profile is transformed into water vapour volume mixing ratio using the standard WMO conversion (World Meteorological Organization (WMO), 2014). The calibration range extends from the surface (491 m ASL) to roughly 7 km ASL depending on the profile. The bottom limit is the first lidar altitude bin at the surface, and the final calibration altitude is determined by the SNR and integration limits we impose. We remove scans at all altitudes where the trajectory spent less than 5 min in the lidar region due to their SNR values being less than 2. This cutoff typically results in the calibration region extending to 7 to 8 km altitude. To ensure that the calibration constant is not biased by a vertical displacement of the air parcel between the lidar and the radiosonde volume, we require the resulting uncalibrated lidar and the radiosonde mixing ratio profile to be correlated to greater than 90% to have a correlation coefficient which minimizes the variance of the fit's residuals and must be higher than 0.75. Several fits are made using correlation coefficient thresholds between 0.75 and 0.9 and the fit with the minimum variance in the residuals is chosen for the final calibration constant.

A moving window of 300 m is run over both the radiosonde and lidar profile, and the cross correlation between the two profiles inside each window is determined. To reduce the effect of noise on the cross correlation, both profiles are smoothed beforehand with a boxcar filter of 101.5 m width. In less than one-third of the cases, when the radiosonde leaves the lidar region early, or the wind is such that the air is spending less than 5 min in the lidar region, a large portion of the profile may be cut off. If there are less than 3 windows (900 m) available for calibration then the radiosonde and the lidar are smoothed to 22.5 m and windows of 100 m are used instead.

If the correlation between the radiosonde and lidar mixing ratios within each window is higher than the correlation threshold then that window's altitude range is accepted for calibration (Dionisi et al., 2010; Whiteman et al., 2012). If there is less than 900 m of data available for calibration at the end of the correlation process, then we do not use that night for calibration as it does not have enough data with which to accurately calibrate. This criterion caused 2 out of the 76 nights to be rejected. While the correlation is calculated on the smoothed profiles, the fit is done by using the native resolution of the lidar inside the accepted calibration windows with a requirement of at least 243 points. The least squares fit is conducted over all of the points selected in the cross-correlation procedure. Each fitting point is weighted by the inverse of the sum of the variances of the water vapour mixing ratio percent uncertainty and the average radiosonde mixing ratio uncertainty. The lidar water vapour mixing ratio statistical uncertainty ($\sigma_{w,stat}$) is propagated from the water vapour and nitrogen channel statistical uncertainties using Eq. 5 (Melfi, 1972). The radiosonde mixing ratio percent uncertainties ($\sigma_{MR,Radiosonde}$) are calculated using the total relative humidity, temperature, and pressure uncertainties reported for each GRUAN flight and propagating through the Hyland and Wexler (1983) equations for saturation vapour pressure to calculate mixing ratio while assuming the relative humidity to be over water (Dirksen et al., 2014; Immler et al., 2010). However, the GRUAN processing occasionally does not report pressure uncertainties below 15 km. Therefore, it was necessary to create a nightly average pressure uncertainty profile which was used on nights where pressure uncertainties were not reported. The variation in the pressure uncertainties was on the order of 0.01%, therefore this assumption is justified. The calibration constant is then determined by using a one-parameter weighted least-squares fit of the form shown in Eq. 1.

4 The Radiosonde Trajectory Method

The radiosonde trajectory method begins with the same procedure as the traditional method, where each of the scans are filtered for clouds or abnormally large background levels, as is discussed in Sect. 2.2. However, instead of choosing the first 30 scans after the radiosonde launch, the scans are chosen based on the radiosonde's movement with respect to the air mass and the wind direction measured by the radiosonde.

4.1 Tracking Air Parcels

The flow chart of the calibration process for both methods is shown in Fig. 4. This section will discuss each step in the calibration process in three pieces. First, we will explain the air parcel tracking and how we choose the lidar measurements which match the radiosonde's. Second, we will discuss the conversion of the lidar measurements into water vapour mass mixing ratio units. Lastly, we discuss how we choose the relevant regions for calibration using cross-correlation and the least square fit.

After the measurements have been cloud-filtered and put on the same altitude grid, we use the radiosonde wind speed measurements to track the air parcels measured by the radiosonde. First, we use the latitude and longitude of the radiosonde, as calculated by the on-board GPS system, as the initial position for air parcel tracking. The air parcel is then assumed to have traveled in a straight line until the time it was measured by the radiosonde. All GRUAN radiosondes report their geophysical coordinates, however radiosondes before 2011 were not GRUAN-processed and did not report their latitude or longitude coordinates. If the geophysical coordinates of the radiosonde are reported, We then transform the coordinates onto a Euclidean grid with the lidar located at the origin using the local flat Earth approximation, which is appropriate when distances are shorter than 20 km (Daidzic, 2017; Smart, 1977) Otherwise, when the geophysical coordinates are not available, we infer the radiosonde position from the wind measurements and the coordinates of the launch site. When inferring the radiosonde position, we assume linear geometry due to the relatively short distances traveled by the radiosonde between each measurement. The radiosonde positions are required to calculate the 2D air parcel trajectories. We do not explicitly consider the vertical movement of the air parcel in this method. Users may need to consider other distance conversion methods, such as the haversine conversion (Smart, 1977), if utilizing radiosondes which are not launched from the same site as the lidar station. Additionally, it may be more appropriate in such cases to use a wind field measurement or model to conduct the trajectory calculation.

RALMO's field-of-view projects to a circle of approximately 1 m diameter at 5 km altitude, an area too small for most trajectories to pass directly through. Therefore, it was necessary to construct a region of assumed horizontal homogeneity in which the water vapour mixing ratio is constant. In order to maintain significant lidar SNR, we defined the homogeneous region, hereafter called the "homogeneous lidar region", to be a circle with a radius of 3 km radius around the lidar. The size of the homogeneous lidar region was chosen by varying the radius of the cylinder from a range of 1 - 25 km and finally increasing it to infinity. Radii below 3 km had very low SNR's and did not have enough altitude coverage to do an effective calibration. Radii above 3 km had large enough SNRs, however, they started to exhibit biases due to over-integrating at some altitudes and lost small features that had previously been visible. The 3 km radius provided the highest SNR, with profiles closest to the radiosonde measurements. In order to maintain a SNR in the water vapour channel greater than 2 above 7 km altitude for the majority of the cases, we defined the homogeneous lidar region to be a circle of 3 km radius centered around the lidar. The size of the homogeneous region was chosen by varying the radius from a range of 1 - 25 km and finally increasing it to infinity. Radii below 3 km resulted in SNRs smaller than 2 below 7 km and in some cases halved the SNR of the water vapour channel at altitudes below 5 km, which decreased the altitude coverage for the calibration and increased the noise in the primary calibration region. While radii above 3 km resulted in SNRs larger than 2 above 7 km, the water vapour profiles started to exhibit biases due to using too long integration times at certain altitudes and losing small features which had previously been visible. The 3 km radius provided the most altitude coverage with profiles closest to the radiosonde measurements and is the best compromise.

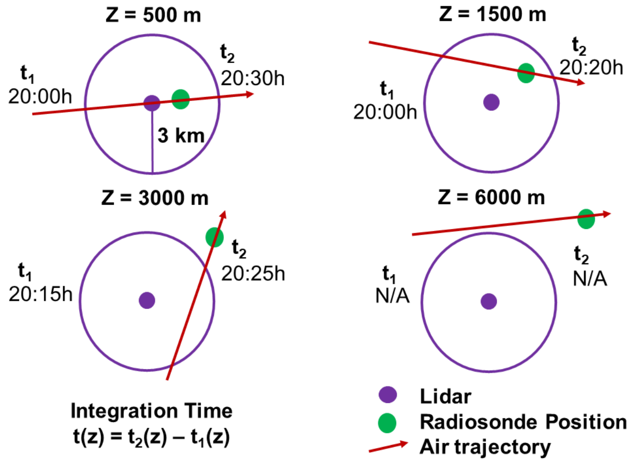


Figure 1. Trajectory calculation and scan selection hypothetical example. The purple circle around the lidar has a 3 km radius and represents the region in which we assume the humidity field is horizontally homogeneous. The green dot is the radiosonde position, the purple dot is the lidar position, and the red arrow is the air parcel trajectory. The variable $\mathbf{z}$ refers to altitude, $\mathbf{t_1}$ is the entry time, and $\mathbf{t_2}$ is the exit time from the 3 km radius. The integration time, $t(z)$, is the total time that the air parcel spends inside the homogeneous region. When the air parcel trajectory does not intersect with the circle, then no data is available for calibration.

Fig. 1 shows how air parcels will always be “seen" by the lidar if the radiosonde remains inside the 3 km radius, whereas any air measured outside the radius may not intersect with the lidar region. If the trajectories do not enter the region, we do not use these altitudes for calibration. The entry and exit times from the homogeneous region mark the first and final scans used to calculate the lidar water vapour mixing ratios, with a maximum of 30 min of integration in order to accurately compare with the traditional technique, which uses a standard 30 min summation across all altitudes (Dinoev et al., 2013; Leblanc et al., 2012; Whiteman et al., 1992; Melfi, 1972). The standard thirty minute integration is the average time it takes a radiosonde to reach the tropopause at mid-latitudes, and therefore generally covers the primary calibration altitudes. Integrating for longer than 30 min is too long to capture the water vapour field variability viewed by the radiosonde. If the total time spent inside the homogeneous region exceeds 30 min, we take ±15 min around the time of closest approach to the lidar. The variation of the integration length with altitude is shown in Fig. 2. The integration time will decrease with altitude for two reasons: higher wind speeds and the air parcel trajectories may intersect with the outer edges of the homogeneous region and are therefore inside for shorter time spans (Fig. 1). The decrease in integration time with altitude will also change depending on the rate at which the radiosonde moves away from the lidar. The majority of nights had integration times less than 5 min above 7 km. However, if the wind is strong at a particular altitude, sharp decreases in integration times may be seen as the radiosonde moves quickly away. It is also possible to see the integration times decrease and then increase again as the radiosonde drifts in and out of the homogeneous lidar region.

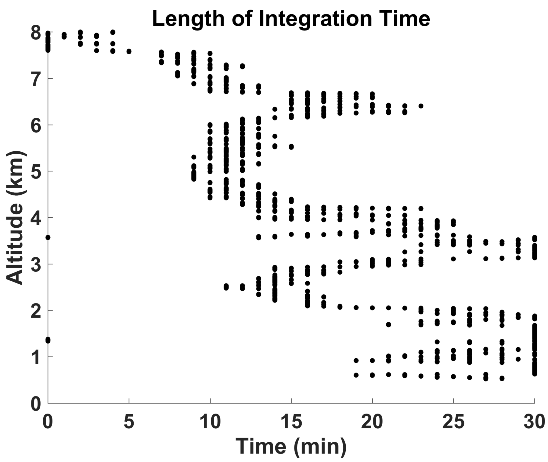


Figure 2. Example integration times from 22 July 2015. The lidar water vapour integration period is determined by the length of time the air parcels spend inside the homogeneous region. The integration time will decrease with altitude due to higher wind speeds. The maximum integration time is 30 min, in order to properly compare with the traditional analysis.

Once the appropriate scans have been chosen by the trajectory analysis, they are integrated to form the raw water vapour and nitrogen profile. The same procedure as in the traditional method (Sect. 4.2 and Sect. 4.3) are then followed to calculate the ratio of the two channels' profiles, find the appropriate calibration regions, and derive the calibration constant. The final calibrated water vapour profile for 22 July 2015 is shown in Fig. 3. The correlation algorithm selected 84% of the profile above 1.5 km to use for the calibration while regions with high variability were excluded from the calibration. The calibrated profile closely follows the radiosonde profile, with differences fluctuating between 5% and 20% over all altitudes. The standard error uncertainty of the slope from the weighted fit is the uncertainty in the calibration constant due to measurement noise. The accuracy to which we know the calibration constant will be discussed further in Section 6.

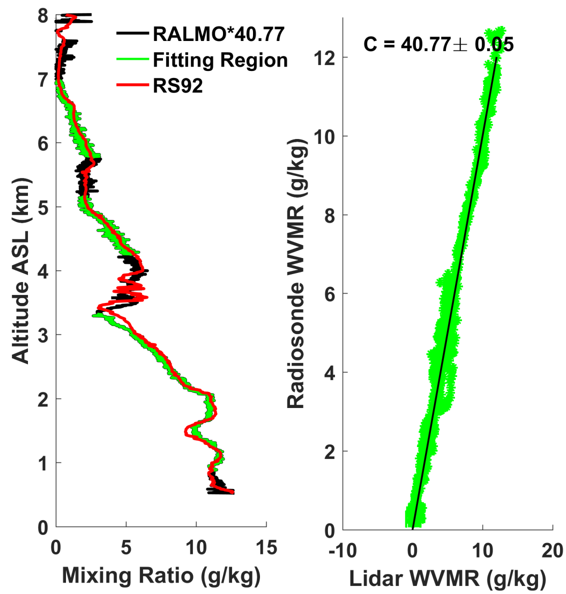


Figure 3. Left: The final trajectory-calibrated profile for 22 July 2015. The lidar profile is in black, the radiosonde is in red. The correlation calibration regions are shown by the overlaid green points. Right: The least square fit of the green points in the left panel. The uncertainty of the calibration constant is the standard error of the slope calculated from the weighted least squares fit.

For clarity on the similarities and differences between the traditional and trajectory methods, a flow chart of the calibration process for both methods is shown in Fig. 4. The main difference is the selection of the appropriate scans for calibration at the beginning. Both methods use the same process for choosing the appropriate calibration regions and calculating the calibration constant.

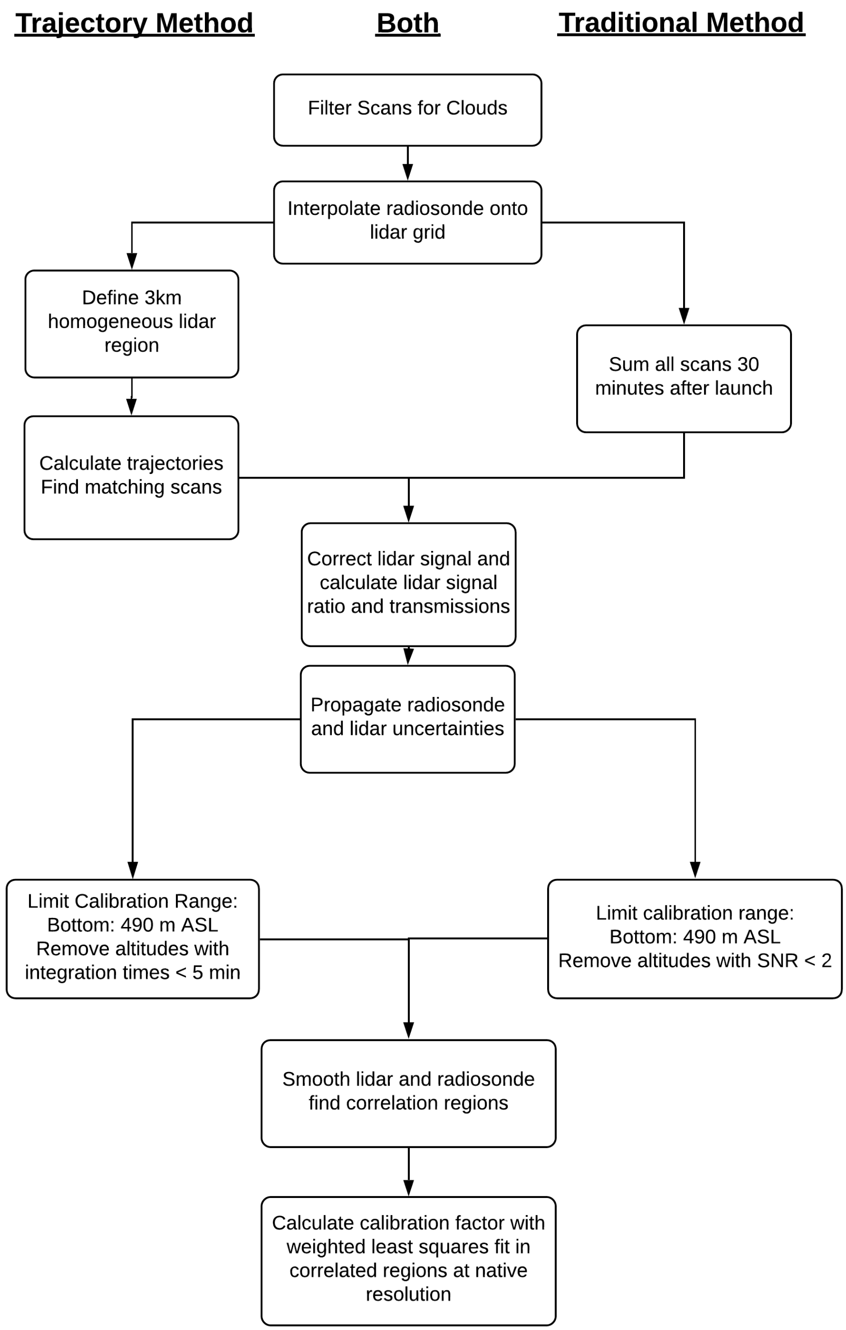


Figure 4. Flowchart of the steps to calibrate the RALMO lidar by the Trajectory Method (left) and the Traditional Method (right).

4.2 Calculation of the Water Vapour Mixing Ratio for RALMO Measurements

The water vapour mixing ratio (w) for the RALMO is calculated from the background- and saturation-corrected lidar signals using the water vapour Raman lidar equation (Melfi, 1972; Whiteman et al., 1992; Whiteman, 2003):

$$w(z) = C_w \frac{N_{\mathrm{H_2O}}(z)}{N_{\mathrm{N_2}}(z)} \frac{\Gamma_{\mathrm{N_2}}(z)}{\Gamma_{\mathrm{H_2O}}(z)} \tag{5}$$

where $N_{\mathrm{H_2O,N_2}(z)}$ is the background- and saturation-corrected water vapour and nitrogen photon signals as a function of altitude (z) and $\Gamma_{\mathrm{H_2O,N_2}(z)}$ is the downward transmissions for the water vapour and nitrogen channels. The transmission values are calculated using the GRUAN-corrected temperature and pressure profiles from the corresponding radiosonde and the Rayleigh cross-sections are determined using the Nicolet (1984) formulae (Nicolet, 1984). We do not correct for aerosols as they are considered to have a very small contribution to the overall mixing ratio (Whiteman et al., 1992). RALMO uses a polychromator with a bandpass of 0.3 nm (Simeonov et al., 2014). RALMO was designed to minimize temperature dependence and the central wavelengths of the water vapour and nitrogen channels were chosen accordingly. The central wavelengths of RALMO's water vapour and nitrogen channels were chosen to minimize temperature dependence. Dinoev et al. (2013) showed that the nitrogen channel had a relative change in transmitted intensity of 0.4% per 100 K and the water vapour channel intensity changed by roughly 1% when varied between $-60°$C and $+40°$C.

The calibration constant C_w is defined as (Whiteman et al., 1992):

$$C_w = 0.781 \frac{M_{\mathrm{H_2O}}}{M_{Air}} \frac{\eta_{\mathrm{N_2}}}{\eta_{\mathrm{H_2O}}} \frac{O_{\mathrm{N_2}(z)}}{O_{\mathrm{H_2O}}(z)} \frac{\sigma_{\mathrm{N_2}}(T(z))}{\sigma_{\mathrm{H_2O}}(T(z))} \frac{F_{\mathrm{N_2}}(T(z))}{F_{\mathrm{H_2O}}(T(z))} \tag{6}$$

The calibration constant contains all unknown factors, such as: the fraction of nitrogen molecules in air, 0.781, the molecular weights of water and dry air ($M_{H_2O,Air}$), the system efficiency of the nitrogen and water vapour channels ($\eta_{\mathrm{N_2,H_2O}}$), the overlap function for both channels ($O_{\mathrm{N_2,H_2O}}(z)$),the Raman cross-section for each molecular species ($\sigma_{\mathrm{N_2,H_2O}}(T(z))$), and the temperature dependency of the Raman cross-section ($F_{\mathrm{N_2,H_2O}}(T(z))$). In RALMO's case, the ratio of the overlap between the two channels is designed to be unity (Dinoev et al., 2013; Simeonov et al., 2014).

4.3 Correlated and Weighted Least Squares Fitting of the Lidar Water Vapour Mixing Ratio to the Radiosonde Water Vapour Mixing Ratio

After calculating the ratio of the corrected lidar signals the uncalibrated mixing ratio profiles from the ratio of the two lidar signals (Section 2.2), we use a correlated and weighted least squares fit to normalize the lidar profile to the radiosonde and find the calibration constant (Dionisi et al., 2010; Whiteman et al., 2012). The radiosonde relative humidity profile is transformed into water vapour volume mixing ratio using the standard WMO conversion (World Meteorological Organization (WMO), 2014). The calibration range extends from 500 m above sea level (ASL) to roughly 7 km ASL depending on the profile. The bottom limit is the first lidar altitude bin at 490 m above sea level, and the final calibration altitude is determined by the SNR and integration limits we impose. We remove scans at all altitudes where the trajectory spent less than 5 min in the lidar region due to their low SNRs which typically results in the calibration region ending around 7 or 8 km. To ensure that the calibration constant is not biased by a vertical displacement of the air parcel between the lidar and the radiosonde volume, we require the resulting uncalibrated lidar and the radiosonde mixing ratio profile to be correlated to greater than 90%.

A moving window of 300 m is run over both the radiosonde and lidar profile, and the cross correlation between the two profiles inside each window is determined. To reduce the effect of noise on the cross correlation, both profiles are smoothed beforehand with a boxcar filter of 101.5 m width. In less than one-third of the cases, when the radiosonde leaves the lidar region early, or the wind is such that the air is spending less than 5 min in the lidar region, a large portion of the profile may be cut off. In which case, if there are less than 3 windows (900 m) available for calibration then the radiosonde and the lidar are smoothed to 22.5 m and windows of 100 m are used instead.

If the correlation between the radiosonde and lidar mixing ratios within each window is higher than 90% then that window's altitude range is accepted for calibration (Dionisi et al., 2010; Whiteman et al., 2012). If there is less than 900 m of data available for calibration at the end of the correlation process, then we do not use that night for calibration as it does not have enough data with which to accurately calibrate. While the correlation is calculated on the smoothed profiles, the fit is done by using the native resolution of the lidar inside the accepted calibration windows with a requirement of at least 243 points. The least squares fit is conducted over all of the points selected in the cross-correlation procedure. Each fitting point is weighted by the sum of the variances of the water vapour mixing ratio percent uncertainty and the average radiosonde mixing ratio uncertainty. The lidar water vapour mixing ratio statistical uncertainty ($\sigma_{w,stat}$) is propagated from the water vapour and nitrogen channel statistical uncertainties (Melfi, 1972) using Eq. 5. The radiosonde mixing ratio percent uncertainties ($\sigma_{MR,Radiosonde}$) are calculated using the total relative humidity, temperature, and pressure uncertainties reported for each GRUAN flight and propagating through the Highland and Wexler (1983) mixing ratio conversion while assuming the relative humidity to be over water (Hyland

and Wexler, 1983; Dirksen et al., 2014; Immler et al., 2010). However, the GRUAN processing occasionally does not report pressure uncertainties below 15 km. Therefore, it was necessary to create a nightly average pressure uncertainty profile which was used on nights where pressure uncertainties were not reported. The variation in the pressure uncertainties was on the order of 0.01%, therefore this assumption is justified. The calibration constant is then determined by using a one-parameter weighted least-squares fit of the form shown in Eq. 5.

Once the appropriate scans have been chosen by the trajectory analysis, they are integrated to form the raw water vapour and nitrogen profile. The same procedure as in the traditional method (Sect. 4.2 and Sect. 4.3) are then followed to calculate the ratio of the two channels' profiles, find the appropriate calibration regions, and derive the calibration constant. The final calibrated water vapour profile for July 22, 2015 is shown in Fig. 3. The correlation algorithm selected 84% of the profile above 1.5 km to use for the calibration while regions with high variability were excluded from the calibration. The calibrated profile closely follows the radiosonde profile, with differences fluctuating between 5% and 20% over all altitudes. The standard error uncertainty of the slope from the weighted fit is the uncertainty in the calibration constant due to measurement noise. The accuracy to which we know the calibration constant will be discussed further in Section 6.

For clarity on the similarities and differences between the traditional and trajectory methods, a flow chart of the calibration process for both methods is shown in Fig. 4. The main difference is the selection of the appropriate scans for calibration at the beginning. Both methods use the same process for choosing the appropriate calibration regions and calculating the calibration constant.

5 Comparing the Traditional and Trajectory Methods

We applied the trajectory technique to 76 nights between January 2011 and December 2016 in which 31 were removed due to lack of lidar measurements during the radiosonde launch window, primarily due to precipitation or repairsroutine maintenance. From the 45 remaining nights, the trajectory calibration and traditional method automatically removed 8 nights due to abnormally high background values above 0.01 counts/bin/s. An additional 13 nights were removed from both the trajectory and traditional calibrations due to low signal-to-noise levels (below 1 SNR) and and the presence of clouds. The filtering process removed all of the nighttime flights from 2008 to 2011 due to significant cloud cover coincident with the radiosonde launch. A final list of the nights with their calibration constants is shown in Table 1. We found 12 out of the 24 remaining calibration nights exhibited significant disagreement between the traditional and the trajectory calibrations, the reasons for which are discussed below. While comparing the calibration constants from the two methods, it became apparent that we could separate them into two groups when observing the water vapour mass mixing ratio contours over the course of the calibration period.

One set of nights exhibited water vapour fields which were horizontally homogeneous around the lidar over the course of the 30 minute calibration period, and were thereby dubbed "homogeneous" nights. The second set of nights showed movement of water vapour layers over 100 m in altitude over the course of 30 min, and were called "heterogeneous" nights. Table 1 has been divided into the two categories and shows the calibration constants for each night and calibration technique.

Homogeneous	$C_{trad} \pm \Delta C_{trad}\%$	$C_{traj} \pm \Delta C_{traj}\%$	**Abs(Difference)**	**Abs(Percent Difference)**	**Comments**
2011.10.05	38.61 ± 4.5	38.42 ± 5.4	0.19	0.49	
2012.07.18	39.67 ± 4.5	39.75 ± 4.5	0.08	0.20	
2012.08.09	40.42 ± 4.7	40.12 ± 4.7	0.3	0.75	
2012.08.29	39.23 ± 4.4	39.32 ± 4.4	0.09	0.23	
2012.12.13	40.30 ± 5.3	40.57 ± 5.3	0.27	0.67	
2013.04.24	41.77 ± 4.8	41.49 ± 4.8	0.28	0.67	
2013.06.05	41.77 ± 4.4	41.67 ± 4.4	0.1	0.24	
2014.01.23	41.31 ± 4.5	41.36 ± 4.6	0.05	0.12	
2014.03.21	40.39 ± 5.5	38.31 ± 5.2	2.08	5.42	Trajectory calibration includes points below 1 km and Traditional does not.
2014.07.18	40.05 ± 5.2	39.93 ± 5.2	0.12	0.30	
2015.11.11	41.50 ± 3.7	41.72 ± 3.6	0.22	0.53	
2016.03.09	45.42 ± 4.5	45.67 ± 4.5	0.25	0.55	
2016.08.24	43.95 ± 6.2	44.13 ± 6.3	0.18	0.41	
Heterogeneous	$C_{trad} \pm \Delta C_{trad}\%$	$C_{traj} \pm \Delta C_{traj}\%$	**Abs(Difference)**	**Abs(Percent Difference)**	**Comments**
2012.02.29	36.50 ± 4.8	35.83 ± 4.8	0.67	1.87	
2012.05.25	37.94 ± 4.6	37.11 ± 4.6	0.83	2.24	
2012.07.27	39.90 ± 5.1	39.39 ± 5.1	0.51	1.29	
2013.06.18	40.72 ± 4.5	39.68 ± 4.6	1.04	2.62	
2015.06.26	41.05 ± 3.8	41.16 ± 3.8	0.11	0.26	Calibration done over the same stable homogeneous regions.
2015.07.22	39.99 ± 3.9	41.33 ± 3.8	1.34	3.24	
2016.03.23	45.09 ± 5.2	45.41 ± 5.2	0.32	0.71	
2016.04.07	38.97 ± 2.7	39.95 ± 2.7	0.98	2.45	
2016.09.09	44.37 ± 4.5	43.72 ± 4.5	0.65	1.49	
2016.10.06	44.30 ± 3.8	44.59 ± 3.8	0.29	0.65	
2016.11.17	46.53 ± 3.5	48.07 ± 3.4	1.54	3.20	

Table 1. A comparison of the calibration constants of all nights used in this study. The table is broken into two sections- homogeneous and heterogeneous calibration nights. Column 1 is the date on which the radiosonde was launched. Column 2 or C_{trad} is the traditional calibration constant with its total percent uncertainty. Column 3 or C_{traj} is the trajectory method calibration constant with its percent uncertainty. Column 4 is the absolute value of the difference between the two constants. Column 5 is the absolute value of the percent difference of the two constants with respect to the traditional calibration constant. Column 6 is for comments regarding the differences. Two nights in the homogeneous section presented larger differences from the rest of the nights due to using different calibration regions. Three nights in the heterogeneous group had very small differences in their calibration constant due to using similar regions for calibration, despite the variability in the water vapour.

We compared the trajectory method result to the traditional technique discussed in the previous section in which the radiosonde movement is not taken into account, and all altitudes are integrated for 30 min after the launch. For the homogeneous nights, we hypothesized that if It became apparent that if the water vapour field is stable for long periods of time and experiences very little change over the distance traveled by the radiosonde, then the radiosonde and the lidar should measure roughly similar water vapour content. Therefore, we should see good agreement small differences between the traditional and trajectory methods' and small changes in the calibration constants. Of the 24 calibrations that passed cloud filtering and had enough calibration regions, 12 dates showed good agreement in their profiles when compared to the radiosonde due to stable or homogeneous water vapour conditions around the location of the lidar. A subset of these nights are shown in Fig. 6 and we have labeled these nights as "homogeneous" or "stable" nights in Table 1. While both methods will produce similar profiles and calibration constants on stable homogeneous nights, the two may not share the same calibration constants due to using different lidar scans (Fig. 5). The traditional method uses all profiles from the radiosonde launch to 30 min after launch, which are all scans inside the two dashed red lines in Fig. 5. The trajectory technique will choose the appropriate calibration scans based on each air parcel's trajectory and its position of closest approach shown by the scans between the magenta dots. Consequently, the trajectory method will not include measurements from altitudes where the air parcel trajectories do not intersect with the homogeneous lidar region. Table 1 is divided into homogeneous and heterogeneous nights. The majority of the homogeneous nights have a percent difference in their calibration constants from the traditional method of less than 0.51%. However, two nights (December 12 2012 and April 21 2014) show large differences and this is due to using different calibration regions in the trajectory method. The average percent difference in the calibration constants is $0.4 \pm 0.3\%$ when not considering the two anomalous nights, but increases to $1.2 \pm 2.0\%$ when they are included.

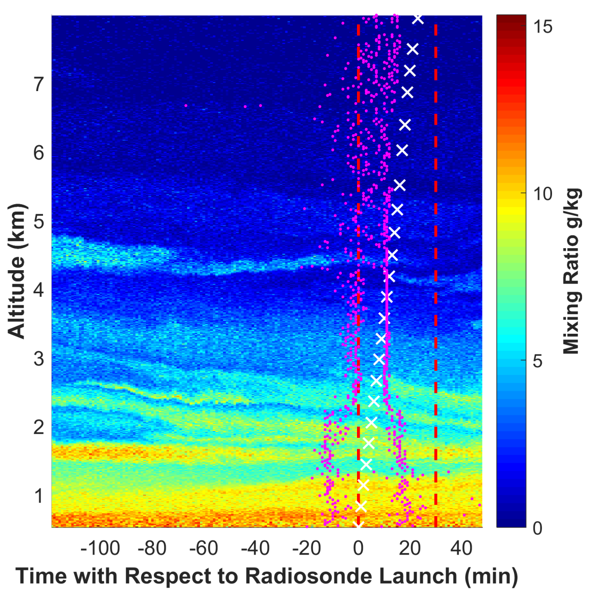


Figure 5. Lidar water vapour mixing ratio measurements on 2012-07-27 00:00 UTC. The time axis is measured relative to the radiosonde launch at 0 m. The traditional method uses all scans between the two red dashed lines. The trajectory method uses all measurements between the magenta dots. The white "x" markers show the height of the radiosonde with time.

A subset of the homogeneous nights are shown in Fig. 6. The first column shows the percent difference from the mean water vapour profile over the 2 hours shown, with the calibration time between the two dashed red lines. The second and third columns are the percent difference between the calibrated lidar water vapour mixing ratio measurements and the radiosonde measurements for the traditional and trajectory techniques, respectively. The pink regions in the figures of the second and third columns are the calibration regions used for each method. Both methods use similar calibration regions, due to the fact that the methods produce similar lidar water vapour profiles. The strength of the trajectory technique is shown by the reduction of the number of regions with large differences between the lidar and the radiosonde in the traditional method (see the second and third columns of Fig.6). For example, on 18 July 2012, the large difference between the radiosonde and the lidar at approximately 3.8 km is reduced by 20% in the trajectory method. The large difference is caused by the appearance of a water vapour layer halfway through the calibration technique and can be seen in the water vapour contour. Sharp features, as shown on 9 August 2012 at 4 km are produced in both methods due to the sudden stratification of the water vapour layers. The large 10% difference between the radiosonde on 9 August 2012 at 2 km is also reduced by the trajectory method by 5%. On 24 April 2013, there is a large and increasing bias between the radiosonde and the lidar above 4 km which is reduced to oscillate around 0% in the trajectory method. While the trajectory method does reduce the bias in the traditional method on this night at altitudes above 4 km, it does produce larger RMS values at the same altitudes. The increase in RMS is due to the smaller integration times at those altitudes due to the distance of the radiosonde from the lidar. Indeed, on 24 April 2013 the radiosonde is 4 km away from the lidar at 4 km altitude and 12 km away at 8 km altitude. Fast winds and larger distances from the lidar decrease the time the air spends in the lidar region and decreases the chances of intersection which results in shorter integration periods. The majority of the homogeneous nights have a percent difference in their calibration constants from the traditional method of less than 0.51%. However, one night (21 March 2014) showed large differences and this is due to using different calibration regions in the trajectory method. The average percent difference in the homogeneous calibration constants is $0.43 \pm 0.21\%$ when not considering the anomalous night, but increases to $0.81 \pm 1.4\%$ when they are included.

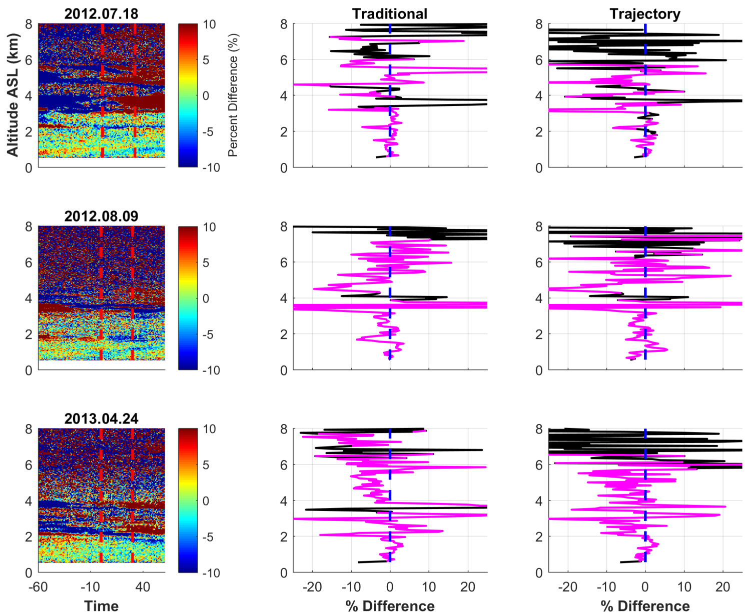


Figure 6. A subset of the dates with largely homogeneous conditions showing the differences between the traditional and trajectory calibration techniques. The first column is the water vapour mixing ratio time series averaged to 15 m altitude bins the percent difference from the mean water vapour mixing ratio profile over the two hours and averaged to 15 m altitude bins. The first red line is the time when the radiosonde was launched. The second red line is 30 min after radiosonde launch and indicates the last profile used for the traditional method. White vertical regions are where scans have been filtered. The second column is the percent difference between the radiosonde and the profile produced using the traditional method. The third column is the percent difference between the radiosonde and the profile produced by the trajectory method. Pink Magenta regions are regions where the correlation between the radiosonde and the lidar are above 90%. During homogeneous conditions, the trajectory and traditional methods show good agreement, with similar percent differences with respect to the radiosonde. Large spikes are regions where the lidar and the radiosonde disagree on layer heights.

Both methods produce profiles that agree well with the radiosonde and have an average bias around 0% with the exception of the night of August 8, 2012 which has an offset of 5% difference with altitude when using the traditional method. The bias on that night is reduced when using the trajectory method. Both methods have difficulty matching the radiosonde at the altitudes where there are sharp changes in water vapour density as shown by the large spikes in Fig. 6.

When the water vapour field is horizontally heterogeneous, meaning water vapour at a given pressure surface fluctuates by 50% or more layers moved over 100 m in altitude over the course of the 30 min traditional calibration period, the trajectory method should better represent the air sampled by the radiosonde than the traditional technique. We define a “heterogeneous" field by movement of water vapour layers over 100 m in altitude over the course of 30 min. Layers on the order of several hundred meters thickness can change in altitude over this period, resulting in water vapour mixing ratios changing over 30% at a given height.

Similarly to the homogeneous nights, a subset of the heterogeneous nights is shown in Fig. 7. The contour of the percent difference from the mean water vapour profile for each night show water vapour layers which change rapidly over the course of the 30 min calibration period (column 1 of Fig. 7). These rapid changes produces large differences in the radiosonde and lidar mixing ratio profiles if the movement of the radiosonde with respect to the air mass is not taken into account. These differences can be on the order of 15% - 20%, as is shown in the second column of Fig. 7, particularly on the night of 22 July 2015. Large differences on this night on the order of ± 10% are reduced to less than 5% by the trajectory method (third column of Fig. 7). Both methods on that night produce sharp differences at 3.8 km due to the sharp change in water vapour content. Above 4 km There is a constant bias between the radiosonde and the lidar in the traditional method of 10% which is reduced to 5% in the trajectory technique. On 27 July 2012 there are large difference features present throughout the entire percent difference profile for the traditional method on the order of 10%. These are similarly reduced to less than 5%, with the exception of the larger spike at 1.5 km caused by the sharp change in water vapour concentration. The night of 25 May 2012 shows less variation than the other two nights, but does have large differences at 4.2 km and 2 km. The higher feature is reduced from -25% difference to 10% by the trajectory method, while the percent difference in the lower feature changes from +20% to +10%.

Similarly to the homogeneous profiles, we do see an increase in noise at the higher altitudes of the trajectory method profiles. This is again cause by the large drift in the radiosonde's position, as well as wind speed. However, below 4 km we do see that the trajectory method reduces the differences between the radiosonde and the lidar by up to 15%.

In general, the percent difference between the radiosonde and the trajectory-calibrated profile on heterogeneous nights is much smaller than the difference between the radiosonde and the traditional method. The trajectory method profile has a smaller standard deviation with altitude and has an average bias of 0%. The traditional method cannot compensate for the rapid changes during the half-hour calibration time, and this results in larger differences between the lidar and radiosonde, on the order of 10 - 20%. The trajectory method does have these large differences above 4 km altitude, as during periods where water vapor is rapidly changing it uses shorter integration periods.

The differences between the calibration constants on the heterogeneous nights is larger than the homogeneous nights due to the difference in calibration regions (Table 1). The average difference in the calibration constants on heterogeneous nights is $2.0 \pm 1.4\%1.82 \pm 1.02$ from the traditional method calibration constant. One night out of the 11 in the heterogeneous nights showed very small differences in the calibration constant despite structural changes throughout the calibration period. This night used similar calibration regions that were also stable over the course of the calibration in both methods. If this night is not included, then the average difference becomes $1.92 \pm 0.93\%$.

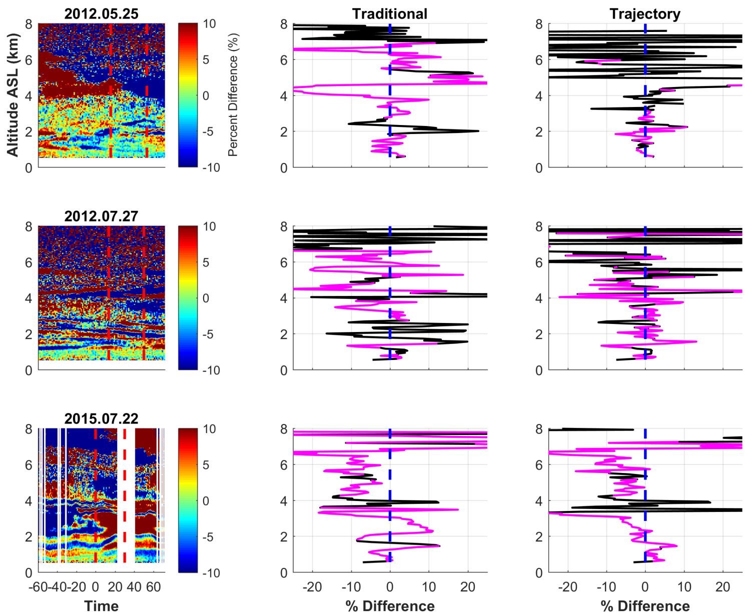


Figure 7. A subset of the dates with largely heterogeneous conditions showing the differences between the traditional and trajectory calibration techniques. The first column is the water vapour mixing ratio time series and the first red line is the time when the radiosonde was launched. The second red line indicates the last scan used in the traditional method. White vertical regions are where scans have been filtered. The second column is the percent difference between the radiosonde and the profile produced using the traditional method. The third column is the percent difference between the radiosonde and the profile produced by the trajectory method.This figure follows the same format as Fig. 6 and This figure shows that when the water vapour field changes over the 30 min traditional calibration period, the traditional water vapour profile can look significantly different from the radiosonde. The trajectory method produces a profile with a smaller percent difference with respect to the radiosonde.

The average and the standard deviation of all percent difference profiles with the radiosonde from the trajectory and traditional method profiles are shown in Fig. 8. The average trajectory bias oscillates around 1%, but the variability increases above 5 km 4.5 km. This is due to the shorter integration times and smaller SNRs at higher altitudes (Fig. 8). The average traditional bias also oscillates around -0.7%, however, the average profile deviates farther from the center than the trajectory method (Fig. 8). The standard deviation of all of the percent difference profiles shows that the trajectory method more accurately fits the radiosonde profile above 2 km on a profile-by-profile basis and will more consistently provide better fits. The standard deviation of the ensemble of percent difference profiles between both calibration methods and the radiosonde shows that the trajectory method has less variability with respect to the radiosonde profile above 2 km. Below 2 km the traditional and trajectory methods produce similar profiles on average, with similar consistency.

While both methods will produce similar profiles on stable nights, the two may not share the same calibration constants due to using different lidar scans (Fig. 5). The traditional method uses all profiles from the radiosonde launch to 30 min after

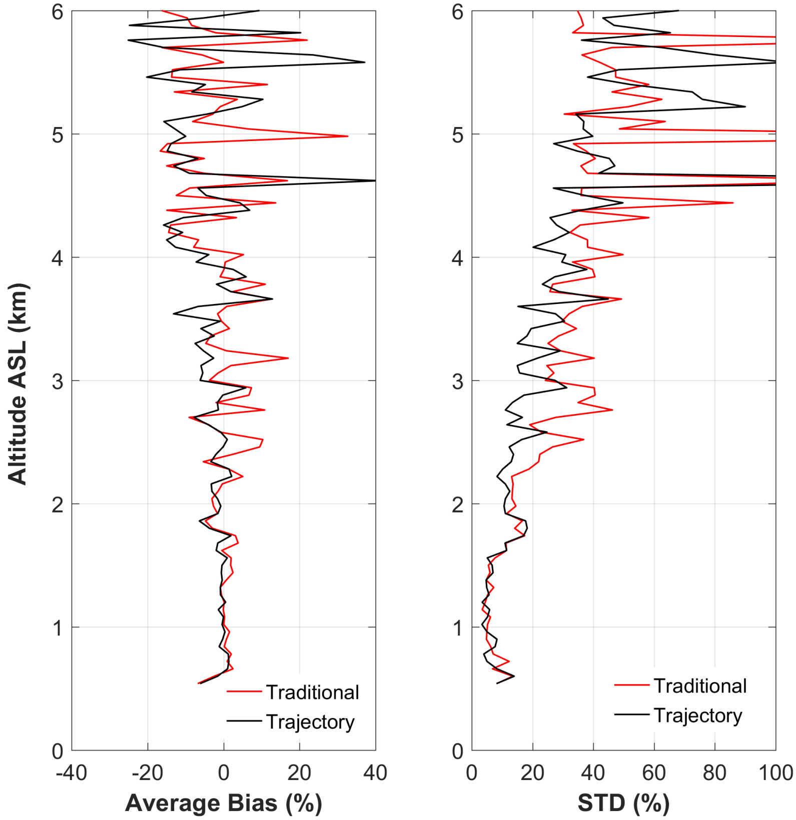


Figure 8. Left panel: The average bias between the radiosonde and the trajectory method calibrated profiles at 25 m vertical resolution for both the trajectory (black) and traditional methods (red). Right panel: The standard deviation of all trajectory percent difference profiles at 25 m resolution for both trajectory (black) and traditional (red) methods.

launch. The trajectory technique will choose the appropriate calibration scans based on each air parcel's trajectory and its position of closest approach. The trajectory method will remove measurements from altitudes where the air parcel trajectories do not intersect with the homogeneous region. Table 1 is divided into homogeneous and heterogeneous nights. The majority of the homogeneous nights have a percent difference from the traditional method of less than 0.5%. However, two nights show large differences and this is due to using different calibration regions in the trajectory method. The average percent difference in the calibration constants is $0.4 \pm 0.3\%$ when not considering the two anomalous nights, but increases to $1.2 \pm 1.95\%$ when they are included.

6 Lidar Calibration Uncertainties for Trajectory and Traditional Methods

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 (Ferrare et al., 1995; Turner et al., 2002; Whiteman et al., 2006; Leblanc and Mcdermid, 2008; Dionisi et al., 2010; David et al., 2017). 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 (or other reference calibration instruments) 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 during a campaign period 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 and its uncertainty is extremely important for long term trend analyses, since such a drift/change could easily be larger than the uncertainty of the calculated trend (Whiteman et al., 2011b). 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.

Previous studies have shown that the largest uncertainty is typically the uncertainty of the reference instrument (Leblanc and Mcdermid, 2008). It was not until recently that such detailed uncertainty budgets became available routinely for radiosonde measurements. The GRUAN radiosonde products are the first radiosonde profiles to have a published uncertainty budget for each measurement as a function of altitude (Dirksen et al., 2014). By using the GRUAN radiosonde product, we are now able to calculate the uncertainty in the calibration constant due to the radiosonde's uncertainties.

We investigated three five major sources of uncertainty in the determination of the calibration constant for both methods: the lidar statistical uncertainty, GRUAN radiosonde mixing ratio uncertainty, and dead time uncertainty, aerosol extinction, and Ångstrom coefficient uncertainties. Most of the calibration uncertainty is due to that of the reference instrument (Leblanc and Mcdermid, 2008). The uncertainty in the calibration constant, the lidar statistical uncertainties, and dead time were identified as the major sources of uncertainty in RALMO water vapour measurements by Sica and Haefele (2016), who also retrieved aerosol extinction, Ångstrom exponents, and their associated uncertainties. In the traditional method, the mixing ratio requires the uncertainty due to aerosol extinction and the Ångstrom exponent to be evaluated, as discussed in Whiteman (2003) and

Kulla and Ritter (2019). The GRUAN radiosonde water vapour mixing ratio uncertainties were calculated from the reported GRUAN total uncertainties for pressure, temperature, and relative humidity, calculated using the Hyland and Wexler 1983 formula for saturation vapour pressure over water The GRUAN radiosonde water vapour mixing ratio uncertainties were calculated using the reported GRUAN total uncertainties (combined statistical and systematic) for pressure, temperature, and relative humidity and by propagating through the Hyland and Wexler 1983 formula for saturation vapour pressure (Hyland and Wexler, 1983; Dirksen et al., 2014). We use an average pressure uncertainty profile calculated from all the nights when the pressure uncertainty is not reported for less than one-third of the nights. The radiosonde relative humidity uncertainties vary between 5% and 10% RH in the troposphere. The pressure uncertainties are on the order of 10^{-3} hPa in the troposphere, and the total temperature uncertainty varies between .1 and .3 K in the troposphere. The radiosonde mixing ratio uncertainties are linearly interpolated onto the lidar's 3.75 m resolution grid for the uncertainty determination.

The lidar mixing ratio statistical uncertainties are propagated through Equation 5 using the random uncertainties from both the water vapour and nitrogen signals. The lidar statistical uncertainties from the trajectory method are smaller than the radiosonde uncertainties below 3 km but are larger than the radiosonde uncertainties, varying from 10% to 20% at and above 4 km from profile to profile.

Both the lidar statistical and radiosonde uncertainties were used as the weights for the least squares fit performed in Sect. 3.2, defined by Eq. 7 (Bevington and Robinson, 2003).

$$C_w = \frac{\sum_{i=1}^{K} \frac{R_i L_i}{\sigma_i^2}}{\sum_{i=1}^{K} \frac{L_i^2}{\sigma_i^2}}, \quad (7)$$

where C_w is the calibration constant, K is the number of points used in the fit, R_i are the radiosonde mixing ratio points used in the calibration, and L_i are the saturation and transmission corrected ratio of water vapour and nitrogen signals, and σ_i are the weights. Using the variance of the residuals of the least-squares fit, one can calculate the uncertainty in the fit, or "fitting uncertainty." This fitting uncertainty is the result of the amount of photon counting noise in the lidar measurements, and can be treated as the uncertainty in the calibration due to the lidar photon counting statistics. The fitting uncertainty is calculated using the standard equations for the slope of a line (Bevington and Robinson, 2003). Each calibration value has some standard uncertainty associated with the weighted fit, hereafter called the fitting uncertainty. The fitting uncertainty arises due to the lidar's photon counting statistics from the lidar's digital water vapour channels. The average trajectory method fitting uncertainty is 0.4% of the average calibration constant. The average fitting uncertainty for the traditional method is 0.3% of the average calibration constant. The traditional method has smaller fitting or statistical uncertainties than the trajectory method due to the larger number of scans used per altitude, on average, compared to the trajectory method. The fitting uncertainty does not encompass the entire uncertainty of the calibration constant, since it is due only to the photon counting noise.

The calibration of a lidar using a radiosonde is limited primarily by the accuracy of the radiosonde measurement. The uncertainty of the water vapour calibration constant due to lidar's random uncertainty and radiosonde's total uncertainty (both systematic and random) systematic and random uncertainties was determined using the uncertainty propagation in Eq. 8 (JCGM, 2008).

$$U_{C_w} = \sqrt{\Sigma_{n=1}^{N}\Sigma_{m=1}^{N}\frac{\partial C_w}{\partial X_n}\frac{\partial C_w}{\partial X_m}cov(X_n,X_m)} \tag{8}$$

where X is the measurement vector including both the radiosonde and lidar measurements (e.g. $X = [L_i, ...Ri, ...]$) used to calculate the calibration constant from Eq. 7 with length $N = 2K$. We make several assumptions in Eq. 8. First, by definition, the covariance of a radiosonde or lidar measurement uncertainty with itself is simply the variance. Second, we assume that the lidar photon counting uncertainties are uncorrelated with each other. Third, we assume that the radiosonde measurement uncertainties are uncorrelated with lidar measurement uncertainties. Lastly, we assume that the radiosonde measurement uncertainties are correlated with each other with a correlation coefficient of $r = 1$. Choosing r equal to unity implies that we are assuming complete correlation and therefore the maximum possible uncertainty. With these assumptions, we can simplify Eq. 8 becomes: as follows:

$$U_{C_w} = \sqrt{\Sigma_{i=1}^{K}(\frac{\partial C_w}{\partial R_i})^2U_R^2 + \Sigma_{i=1}^{K}(\frac{\partial C_w}{\partial L_i})^2U_L^2 + 2\Sigma_{i=1}^{K-1}\Sigma_{j=i+1}^{K}\frac{\partial C_w}{\partial R_i}\frac{\partial C_w}{\partial R_{i+1}}r_{ij}U_iU_j} \tag{9}$$

where $U_{L,R}$ are the corresponding lidar and radiosonde mixing ratio uncertainties, and $U_{i,j}$ are the uncertainties corresponding to the measurement vector X.The derivatives are calculated from Eq. 7. Note that the second term in Eq. 9 is the uncertainty due to the lidar's photon counting uncertainty. This term is the same as the fitting uncertainty discussed in the previous paragraph, and the values agree with each other within a tenth of a percent. The combined uncertainty in the calibration constant due to the radiosonde and lidar uncertainties is an average of 4% for both the trajectory and traditional techniques with signal levels below 15 MHz.

The dead time uncertainty can be large for RALMO, particularly during the daytime. Thus, Eq. 8 must be modified to account for this contribution when it is present. The dead time uncertainty is propagated through Eq. 5 assuming a non-paralyzable system and using Eq. 8. For RALMO, we assume a deadtime uncertainty (U_γ) of 5% or 0.2 ns, which was the standard deviation of the retrieved dead times for all of these nights when using the Optimal Estimation method of Sica and Haefele (2016). The average calibration constant uncertainty due to dead time uncertainty is then 0.3% of the calibration value for both the trajectory and traditional techniques, about equal to the fitting uncertainty. The average total accuracy of both the trajectory and traditional calibration constant including dead time effects is 4% when considering the lidar statistical uncertainty, the radiosonde measurement uncertainties, and the dead time uncertainty, with most of the uncertainty due to the radiosonde measurement.

The uncertainty in the calibration constant due to the uncertainty in the extinction profile is calculated using Eq. 8. The uncertainties for the extinction were assumed to be 100% to determine an upper-limit uncertainty contribution. However, the derivatives of the calibration constant with respect to the individual extinction values were so small that the uncertainty contribution from the extinction was consistently less than 0.01% for all cases. The larger uncertainty component in the extinction is the calibration uncertainty due to the assumption of the Ångstrom exponent. The uncertainties in the Ångstrom exponent values were estimated by detrending the time series of these measurements from 2011-2015 using a summation of a 6 and 12 month sinusoid to the Ångstrom exponent measurements. The standard deviation of the fit's residuals was 0.34. The uncertainty in the calibration constant due to the uncertainty in the Ångstrom exponent was then calculated to be $0.4 \pm .5\%$. While on average it is only an order of magnitude larger than the extinction uncertainty component, the Ångstrom exponent contributes more to the uncertainty when more aerosols are present. The maximum contribution from the Ångstrom exponent was 1.8% on 23 March 2016 due to the presence of a stronger aerosol layer. The rest of the nights had either no aerosols present or weakly interacting aerosol layers resulting in lower uncertainty contributions from the Ångstrom exponent.

Another possible contributor to the total calibration constant uncertainty is the overlap function and aerosol layers. RALMO is designed to have no differential overlap in the water vapour and nitrogen channels and an overlap ratio between the nitrogen and water vapour signals of unity. However, a small differential overlap could result from chromatic aberration from the protective windows and edge filters (Dinoev et al., 2013).

In many situations the uncertainty of the calibration constant is estimated using the standard deviation of the calibration constant over a time period of days (Leblanc and Mcdermid, 2008; Dionisi et al., 2010). However, using GRUAN radiosonde measurement uncertainties allows us to calculate uncertainties on a nightly basis and does not require a time series. We also compared the average of the nightly calibration results to the standard deviation of the entire 24–night RALMO calibration time series used in this study. The RALMO system is known to have differential aging of its photomultipiers which causes the calibration to drift (Simeonov et al., 2014). A linear fit was made to the calibration time series and then removed The calibration time series was de-trended to calculate the standard deviation of the calibration over 10 6 years. The standard deviations for both the traditional and trajectory time series techniques were 4.5%, thereby agreeing with the average nightly uncertainty. The results from calculating the standard deviation of the time series shows that the typical methods used in calibration campaigns will generally give the same result as taking the average uncertainty of the individual uncertainties. However, we would suggest that taking the individual uncertainties is a better approach for long term analysis and maintaining consistency throughout a time series of measurements.

Two other possible contributor to the total calibration constant uncertainty is the overlap function and aerosol layers. RALMO is designed to have no differential overlap in the water vapour and nitrogen channels and an overlap ratio between the nitrogen and water vapour signals of unity. However, a small differential overlap could result from chromatic aberration from the protective windows and edge filters (Dinoev et al., 2013). We also do not consider aerosols in our transmission calculation, but we expect this to make a small contribution to the uncertainty for the nights with typical aerosol loading used in this study. (Whiteman et al., 1992; Whiteman, 2003; Dinoev et al., 2013) and others have shown aerosol effects on lidar signals to be less than 5% during mostly clear conditions, which would translate to a negligible effect on the calibration constant.

7 Summary

We have presented a new method, using GRUAN-corrected radiosondes, to calibrate Raman-scattering water vapour lidar systems that incorporates geophysical variability into the determination of the calibration constant. The trajectory method tracks the air parcels measured by the radiosonde and matches them with the appropriate lidar measurement time; thus, the integration time varies with height. We compared this method to the traditional lidar calibration technique where we sum 30 min of lidar measurements and fit them to a radiosonde profile.

The difference between the traditional and trajectory method calibration coefficients is due to the different lidar profiles used difference in 1-minute lidar scans selected by the methods, as well as the difference in correlation regions used to determine the calibration coefficient from these profiles. This difference means We found that when the water vapour field is homogeneous, the traditional method and trajectory method profiles will produce similar profiles, with slight differences due to the correlation regions included. We found that The homogeneous nights had an average difference of 0.4% from the traditional calibration constant value. In contrast, the heterogeneous nights, or nights with significant structural changes over the 30 min traditional calibration period, had an average difference of 2% with respect to the traditional constant. with the traditional constant typically smaller.

The average fitting statistical uncertainty in the calibration constants produced by the trajectory technique is 0.4%, as opposed to the traditional method error uncertainty of 0.3%. However, although the fitting uncertainty is negligible relative to the uncertainty of the calibration constant due to the uncertainty of the radiosonde measurements , which is 4% on average. which average 8% of the mixing ratio from the ground until 10 km. We have also shown that using trajectories to track the air sampled by the radiosonde more accurately reproduces the radiosonde profile when the water vapour field is variable and decreases the percent difference between the lidar and radiosonde measurements by 5 - 10%. In summary, we found the following:

1. The traditional and trajectory methods agree when the water vapour field is homogeneous during the radiosonde flight. The average difference between their calibration constants (when not considering the single outlier) was $0.43 \pm 0.21\%$.

2. The trajectory method provides a better fit with the radiosonde when the water vapour field changes appreciably over the time of the radiosonde flight. For these cases the calibration constants calculated by the trajectory method resulted in an average of $1.92 \pm 0.93\%$ difference. with the trajectory method typically larger than the traditional.

3. The trajectory method produces a smaller average bias between the radiosonde and the lidar than the traditional method below 4 km (Fig. 8). Adding points above 4 km does not change the calibration constant significantly as the photon counting uncertainty becomes large at these altitudes.

4. The combined lidar statistical and radiosonde mixing ratio uncertainties contribute an average of 4.5% uncertainty in the calibration constant determination for both calibration methods, where the radiosonde mixing ratio uncertainty is the dominating factor.

5. The uncertainty in the calibration coefficient due to the uncertainty in dead time contributes an average of 0.3% in the calibration constant coefficient for a 5% 4% dead time uncertainty in the knowledge of the dead time.

6. The uncertainty in the calibration constant due to the uncertainty in the extinction is less than 0.01%. The uncertainty in the calibration constant due to the Ångstrom exponent's uncertainty is larger and is on average 0.4%, but can reach higher than 1% when strongly-attenuating aerosol layers are present.

7. The trajectory method has an RMS uncertainty of 5% while the traditional method has an RMS of 4% on average over all calibration nights, which is consistent with the total uncertainties calculated using Eq. 8.

A summary of the uncertainty components in the calibration constant is shown below in Table 2.

8 Discussion and Conclusions

The trajectory calibration technique attempts to more realistically represent the physical processes taking place during a radiosonde - lidar calibration, by ensuring the radiosonde and lidar sample the same air mass. This tracking method was built

Parameter	Parameter Uncertainty	Avg Uncertainty in Calibration Constant
Lidar Photon counting	Varies with Altitude 5 - 40%	<0.5%
Sonde Mixing Ratio	Varies with Altitude 0.5-40%	4%
Dead time	5%	0.3%
Extinction	100%	< 0.01%
Ångstrom Exponent	0.34	0.4%

Table 2. Components of the calibration uncertainty, their inherent uncertainty, and their contribution to the uncertainty of the calibration constant.

upon the methods suggested in Whiteman et al. (2006), Leblanc and Mcdermid (2008), Adam et al. (2010), and Herold et al. (2011). Similarly to Whiteman's “Track" technique and Leblanc's “radiosonde-tracking" technique the techniques discussed in these studies, we match the measurements at each altitude with the radiosonde. However, Whiteman et al. (2006) assumed a horizontally homogeneous and uniformly translating atmosphere and did not consider varying wind speed and direction. In Whiteman et al. (2006) the integration time was varied with altitude in order to keep the random uncertainty below 10%, however, the position of the air parcels was not considered. This technique was ultimately found to be not as accurate as other methods, and was later improved upon using the correlation comparisons in Whiteman et al. (2012). Our method does not assume a uniformly translating atmosphere, however, we do consider a homogeneous region around the lidar and the integration time is varied as a function of the time the air parcels spend inside the homogeneous region. Leblanc and Mcdermid (2008) used four methods to match the radiosonde and the lidar measurements: 1) no matching, summing 2 hours of lidar profiles 2) using all lidar scans before the radiosonde reaches 10 km, about 30 min of scans, similar to our "traditional method", 3) only altitudes with minimum water vapour variability over 2 hours are used to calibrate, and 4) only using scans which were coincident with the radiosonde altitude - similar to the Whiteman et al. (2006) “Track” technique and our trajectory method. However, method 4 did not track the air parcels as we did. Leblanc et al. (2012) found that the second method provided the smallest variation in their calibration constant, but did mention that the other methods produced very close results and could be used as well.

Another way to attempt to correct for the movement of the air mass or radiosonde is to follow the methods in Dionisi et al. (2010); Whiteman et al. (2012) which look for regions of high correlation as in these regions it is more likely the radiosonde and lidar are sampling the same air mass. The traditional method, using the correlation algorithm, does provide similar calibration constants to the trajectory method on homogeneous nights. However, using the combined correlation algorithm with the trajectory tracking can provide more regions of high correlation for the calibration, particularly on heterogeneous nights when the air mass changes rapidly as it passes over the lidar. In some cases, of course, it may not provide more regions for calibration particularly if the wind speeds are high and the radiosonde quickly leaves the 3 km homogeneous lidar region.

Using our new trajectory method has several advantages over the traditional technique. The first advantage is that the method presents an automatic and new scheme to calibrate with non-co-located radiosondes. The trajectory method does not rely on the radiosonde's location, but instead relies of the direction of the air measured by the radiosonde. The trajectory method will automatically find the appropriate calibration times as a function of altitude for the lidar. Lidar stations may then be able to use radiosondes launched farther away more effectively, thus allowing more frequent calibrations over the year as well as reducing the need for expensive calibration campaigns. Lidar stations who use our technique with radiosondes located several kilometers away may find it necessary to expand their "lidar region" to greater than 3 km. Secondly, this method allows for calibration if the water vapour field changes rapidly in space and time, allowing more nights to be used for calibration when they would otherwise be discarded due to large differences between the traditional lidar profile and the radiosonde. Lidars with drifts or fluctuations in their calibration constant that may require many calibrations might also find this technique useful. Additionally, frequent and accurate lidar calibrations are critical for detecting water vapour trends and small changes in water vapour in the UTLS region. The tracking method also removes the representative uncertainty which is a large component of the uncertainty budget in the traditional calibration constant calculation. We consider the representation uncertainty to be small in the trajectory method because we are now considering the location of the radiosonde relative to the lidar. Lastly, this technique provides an automatic, objective and quantitative method of determining acceptable calibration nights. This method could conceivably be expanded to work with ozonesondes or tracking other conserved quantities such as aerosols. We have not attempted to expand this technique, but leave it up to others who may find it useful. The method could also be further expanded to work with wind field measurements that include vertical wind speeds.

A significant new aspect of our study is using calibrated GRAUN radiosondes whose analysis includes a complete uncertainty budget. The full uncertainty budget shows the radiosonde measurement is the dominant uncertainty source as compared to the uncertainty in the regression line on an individual night derived from the uncalibrated lidar measurements and sonde. The uncertainty in the lidar measurements, the dead time, and the extinction components contribute an order of magnitude smaller uncertainty than the radiosonde. However, the Ångstrom exponent can contribute on the same order of magnitude uncertainty as the radiosonde if there are strongly-interacting aerosols present during the calibration. Using the GRAUN sondes allows a calibration to be determined with a full uncertainty budget on an individual night, as opposed to requiring a times series of nights to calculate a statistical calibration variation. The uncertainties in our calibration determinations could be reduced using the hybrid method of Leblanc and Mcdermid (2008), as a refinement of our method would be to combine the trajectory calibration with an internal lamp source or some sort of internal calibration technique which could further reduce the variation in the calibration over time.

Eleven of the calibration nights in this study showed significant structural variations in water vapour over the 30 min traditional calibration period. These nights had an average of 2% difference in the calibration constant, which is less than the average calibration uncertainty of 4.5%. Therefore, the trajectory and traditional methods do not produce statistically different calibration values. However, the trajectory method does more accurately reproduce the radiosonde profile than the traditional method below 4 km and above 4 km the methods do equally well. (Fig. 8). The water vapour content below 4 km for the nights in this study was an average of 87% of the total content measured by the radiosonde. Therefore, we believe that calibration should be limited to below 4 km where the signal is highest and the trajectory method performs best. Additionally, the points above 4 km do not make a significant difference in the calibration factor obtained.

The RALMO has an average of 50% uptime over the last 10 years, making it an ideal database for the detection water vapour trends in the free troposphere. In addition to frequent measurements, trend analyses also require minimal uncertainty and well-characterized retrievals. The aim of this work was to develop a calibration method that characterized the uncertainty of the calibration constant as well as making sure it was physically consistent with the reference instrument. The trajectory calibration technique will be used in conjunction with an internal calibration method to produce a 10 year water vapour climatology and UTLS trend analysis using RALMO measurements .

Data availability. All GRUAN data is accessible on www.gruan.org and access may be requested through them. MeteoSwiss lidar data may be requested by contacting Dr. Alexander Haefele (Alexander.Haefele@meteoswiss.ch).

Acknowledgements. We would like to thank the GRUAN support team for providing the GRUAN-processed radiosonde measurements. Shannon Hicks-Jalali would also like to thank Ali Jalali who spent time reading the paper and providing helpful scientific discussions and suggestions throughout the entire process. This project has been funded in part by the National Science and Engineering Research Council of Canada through a Discovery Grant (Sica) and a CREATE award for a Training Program in Arctic Atmospheric Science (K. Strong, PI), and by MeteoSwiss (Switzerland).

References

Adam, M., Demoz, B. B., Whiteman, D. N., Venable, D. D., Joseph, E., Gambacorta, A., Wei, J., Shephard, M. W., Miloshevich, L. M., Barnet, C. D., Herman, R. L., Fitzgibbon, J., and Connell, R.: Water vapor measurements by Howard university Raman lidar during the WAVES 2006 campaign, Journal of Atmospheric and Oceanic Technology, 27, 42–60, https://doi.org/10.1175/2009JTECHA1331.1, 2010.

Avila, G., Fernández, J., Tejeda, G., and Montero, S.: The Raman spectra and cross-sections of H2O, D2O, and HDO in the OH/OD stretching regions, Journal of Molecular Spectroscopy, 228, 38–65, https://doi.org/10.1016/j.jms.2004.06.012, 2004.

Bevington, P. R. and Robinson, D. K.: Data Reduction and Error Analysis for the Physical Sciences, 3rd Edition, McGraw-Hill Companies, Inc., https://doi.org/10.1063/1.4823194, 2003.

Brocard, E., Philipona, R., Haefele, A., Romanens, G., Mueller, A., Ruffieux, D., Simeonov, V., and Calpini, B.: Raman Lidar for Meteorological Observations , RALMO – Part 2 : Validation of water vapor measurements, Atmospheric Measurement Techniques, 6, 1347–1358, https://doi.org/10.5194/amt-6-1347-2013, 2013.

Daidzic, N.: Long and short-range air navigation on spherical Earth, International Journal of Aviation, Aeronautics, and Aerospace, 4, 1–54, https://doi.org/10.15394/ijaaa.2017.1160, 2017.

David, L., Bock, O., Bosser, P., Thom, C., and Pelon, J.: Study and mitigation of calibration factor instabilities in a water vapor Raman lidar, Atmospheric Measurement Techniques, 10, 2745–2758, 2017.

Dinoev, T., Simeonov, V., Arshinov, Y., Bobrovnikov, S., Ristori, P., Calpini, B., Parlange, M., and Van Den Bergh, H.: Raman Lidar for Meteorological Observations , RALMO – Part 1 : Instrument description, Atmospheric Measurement Techniques, 6, 1329–1346, https://doi.org/10.5194/amt-6-1329-2013, 2013.

Dionisi, D., Congeduti, F., Liberti, G. L., and Cardillo, F.: Calibration of a multichannel water vapor Raman lidar through noncollocated operational soundings: Optimization and characterization of accuracy and variability, Journal of Atmospheric and Oceanic Technology, 27, 108–121, https://doi.org/10.1175/2009JTECHA1327.1, 2010.

Dirksen, R. J., Sommer, M., Immler, F. J., Hurst, D. F., Kivi, R., and Vömel, H.: Reference quality upper-air measurements: GRUAN data processing for the Vaisala RS92 radiosonde, Atmospheric Measurement Techniques, 7, 4463–4490, 2014.

Ferrare, R. A., Melfi, S. H., Whiteman, D. N., Evans, K. D., Schmidlin, F. J., and Starr, D. O.: A Comparison of Water Vapor Measurements Made by Raman Lidar and Radiosondes, Journal of Atmospheric and Oceanic Technology, 12, http://adsabs.harvard.edu/abs/1995JAtOT..12.1177F, 1995.

Foth, A., Baars, H., Di Girolamo, P., and Pospichal, B.: Water vapour profiles from raman lidar automatically calibrated by microwave radiometer data during HOPE, Atmospheric Chemistry and Physics, 15, 7753–7763, https://doi.org/10.5194/acp-15-7753-2015, 2015.

Godin-Beekmann, S., Porteneuve, J., and Garnier, A.: Systematic DIAL lidar monitoring of the stratospheric ozone vertical distribution at Observatoire de Haute-Provence (43.92°N, 5.71°E), Journal of Environmental Monitoring, 5, 57–67, https://doi.org/10.1039/b205880d, 2003.

Han, Y., Snider, J. B., Westwater, E. R., Melfi, S. H., and Ferrare, R. A.: Observations of water vapor by ground-based microwave radiometers and Raman lidar, Journal of Geophysical Research, 99, 695–702, 1994.

Herold, C., Althausen, D., Müller, D., Tesche, M., Seifert, P., Engelmann, R., Flamant, C., Bhawar, R., and Di Girolamo, P.: Comparison of Raman Lidar Observations of Water Vapor with COSMO-DE Forecasts during COPS 2007, Weather and Forecasting, 26, 1056–1066, https://doi.org/10.1175/2011waf2222448.1, 2011.

Hogg, D., Guiraud, F., Snider, J., Decker, M., and Westwater, E.: A Steerable Dual-Channel Microwave Radiometer for Measurement of Water Vapor and Liquid in the Troposphere, Journal of Climate and Applied Meteorology, 22, 1983.

Hyland, R. and Wexler, A.: Formulations for the thermodynamic properties of the saturated phases of H2O from 173.15 K to 473.15 K, ASHRAE Tran., 89, 500–519, 1983.

Immler, F. J., Dykema, J., Gardiner, T., Whiteman, D. N., Thorne, P. W., and Vömel, H.: Reference quality upper-air measurements: Guidance for developing GRUAN data products, Atmospheric Measurement Techniques, 3, 1217–1231, https://doi.org/10.5194/amt-3-1217-2010, 2010.

JCGM: Evaluation of measurement data — Guide to the expression of uncertainty in measurement, International Bureau for Weights and Measures (BIPM), 50, 134, https://doi.org/10.1373/clinchem.2003.030528, http://www.bipm.org/en/publications/guides/gum.html, 2008.

Kämpfer, N.: Monitoring Atmospheric Water Vapour, 2013.

Kulla, B. and Ritter, C.: Water Vapor Calibration: Using a Raman Lidar and Radiosoundings to Obtain Highly Resolved Water Vapor Profiles, Remote Sensing, 11, 616, https://doi.org/10.3390/rs11060616, https://www.mdpi.com/2072-4292/11/6/616, 2019.

Leblanc, T. and Mcdermid, I. S.: Accuracy of Raman lidar water vapor calibration and its applicability to long-term measurements, Journal of Applied Optics, 47, 2008.

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, Atmospheric Measurement Techniques, pp. 17–36, https://doi.org/10.5194/amt-5-17-2012, 2012.

Melfi, S. H.: Remote Measurements of the Atmosphere Using Raman Scattering, Journal of Applied Optics, 11, 1605–1610, 1972.

Miloshevich, L. M., Vömel, H., Whiteman, D. N., and Leblanc, T.: Accuracy assessment and correction of Vaisala RS92 radiosonde water vapor measurements, Journal of Geophysical Research, 114, 1–23, https://doi.org/10.1029/2008JD011565, 2009.

Nicolet, M.: On the molecular scattering in the terrestrial atmosphere : An empirical formula for its calculation in the homosphere, Planetary and Space Science, 32, 1467–1468, https://doi.org/10.1016/0032-0633(84)90089-8, 1984.

Penney, C. and Lapp, M.: Raman-scattering cross sections for water vapor, Optical Society of America, 66, 422–425, 1976.

Ross, R. J. and Elliott, W. P.: Tropospheric WAter Vapor Climatology and Trends over North America: 1973-93, Journal of Climate, 9, 3561 – 3574, 1996.

Sica, R. J. and Haefele, A.: Retrieval of water vapor mixing ratio from a multiple channel Raman-scatter lidar using an optimal estimation method, Journal of Applied Optics, 55, 2016.

Sica, R. J., Sargoytchev, S., Argall, P. S., Borra, E. F., Girard, L., Sparrow, C. T., and Flatt, S.: Lidar measurements taken with a large-aperture liquid mirror. 1. Rayleigh-scatter system., Applied optics, 34, 6925–36, https://doi.org/10.1364/AO.34.006925, 1995.

Simeonov, V., Fastig, S., Haefele, A., and Calpini, B.: Instrumental correction of the uneven PMT aging effect on the calibration constant of a water vapor Raman lidar, Proc. of SPIE, 9246, 1–9, https://doi.org/10.1117/12.2066802, 2014.

Smart, W.: Textbook on Spherical Astronomy 6th Ed., 6, Cambridge University Press, Cambridge, 1977.

Trenberth, K. E., Fasullo, J., and Smith, L.: Trends and variability in column-integrated atmospheric water vapor, Climate Dynamics, 24, 741–758, https://doi.org/10.1007/s00382-005-0017-4, 2005.

Turner, D. D., Ferrare, R. A., Heilman Brasseur, L. A., Feltz, W. F., and Tooman, T. P.: Automated retrievals of water vapor and aerosol profiles from an operational raman lidar, Journal of Atmospheric and Oceanic Technology, 19, 37–50, https://doi.org/10.1175/1520-0426(2002)019<0037:AROWVA>2.0.CO;2, 2002.

Venable, D. D., Whiteman, D. N., Calhoun, M. N., Dirisu, A. O., Connell, R. M., and Landulfo, E.: Lamp mapping technique for independent determination of the water vapor mixing ratio calibration factor for a Raman lidar system, Journal of Applied Optics, 50, 2011.

Weatherhead, E. C., Reinsel, G. C., Tiao, G. C., Meng, X.-L., Choi, D., Cheang, W.-K., Keller, T., DeLuisi, J., Wuebbles, D. J., Kerr, J. B., Miller, A. J., Oltmans, S. J., and Frederick, J. E.: Factors affecting the detection of trends: Statistical considerations and applications to environmental data, Journal of Geophysical Research, 103, 17 149, https://doi.org/10.1029/98JD00995, http://doi.wiley.com/10.1029/98JD00995, 1998.

Whiteman, D. N.: Examination of the traditional Raman lidar technique II. Evaluating the ratios for water vapor and aerosols, Applied Optics, 42, 2593, https://doi.org/10.1364/AO.42.002593, 2003.

Whiteman, D. N., Melfi, S. H., and Ferrare, R. A.: Raman lidar system for the measurement of water vapor and aerosols in the Earth's atmosphere., Applied optics, 31, 3068–82, https://doi.org/10.1364/AO.31.003068, 1992.

Whiteman, D. N., Russo, F., Demoz, B., Miloshevich, L. M., Veselovskii, I., Hannon, S., Wang, Z., Vömel, H., Schmidlin, F., Lesht, B., Moore, P. J., Beebe, A. S., Gambacorta, A., and Barnet, C.: Analysis of Raman lidar and radiosonde measurements from the AWEX-G field campaign and its relation to Aqua validation, Journal of Geophysical Research, 111, 1–15, https://doi.org/10.1029/2005JD006429, 2006.

Whiteman, D. N., Venable, D., and Landulfo, E.: Comments on " Accuracy of Raman lidar water vapor calibration and its applicability to long-term measurements ", Journal of Applied Optics, 50, 2170 – 2176, 2011a.

Whiteman, D. N., Vermeesch, K. C., Oman, L. D., and Weatherhead, E. C.: The relative importance of random error and observation frequency in detecting trends in upper tropospheric water vapor, Journal of Geophysical Research Atmospheres, 116, 1–7, https://doi.org/10.1029/2011JD016610, 2011b.

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, Atmospheric Measurement Techniques, 5, 2893–2916, https://doi.org/10.5194/amt-5-2893-2012, 2012.

World Meteorological Organization (WMO): Guide to Meteorological Instruments and Methods of Observation: (CIMO guide), Tech. rep., https://library.wmo.int/doc_num.php?explnum_id=4147, 2014.