This is from Referee #3

Thank you for providing additional information and clarifications. Though most of my concerns were addressed adequately in your response, some issues still remain. I wrote additional comments concerning those issues below. Where relevant, I reproduced my original remarks/questions and your responses.

Original comment: Section 3.3.3: Since you do not precisely explain what the Matlab Neural Code does and how the blue trace is derived, I suggest you remove that part and shorten this section.

Author: I have had several discussions at NDACC lidar meetings and at the last IRLC about using machine learning and MatLab's neural network toolbox to estimate lidar profile backgrounds. I think that is plot will be interesting to several people.

Reviewer: I do not question potential interest within the community in machine learning and using Matlab's toolbox. Machine learning is a complex and much hyped topic, but many issues regarding e.g. reproducibility of the results are not yet well understood, as results often depend on how the particular network was trained and what training data were used.

If you do not provide any information about how you set up the neural network and how it was trained, nobody will ever be able to reproduce your findings or compare with other results. Thus, publishing those results will be worthless for the science community.

You may show results obtained with the Matlab Neural Network tool in your paper. However, you should briefly explain what you did. Just saying "I used this toolbox and that is what I got" is no scientific work. On the other hand, you could say "I took the Fourier transform of time series x and here is the spectrum", as the Fourier transform is a well-defined algorithm which will always produce the same results when applied to the same data sets. With neural networks many things can go wrong because their behavior is not so well-defined. In your case that is especially important given your statement "the software requires an exhaustive set of example bad profiles which we cannot supply". So how was the network trained without a representative training data set? Did you use any trick nobody else knows?

Coming up with a precise description of what the neural network did will likely be difficult. For that reason, and because the neural network part is not essential for your paper's main results, I suggested to remove this part. However, you may insist on showing these results. In that case I will insist on a description of the neural network. That description may be brief, but all essential information which allows someone to repeat your steps must be given.

Discussion of Matlab's neural networking tool are removed. Figure06 is edited.

Line 245: "bad scans" -> "bad profiles"?

Changed

Original comment: Lines 267-269: What is the reason for choosing "the point where the signal to noise equals one in the density profile"?

Author: It seems like a reasonable choice for deciding on an arbitrary starting point. We were motivated by getting temperatures in the UMLT. However, the point is well taken. I'm aware that other groups use other definitions. It would be good to see a study devoted specifically on this topic.

Reviewer: I may sound harsh, but your answer "it seems like a reasonable choice" does not convey any scientific value. To be more precise, why is that a reasonable choice? If you did not investigate this question, it is ok to say so (given the limited time, no study can be perfect). I am aware of different groups using different definitions for the starting point, and I was just curious whether you have any convincing arguments for a particular definition.

We will be more precise in describing the quantitative motivations for our choice of SNR=1. I did some basic testing of different SNR values 3 years ago, and found that the results were a compromise between starting the temperature inversion at an altitude which is too high, where the profile contained more noise than signal (SNR<1; resulted in variability at the top of the temperature profiles which is not geophysical) and starting the inversion at an altitude which is too low (SNR=5 for example; resulted in having to start the temperature inversion a few km lower, and with the Hauchecorne-Chanin retrieval requiring yet further altitude to be cut off of the retrieved temperature profiles in the region before the uncertainty converges to acceptable values, we lost significant information in the altitude range we wish to study). Therefore, SNR=1 was chosen, as it is the least restrictive criterion for batch processing our temperatures which still ensured that the temperature inversions were initialized more by information than by noise.

We have added the following text to line 269:
"We chose a cutoff value of SNR=1 because it is the least strict value we could use which ensures that we have more information than noise (or, specifically, more information than

noise plus background counts), at the altitude within the density profile where we begin the downward temperature integration. Had we chosen a criterion which was less strict (SNR<<1), we would expect to see more statistical variability in the top altitudes of the temperature retrieval as a result of starting the temperature integration in a region which contains more noise than signal. Conversely, choosing a criterion which is too strict (SNR>>1) limits the maximum altitude of the temperature retrieval as discussed in section 3.6.1."

Original comment: Section 3.5.3: You should not attempt to "correct" signal induced noise. It is fundamentally impossible to characterize properly signal induced noise in lidar signals because the noise is superposed on the atmospheric signal. Determining the signal induced noise from the background signal above the lidar signal is bound to fail because you are essentially observing the noise at different times outside the period where you actually are interested in. Signal induced noise is highly non-linear and therefore it is impossible to properly correct it. The data should be regarded as corrupt and not be used in lidar analysis. Besides, significant signal induced noise (e.g. blue trace in Figure 9) indicates that detectors are operated outside safe limits or there is a general technical problem with the lidar. If you insist on using the questionable data, you should assess how the retrieved temperature profile changes when you tweak your model representing the signal induced noise (e.g. cubic versus linear). How do your retrieved profiles compare to independent observations e.g. radiosondes at lower altitudes?

Author: I disagree with the conclusion that we should not make the attempt at a SIN correction. You are quite right that a perfect correction might be impossible. However, we have found that a correction of the sort described in the paper, for the types of signal induced noise that we see at OHP, can be adequately applied for the purposes of our temperature retrievals. The effects of this signal induced noise in our profiles, when uncorrected, is to warm the upper altitude regions of the temperature profiles. Conveniently, we have two measurement channels (the high and low gain channels) which make coincident measurements in this region. Typical count rates within this region are are well within the linear response regime of the high gain channel; therefore dead time correction is not required at these altitudes, and we can believe the high gain channel temperature profile in this region. The quadratic correction for signal induced noise in the low gain channel brings the resulting low gain temperatures into agreement with those from the high gain channel at these high altitudes.
While it would be wonderful to eliminate every stray source of noise in the lidar, we cannot do this for the measurements going back 40 years and more - which form a valuable data set. We also point out that the effect of this quadratically-characterized signal induced noise is negligible at low altitudes: For example, in Fig. 9, the SIN contribution at 30 km is less than 100 counts,

compared to a bg + signal value in the tens of MHz (see fig03). In terms of contribution to temperature, this is so small as to not be observable.
I did some initial quality testing between my 3 channel lidar temperature retrieval and the radiosondes launched from the station at Nimes (~150 km west) and the results are reasonable. There's some expected differences but the results can be very good when the sonde travels directly east. That said the focus of this paper is above 30 km and a full radiosonde comparison study with calculated air mass trajectories would be a good project for the next student.

Reviewer: I agree, we can't change the past and need to work with the data at hand. Above you mention the agreement between high and low channel temperatures, which improves when the quadratic correction is applied. That is very valuable information, because it gives credibility to your approach, and should be mentioned in your manuscript as well. On the other hand, your validation works only for the lower channel. In the absence of any validation for the upper channel which shows a completely different behavior (linear versus cubic), you could at least provide an estimate of the magnitude of the correction, e.g. x K at 75 km, where x is the difference between temperature profiles retrieved with and without correction. If x is sufficiently large (e.g. >1 K), that should be acknowledged as potential source of error, as signal induced noise is a dynamic phenomenon which commonly depends on several factors (e.g. peak intensity, average intensity, particular type of detector) and thus likely varies over a broad range of time scales (from pulse-to-pulse to months). The problem is that signal induced noise causes a non-Gaussian error, so integrating longer does not help you. Your correction most likely helps alleviating the problem, but it won't be perfect. How well it really works – we don't know. E.g. you may unknowingly overcorrect the noise resulting in a cold bias, or undercorrect and still retain a warm bias. Only the comparison with an independent data set can tell whether your correction is working as it is supposed to. However, if your x is small (I think it is. Unless both lidar systems show exactly the same behavior, I would expect larger differences in Fig. 15 for a large x.), you may argue that the effect of signal induced noise on temperature is small as well. Because of the problems it causes, most groups try to avoid signal induced noise by limiting the peak count rate to safe levels.

Added near pg17 line 379 (updated version of article):
"We have some confidence that the quadratic background correction to the low gain channel correctly approximates the moderate non-linear signal induced error because we can compare the corrected low gain channels to the high gain channel. In the overlap region we have two channels making coincident measurements and we can safely assume that the response rate for the high gain channel is linear. Therefore, a correction for signal induced noise in the low gain channel which brings the resulting low gain count rates into the closest agreement with the high gain channel count rates at the same altitudes will be

the optimal choice for the correction. In some cases, the quadratic correction for signal induced noise in the low gain channel yields better agreement than the constant or linear corrections, in which case it is employed. The best individual choice (constant, linear, quadratic) is used for each profile. We believe these empirical corrections to be sufficient, because (a) the resulting agreement with the high gain channel improves as compared to the uncorrected profile, and (b) the resulting corrected low gain count profiles are generally equal to the high gain count profiles to within statistical uncertainty, and (c) for the few cases in which the empirical correction ultimately fails, this will be apparent by the corrected signal retaining poor SNR values. The melding procedures of Section 3.6 weight the combined high and low gain Rayleigh channels according to SNR, and so in these cases, the poorly-corrected low gain contributions to the final melded counts profile will be negligible, and all information will be obtained from the high gain channel."

For the high gain channel, there is no extra channel to compare to. Instead, we consider the shape of a standard idealized lidar profile which contains no signal induced noise: That which has only a constant background with no slope nor any curvature. We are confident that the majority of the high gain channel profiles exhibit this behaviour - the constant background is in fact the most commonly fit background for this channel. For the linear and quadratic backgrounds which are sometimes exhibited by this channel, we reason that any correction which can bring the corrected background into closer agreement with the shape of an idealized background (i.e. a constant about zero) will be optimal. The profiles for which we have applied a linear or quadratic background removal do indeed result in such corrected profiles and we deem this to be sufficient for the purposes of the current study.

The ultimate effects of a quadratic background correction on the retrieved temperature profile, as compared to a constant background correction for the same profile, are possible to calculate. However, it is not immediately apparent how to properly characterize the effect of the quadratic non-linear SIN correction as a function of altitude for profiles in general. The magnitude of the change in temperature depends both on the quality of the nightly integrated photon counts profile (a clear night with high signal will receive a correction at a different altitude than cloudy night with lower laser power), slope of the linear correction or the curvature of the quadratic correction, and the underlying 'true' photon counts gradient. Plus there are confounding variables, correcting the background changes the a priori initialisation altitude, the value of the a priori, the convergence of the temperature algorithm, and the choice of vertical integration scale. The effect of a quadratic background correction on the temperature.

You're absolutely correct my correction is far from perfect. I've done my best to correctly identify and subtract the noise and background in the lidar profiles. In Figure 11 where I compare the results of the changes made in the algorithm to the satellite temperatures we see that the corrections work together to minimize the temperature differences. It may be interesting in future to delve into the relative temperature contributions of each of the corrections identified in this paper. For the current project, we ensured that each correction was an improvement within the arena in which we were testing it (i.e. removal of TES, we checked that TES were indeed removed from profiles. For removing background, we check that the resulting counts profile did not have any remaining structure at the background altitudes). We did not run the full temperature analysis with every combination of corrections turned on and off - although this is a parameter space we could certainly explore at some later date.

The peak count rate is well within safe limits. We cover the full dynamic range of photon returns by sharing the altitude range between two Rayleigh channels, each designed specifically to each cover only a part of the dynamic range. This keeps count rates on each PMT within the linear regime. As shown in Figure 3, there is no evidence of large nonlinear response in the raw counts profiles. If there was, we would expect to see a flattening of the profiles at the lower end of each altitude range, in addition to the flattening induced by the intentional electronic blanking of the PMTs. We do not see such an effect. Further, nonlinearity tests were carried out during the critical review of French lidars (REF: Kekhut 1993; section 5b), which determined the linear regime for each PMT, and the regime which is correctible via dead time correction (to within a 5% uncertainty). The count rates shown in Figure 3 of our paper are well within the linear regime of the counting hardware.

Original comment: Figure 13: It is hard to estimate absolute temperature differences. I suggest you use a segmented color bar with 6-10 different colors. Can you provide a plot showing combined temperature error estimates of both lidar data sets? There is a period in mid 2001 with distinct blue color (negative temperature differences) between 30 and 55 km altitude. Could these observations also have been affected by misalignment? A similar area can be found in right after the last marked region in 2011.

Author: The same information is already presented in a more compact way in Fig14 I've added the following text: 'For reference, a typical LTA temperature profile with an effective vertical resolution of 2 km has an uncertainty due to statistical error of 0.2 K at 40 km; 0.4 K at 50 km; 0.6 K at 60 km; 0.7 K at 70 km; 1.8 K at 80 km; and 602 K at 90 km. For reference, a typical LiO3S temperature profile with an effective vertical resolution of 2 km has an uncertainty due to

statistical error of 0.3 K at 40 km; 0.5 K at 50 km; 1.0 K at 60 km; 2.7 K at 70 km; and 10 K at 80 km.' I cannot account for the blue regions in Fig13 based on either lidar uncertainty budget or through geophysical explanations. Yes you're correct the blue bias between 30-50 km is likely due to misalignment. Given 5 mirrors in LTA and 4 mirrors in LiO3S there are many possible ways to be misaligned. As well the severity of the misalignment.

Reviewer: If blue (or red) biases outside the boxes may be caused by misalignment, then misalignment is obviously a major source of error which ultimately limits the accuracy (and, depending on time scale, also precision) of your measurements.

We agree. This is a limitation of our lidar system. Prior to this work, the effects of misalignment had not been identified in the long-term OHP data set. Therefore, identifying that this is a contributing factor to the OHP temperature analyses (as shown here), and making a first attempt to identify these regions programmatically in the data, is a first step to mitigating this factor. In future we will follow up with more sophisticated tests to address this important issue.

We have added the following text to acknowledge this in the manuscript, at line 450:
"It is possible that the criteria described above for identifying periods of misalignment is not yet stringent enough. Therefore, one limitation of the OHP measurements in terms of accuracy, and depending on time scale, also precision, is the influence of periods of misalignment that have not been programmatically identified. An ideal solution would be to have an independent method of monitoring mirror alignment during atmospheric measurements (e.g. installation of a small sighting telescope to measure the alignment coupled with an automatic fiber optic alignment system). With the existing data set from OHP extending back two decades, we unfortunately cannot retrospectively address such a hardware goal, but there may be opportunities in future to look into the effects of choosing different criteria to identify periods of misalignment."

This is from Referee #2

This review is on the revised manuscript of the Wing et al. paper. The authors addressed the comments of the previous review in the extensive (appreciated!) reply and in the new manuscript. The paper is improved and much better comprehensible, now. Nevertheless, there are still some remarks, either to be repeated or new topics coming up with the new version.

The general challenge of the paper is the combination of different aspects (or goals of the paper) and their proper description in the text, that I see better now after reading the review reply. I find three goals: i) the introduction of the long-term data set and the instrumental changes, iii) treatment of this heterogeneous data set – or quality assurance - for the use in the accompanying paper, iii) improvement of the temperature algorithm and reduction of the bias compared to satellite soundings. Of course, these goals cannot be completely separated from each other, but they often affect different altitude sections. I recommend to make these goals clearer, try to subdivide the text appropriately or refer each (sub)section to its particular goals. As an example, the instrumental achievements for background reduction (Sec. 3.4) are dedicated to the first goal, but the deadtime correction (3.5.1) to the second.

The following text has been added to
This work follows three main goals: i) the introduction of the long-term data set and the instrumental changes, ii) treatment of this heterogeneous data set for the use in the accompanying paper, and iii) improvement of the temperature algorithm and reduction of the bias compared to satellite soundings. These goals cannot be completely separated from each other, but goal i) is broadly addressed in sections 2.1-3.2 and 3.4, goal ii) is addressed in sections 3.3.2-3.4 and again in sections 3.5-4, goal iii) is addressed in section 5.

Line 45 – end: Please refer "first part" (second, third) to the Section numbers used in the manuscript.

Changed.

---

l. 72: Please explain shortly how your Mie channel is working if it uses a similar filter as the Rayleigh channels.

We have specified on line 72 that the Mie channel is 532 nm.

The Mie aerosol channel and Raman water vapour channel are not central to this work. A full discussion of OHP aerosol lidar can be seen in the following reference, which we have included at the end of section 2.1:

Khaykin et al., *Variability and evolution of the midlatitude stratospheric aerosol budget from 22 years of ground-based lidar and satellite observations*, Atmospheric Chemistry and Physics, vol. 17, no. 3, pg. 1829--1845, 2017. 10.5194/acp-17-1829-2017

Table 1: Please check the dimension of the mirror diameters.

'mm' changed to 'cm'

Section 3.2: I got confused (and the general reader may also) about the altitude resolutions, integration times, and potential smoothing. I find raw data with 75 m resolution, temperate data with 300 m and 2 km, nightly means and individual profiles. I recommend showing raw data only at the resolution used for the quality control, and temperatures as nightly means with the altitude resolution used for the accompanying paper.

We have presented plots at the relevant resolution for each step of the data processing procedures. Section 4 (re-written) clearly summarizes the resolutions used for each stage. The quality control steps are not all done at the same resolution.

Altitude resolutions used:

- **Raw measurements are made 75 m. Figure 3 is plotted at 75 m.**
- **Spike and LTS corrections must be done at the native resolution, of 75 m. Figures 4 and 5 are plotted at 75 m.**
- **Further quality control can be done at the processing resolution of 300 m. Therefore, figures 6, 7, and 9 are plotted at 300 m resolution.**
- **The temperature profile is calculated at 300 m resolution. Figure 10 is plotted at 300 m.**
- **Next, for comparison with LiO3S and satellite measurements, the LTA NDACC temperature (black) results are integrated to a effective resolution of 2 km, while the LTA new algorithm (green) results remain at 300 m. Figure 11 is plotted at these resolutions.**
- **LTA new algorithm temperatures are then integrated to 2 km effective resolution for quantitative comparison with LiO3S values, which are at 2 km resolution. Figure 13 is plotted at 2 km resolution.**

We have made a more detailed caption for Fig 10, as it is a nightly mean profile. We have now specified that clearly.

The integration times at OHP are set by hand each night. The times indicated in the plots are correct, although not necessarily constant from night to night.

---

l. 184: I suggest writing "overestimation of the background due to localized signal contaminations" (noise should not be confused with background)

Changed.

---

l. 185: "warmer temperatures" should read "higher temperatures"

Changed.

---

l. 185-187: I suggest removing these sentences, because signal contaminations will normally not result in underestimation of the true background (detector noise, moonlight etc.). The sentences may be moved to Section 3.5.3.

Removed:
The opposite holds true for an underestimation of the background (produces a colder profile).

---

l. 240 to …: Please accept that the reader may not be able to identify three or more groups, but only "high background at low signal" and vice versa. Please explain in more detail or (preferred!) just state after introducing the red lines in Fig. 6 that this case may be simple but you are seeking for a flexible solution for 20 years of data (instrumental changes) and various conditions. Furthermore, express that you are searching for the outliers within a single night, not bad-signal profiles in general.

We included the text in lines 243-244 to address exactly the first issue that you point out: It's quite likely that a given reader may not identify three or more groups, although other readers may do so. We have pointed out the particular profile number ranges for the groups we ourselves identify so that the reader can follow along with one example of a reasonable grouping, but there's no reason that someone else would choose the same groups. It is therefore preferable to have an objective, rather than a subjective, test to

determine outliers within a single night. The point you make is exactly one motivation for this test.

Our use of the term "programmatic solution" on line 245, and all text on lines 249-253 was intended to address the second comment: That we are seeking a solution which can be applied to all sorts of measurements we may acquire in various conditions.

To address the third comment, the first paragraph of section 3.3.3 has been rewritten to read:
"After the removal of lidar profiles which suffer from clear signal contamination, there may still be profiles which ought not be included in a lidar temperature analysis because they are outliers of poor quality compared to other profiles within the same night. Conceptually, 'bad profiles' are lidar profiles with a high background and/or a low signal strength as compared to profiles measured shortly before or after the profile in question. These profiles need to be positively identified as not belonging to the general population of nightly lidar profiles".

---

l. 249-254: I still recommend to delete this section and the blue line. This is not a project report. I doubt that you want to state that the Matlab software cannot be used in general. Therefore this section may at most be interesting for the NDACC people you talked to, but the general reader will be confused.

Done.

---

l. 260-267: I am sorry but I still do not understand how the MWW rank-sum test is applied here. As far as I understand, the test checks which of two distributions is larger. Do you simply want to check whether the background is larger than the signal at 35-45 km? Then you may not need the cumulative sum. How the rank sum is calculated and how does this compare to the cumulative sum of either background or signal? Why the first 13 profiles are discarded? From my point of view this test is a central method of the quality control procedures and therefore should be clearly described.

The first null hypothesis here is that all the lidar profiles belong to a single continuous distribution with a single nightly median. On a clear night with constant laser power and a relatively constant background we can easily fail to reject the null and we have no real need for a non-parametric statistic. However, in our example shown in Figure 6 and 7 we have too much signal and background variation in our example for this to be the case.

Failing this criterion we employ the MWW rank sum test which allows the identification of outliers based on N sub-population medians with non-gaussian distribution. A regular T-test will not work here as the assumptions we need to make about the populations are not true. For example, anomalies in noise are most likely to be high while anomalies in signal are likely to be low but anomalies in SNR can go either way. As another example, consider the SNR responses for transient thin cirrus clouds vs. moonlight filtering through transient optically thick clouds.

By using the cumulative sum we can more easily identity significant changes in a signal. Look and figure 6 and figure 7 and mentally try to look for subpopulations. It's much easier to conceptualize the ordinally ranked and cumulatively summed data in figure 7 than to guess at divisions in figure 6. Cumulative sums are the natural choice of variable to use in a MWW test.

Here's a back of the envelope calculation just looking at Fig 7. Unfortunately, I don't save the intermediate statistical scores - I'm only interested in the end determination and whether I should include a profile or not.

Step 1: Do a cumulative sum for the signal and the noise as seen in Figure 7. Take SNR and arrange largest to smallest This allows us to establish an ordinal rank for the data. Given that all lidar data is positive (no negative photon counts) this rank usually but not always corresponds to profile number. Exceptional cases of poor SNR will be assigned a low ordinal rank and the best SNR will have high ordinal rank.

Step 2: In each of the 4 sections of Figure 6 there are some higher and some lower SNR values. I want to see whether there are significant differences in the sub-population medians. The expectation value for the rank each sub-median is U_all = (number of observations in sub-median)*(total number of profiles +1)/2
Example Profiles 1-13: U_1-13 = 13*(92+1)2 = 604.5

Step 3: Do a Z distribution score Z =((sum of ranks) - (mean of population))/(E_w)
Where E_w is the standard error of the Wilcox sum. E_w = sqrt((number of observations in sub-median)*(total number of profiles -number of observations in sub-median)*(total number of profiles+1)/12)
In our Example E_w = sqrt(13*(92-13)*(92+1)/12) = 89
Sum of ranks is somme of 1 to 13 = 91

Step 4: Z =(91-604.5)/89 = -5.8

Step 5: Using a Z-test lookup table for a lower tailed test p = 0.05 has a value of -2.576. So we reject the null hypothesis that the first 13 ordered profiles belong to the population of lidar nightly lidar profiles which are distributed about the nightly median.

---

l. 263: As mentioned above, please clearly distinguish between "background" (the average count rate above the usable range, i.e. at 120-153 km) and "noise" (the statistical uncertainty of the count rate at a given height). The background count rate has its noise, but also the signal has. Colloquial both terms are often confused but must not in a publication.

Changed line 263:
"Consider that a poor quality lidar profile which has a signal to noise ratio of 1 at 70 km contributes more information from the signal than from the (background + noise) at 60 km, but information from the (background + noise) than from the signal at 80 km."

Changed line 267:
The background (noise) of an individual profile, N_{i}, is expressed as the summation of photon counts in bins which fall between 120 km and 155 km and the nightly background, N_{sum} is the summation of all N_{i} for the night.

Section 3.4: You should make clear that this Section is of different "character" than the other ones. Here you describe some (important) instrumental changes that have been done in the past, not your recent software developments. The removal of SIN is described in Section 3.5.3, i.e. either discard this paragraph in 3.4 or add a reference.

We take your point that this section is a little different than the surrounding sections. Certainly the hardware changes are not part of the improved algorithm. However, we feel this section fits best here, since it is preceded by a discussion about SNR, and 3.4 explains how we've done our best over 20 years to reduce the noise. The noise discussed here is total background and is not SIN, so it requires its own description here.

---

l. 346: It should read "X to Y km", but the text in brackets can also be removed. The remaining text is clear enough.

Changed.

---

l. 423: Please explain (in Section 2?) how you combine the N2 Raman channel with the Rayleigh channel. How is the aerosol correction done? Is the Raman channel simply treated as the molecular channel below 30 km?

The focus of this article is above 30 km.
The Raman channel was treated just like a molecular channel. I used a simple model aerosol profile to correct for transmission at both 532 nm and 607 nm. The main benefit of including the N2 Raman channel in this work is to ensure that the count rates in the low gain Rayleigh channel are linear.

Fig. 10: The paper mainly deals with nightly mean profiles; therefore I recommend to show a mean profile here. Furthermore you should indicate the transitions between the channels. I generally doubt that an error of 30 % is useful, even for statistical analyses. I recommend using much smaller error margins to avoid unrealistic temperature gradients.

Figure 10 caption changed to include the word "nightly average".

The transition altitude ranges vary from night to night based on the relative signal quality in each lidar channel and the vertical integration. It is a gradual transition, produced by an average weighted by uncertainty in each contributing channel at each altitude. (Alpers et al., 2004) provides a good description of the technique.

Readers of the paper will be interested to see the effects of the new algorithm at the higher resolutions presented here, even if the uncertainties are large. It is difficult to find any other long-term dataset for temperatures from 5 to 80 km at 300 m vertical resolution. The presentation here allows readers to see the limits of this method - warts and all. This is not the resolution that we would conduct our geophysics at. Rather it is for the purposes of algorithm testing. The measurements can, of course, be integrated to lower altitude resolution, reducing the uncertainty significantly and smoothing the temperature gradients to more realistic values.

---

l. 452: Please explain in some more detail the differences between the standard NDACC retrieval and yours. I assume that also NDACC has some quality assurance measures like removal of SIN contamination, removal of TES, or deadtime correction.

We have added to the beginning of the re-worked section 4 (see full Section 4 in response to a later comment):

"The NDACC algorithm contains such corrections as deadtime, background, and transmission. The new algorithm improves upon the background correction and identification of bad profiles, and introduces corrections for: signal spikes, TES, identification of good profiles, and noise reduction, all which have not previously been addressed by the NDACC algorithm."

Fig. 11: This Figure somewhat demonstrates the confusion about the goals of the paper. The improved retrieval is effective mainly above 75 km. The validation by the co-located lidar is made below 75 km. The companion paper uses data up to 80 km. I do not criticize this figure, but would like to see some clarifications throughout the manuscript, which particular topic is addressed. This would help the reader finding the context of this long paper.

The re-wording and clarifications in the new Section 4 (see response to other comment below) should help with this matter. Likewise, the description of the goals in the introduction.

Up to 70 km, and even 75 km, the improved LTA algorithm has no effect on retrieved temperatures. There, the LTA and LiO3S results match to within their uncertainties even when both lidars use the NDACC algorithm. So below 75 km, the lidars already agree.

Above 75 km, the changes as a result of the new LTA algorithm are noticeable. From 70 to 80 km, the LiO3S measurements are still present. Therefore, there is a small region of 5 km in which the LTA NDACC result and the LTA new algorithm result can each be compared to the LiO3S NDACC result. The effects of the new LTA algorithm are to make the LTA and LiO3S come into much closer agreement. Therefore we see this as a success of the new algorithm.
Of course, we would prefer to have a co-located lidar which routinely produces temperatures to higher altitudes for comparison.

Considering that Fig 11 shows improvements in the LTA retrieval with respect to satellite and model comparisons which account for almost half of the discrepancies seen between LTA NDACC retrievals and the satellites and model, including at altitudes up to 90 km, we are encouraged that the new LTA algorithm is an improvement at higher altitudes as well.

Figure 11 (second topic): Is the variance of the SABER data relevant? I assume it shows mainly geophysical variation of the temperatures in the data set. Median errors of all profiles are more

important. How many profiles contribute to the ensemble, what are the criteria for temporal and spatial matching?

Yes the variance shows the spread of the data due to both errors and geophysical variability.

Figure 11 has been updated with all ensemble median errors.

N = 212 lidar nights

We have updated the text in Section 4 (see full updated text in response to later comment) to say:
"This possibility is explored further in the companion paper, and all coincidence criteria for the satellite comparisons are available therein (Wing et al., 2018b)."

---

l. 480: It is confusing that you stress the advantages of your retrieval compared to NDACC in the first part of this paper, but then use the NDACC code for comparison between the lidars.

We have addressed this confusion by re-writing section 4.

---

l. 514: Before, a resolution of 300 m has been mentioned.
We have re-written Section 4 to clarify the difference between the vertical resolution of the raw lidar profiles, the algorithm development conducted at 300 m resolution for LTA and the integration of the resulting temperature profiles to the lower 2 km effective resolution for direct comparison with results calculated using the NDACC algorithm.

Section 4 now reads:

"The NDACC algorithm contains such corrections as deadtime, background, and transmission. The new algorithm improves upon the background correction and identification of bad profiles, and introduces corrections for: signal spikes, TES, identification of good profiles, and noise reduction, all which have not previously been addressed by the NDACC algorithm.

The LTA data is collected at 75 m resolution. The spike and TES corrections described in Section 3.3.1 and 3.3.2 are carried out at this resolution. Then the profiles are integrated to 300 m, at which point the remainder of the corrections in Section 3 are applied.

Temperature profiles using the new algorithm are calculated at 300 m resolution for LTA, and are plotted as the green line in Fig 11. This is higher resolution than the standard NDACC temperature resolution, which is 1 km, smoothed to 2 km effective vertical resolution. The LTA NDACC-calculated temperatures (black line in Fig 11) are plotted at 2 km effective resolution. By implementing the new algorithm, we have cooled the UMLT lidar temperature retrievals with respect to the standard NDACC temperature algorithm. The modifications cool the mesospheric retrievals by approximately 5 K near 85 km and 20 K by 90 km. There is no significant difference between the new and the NDACC algorithms for LTA below 70 km.

Temperature profiles calculated for LiO3S are all carried out using the NDACC algorithm at an effective vertical resolution of 2 km, and these are show as the orange line in Fig 11. Whereas the LTA NDACC algorithm results are warmer than the LiO3S NDACC algorithm results above 70 km, we now see that the LTA new algorithm results are cooled sufficiently that they more closely match the LiO3S measurements up to 78 km. Therefore the corrections for LTA proposed in the new algorithm represent a significant improvement over the LTA NDACC algorithm for altitudes above 70 km.

A comparison with temperature retrievals from the satellites MLS (red line in Fig 11) and SABER (blue with median error and shaded ensemble variance), and with the MSIS-90 model (magenta line in Fig 11), also show an improvement in the LTA temperatures retrieved using the new algorithm as compared to the LTA NDACC algorithm. By implementing the techniques described in the sections above we can account for nearly half of the temperature difference between the lidar and the satellites at 90 km. The character change in the difference functions above and below 84 km is in part due to the increasing contributions of the species specific Rayleigh backscattering correction and the corrections to the gravity vector. The remaining temperature difference between the improved lidar temperatures (green) and the satellites and model may be in part due to distortions in the satellite a priori for the geopotential vector. This possibility is explored further in the companion paper, and all coincidence criteria for the satellite comparisons are available therein (Wing et al., 2018b).

It is important to note that additional complications exist when comparing temperatures derived from ground based lidars to temperatures derived from satellite data which have their own calibration concerns. We explore the issues of lidar-satellite comparison in Part B of this paper. A co-located ground-based resonance Doppler or Boltzmann lidar would provide a better comparison data set as resonance lidars have high signal to noise ratios above 75 km (Alpers et al., 2004).”

---

l. 516: Do you really mean 602 K?

Corrected to 6 K

---

l. 543: Here, 300 m resolution is used again. Please explain or unify.

We have unified to 300 m, and have further explained.

Modfied text:
For reference, a typical LTA temperature profile with an effective vertical resolution of 300 m (e.g. as shown in Fig 10; the resolution at which we apply the new correction algorithms, which has a detectable effect above 70 km) has an uncertainty due to statistical error of < 1 K below 50 km; 4 K at 60 km; 10 K at 70 km; 50 K at 80 km. LTA temperature profiles are then integrated to 1 km and then smoothed to an effective vertical resolution of 2 km for comparison with the lower-resolution LiO3S, and for submission to the NDACC database. The same typical LTA temperature profile with an effective vertical resolution of 2 km has an uncertainty due to statistical error of 0.2 K at 40 km; 0.4 K at 50 km; 0.6 K at 60 km; 0.7 K at 70 km; 1.8 K at 80 km; and 6 K at 90 km. A typical LiO3S temperature profile cannot be integrated to the required heights at 300 m resolution due to low signal rates. It is integrated to 1 km, and smoothed to an effective vertical resolution of 2 km, resulting in typical uncertainties due to statistical error of 0.3 K at 40 km; 0.5 K at 50 km; 1.0 K at 60 km; 2.7 K at 70 km; and 10 K at 80 km.

There are a lot of typos and odd formulations throughout the manuscript. I recommend consulting advice.

We have had the paper proofread again following the updates.

Manuscript prepared for Atmos. Meas. Tech.
with version 2014/09/16 7.15 Copernicus papers of the LaTeX class copernicus.cls.
Date: 24 September 2018

Lidar temperature series in the middle atmosphere as a reference data set. Part A: Improved retrievals and a ~~20 year cross-validation~~ 20-year cross validation of two co-located French lidars

Robin Wing[1], Alain Hauchecorne[1], Philippe Keckhut[1], Sophie Godin-Beekmann[1], Sergey Khaykin[1], Emily M. McCullough[2], Jean-François Mariscal[1], and Éric d'Almeida[1]

[1]LATMOS/IPSL, UVSQ Université Paris-Saclay, Sorbonne Université, CNRS, Guyancourt, France
[2]Department of Physics and Atmospheric Science, Dalhousie University, Halifax, Canada

Correspondence to: Robin Wing (robin.wing@latmos.ipsl.fr)

Abstract. The objective of this paper and its companion (Wing et al., 2018b) is to show that ground based lidar temperatures are a stable, accurate and precise dataset for use in validating satellite temperatures at high vertical resolution. ~~Long-term~~ Long term lidar observations of the middle atmosphere have been conducted at the Observatoire de Haute-Provence (OHP), located in southern France (43.93° N, 5.71° E), since 1978. Making use of 20 years of high-quality co-located lidar measurements we have shown that lidar temperatures calculated using the Rayleigh technique at 532 nm are statistically identical to lidar temperatures calculated from the non-absorbing 355 nm channel of a Differential Absorption Lidar (DIAL) system. This result is of interest to members of the Network for the Detection of Atmospheric Composition Change (NDACC) ozone lidar community seeking to produce validated temperature products. Additionally, we have addressed previously published concerns of lidar-satellite relative warm bias in comparisons of Upper Mesospheric and Lower Thermospheric (UMLT) temperature profiles. We detail a data treatment algorithm which minimizes known errors due to data selection procedures, a priori choices, and initialization parameters inherent in the lidar retrieval. Our algorithm results in a median cooling of the lidar calculated absolute temperature profile by 20 K at 90 km altitude with respect to the standard OHP NDACC lidar temperature algorithm. The confidence engendered by the ~~long-term~~ long term cross-validation of two independent lidars and the improved lidar temperature dataset is exploited in (Wing et al., 2018b) for use in multi-year satellite validations.

1 Introduction

Rayleigh lidar remote sounding of atmospheric density is an important tool for obtaining accurate, high resolution measurements of the atmosphere in regions which are notoriously difficult to measure routinely or precisely. A key strength of this technique is the ability to retrieve an absolute temperature profile from a measured relative density profile with high spatio-temporal resolution, accuracy and precision. This kind of measurement is exactly what is required to detect longterm middle atmospheric temperature trends associated with global climate change ~~is a~~ and is of great value for routine satellite and model validation (Keckhut et al., 2004).

Comparisons of middle atmospheric temperatures measured from satellites to those measured from lidars have all noted a relative warm bias in lidar temperatures above 70 km. Several recent examples of lidar-satellite relative warm bias in the upper mesosphere can be found in the work of: (Kumar et al., 2003) ~~([~~5-10 K relative to HALOE~~); (Sivakumar et al., 2011) (~~]; (Sivakumar et al., 2011) [5-10 K relative to HALOE, 6-10 K relative to COSMIC/CHAMP, 10-16 K relative to SABER~~);~~ ~~(Yue et al., 2014) (~~]; (Yue et al., 2014) [13 K at 75 km relative to SABER~~); (García-Comas et al., 2014) (~~]; (García-Comas et al., 2014) [3-4 K at 60 km relative to SABER and MIPAS~~); (Yue et al., 2014) (~~]; (Yue et al., 2014) [13 K at 75 km relative to SABER~~); (Dou et al., 2009) (~~]; (Dou et al., 2009) [4 K at 60 km relative to SABER~~); (Remsberg et al., 2008) (~~]; (Remsberg et al., 2008) [5-10 K at 80 km relative to SABER~~)~~]; and (Taori et al., 2012; Taori et al., 2012) ~~([~~25 K near 90 km relative to SABER~~)~~]. The bias is generally attributed to lidar 'initialization uncertainty' and model a priori contributions to the temperature retrieval but, no systematic attempts are made to fully establish this conclusion. These authors also explore the possible influences of tides, lidar-satellite co-incidence criteria, satellite vertical averaging kernels, and satellite temperature accuracy as possible contributing factors.

The work of this paper is to evaluate the suitability of lidars as a reference dataset and to address the problem of systematic errors due to initialization of the lidar algorithm. The subsequent comparison of the improved lidar temperatures to satellite measurements is conducted in the companion paper (Wing et al., 2018b).

~~The first part of this paper~~ This work follows three main goals: i) the introduction of the long term data set and the instrumental changes, ii) treatment of this heterogeneous data set for use in the accompanying paper, and iii) improvement of the temperature algorithm and reduction of the warm bias compared to satellite soundings. These goals cannot be completely separated from each other, but goal i) is broadly addressed in sections 2.1 to 3.2 and 3.4; goal ii) is addressed in sections 3.3 to 3.4 and again in sections 3.5 - 4, and goal iii) is addressed in section 5.

Section 2 of this paper describes the current experimental setup, the specifications of two OHP lidars, and the measurement cadence of two key NDACC (Network for the Detection of Atmospheric Composition Change) lidar systems.

The second part Section 3 of this paper outlines techniques to minimize the magnitude of the aforementioned lidar-satellite temperature bias by systematically detailing a rigorous procedure for the treatment and selection of raw lidar data and will propose improvements to the standard NDACC lidar temperature algorithm for the UMLT (Upper Mesosphere and Lower Thermospshere) region.

The third part Section 4 of this paper gives the net results of the temperature modifications and system improvements in the LTA lidar at OHP.

Section 5 of this paper compares the lidar temperatures produced by an NDACC certified temperature lidar at 532 nm with temperatures produced by the non-absorbing 355 nm line of a co-located NDACC certified ozone DIAL (DIfferential Absorption Lidar) system. This comparison is conducted using a large database of two co-located lidar systems with the goal of providing confidence in the longterm stability of the lidar technique at both wavelengths. There are currently 10 certified temperature lidars, 6 of which are current in their data submission and have temperature profiles freely accessible online. Similarly, there are 12 certified stratospheric ozone DIAL systems of which 5 systems are current with data submission and are available through the NDACC website. We hope that this work will encourage sites with outstanding data obligations to submit their measurements and for DIAL ozone sites to seek validation for their temperature data products for inclusion in the NDACC database (nda). As an ancillary goal we will show that temperatures produced by the Rayleigh lidar technique are accurate, precise and stable over multiple decades and as such are the ideal type of measurement for use in future ground based validation of satellite temperatures. The result of this demonstration will be used in the companion paper (Wing et al., 2018b) as justification for validating satellite data with lidar temperatures.

2 Instrumentation Description

2.1 Rayleigh Lidar

The OHP Rayleigh-Mie-Raman lidar, LTA (Lidar Température et Aérosols), uses a seeded Nd:YAG to produce a 532 nm laser source with a maximum power of 24 W. The transmitted beam is passed through a 13X beam expander and has a 30 Hz repetition rate, a 7 ns pulse width, and a beam divergence of less than 0.1 mrad.

The receiver assembly consists of a high and low gain elastic channel for 532 nm, a Mie scatter channel at 532 nm for aerosols, a Raman channel at 607 nm for molecular nitrogen, and a Raman channel at 660 nm for water vapour. A schematic of the telescope array is shown in Fig. 1. The high gain Rayleigh channel consists of four telescopes. At the focal point of each telescope is an actuator-mounted 400 μm diameter fibre optic. The four fibre optics are bundled to project a single signal onto a Hamamatsu R9880U-110 photomultiplier. The low gain Rayleigh, nitrogen Raman, water vapour Raman and Mie channels all use a single telescope setup and actuator mounted fibre optic.

The two Raman channels rely on the largest telescope and the signals are separated by a dichroic mirror. Specifications for each telescope are found in Table 1.

LTA	Mirror Diameter (cm)	Focal Length (mm)	Field of View (mrad)	Parallax (mm)	Optical Filter Width (nm)	Filter Maximum Transmission (%)
High Gain Rayleigh	4X 50	1500	0.27	800	0.3	84
Low Gain Rayleigh	20	600-800	1.7	257	0.3	84
Nitrogen Raman	80	2400	0.6	600	1	~ 50
Water Raman	80	2400	0.6	600	1	~ 50
Aerosol Mie	20	600-800	1.7	257	0.3	84

Table 1: Specifications for the LTA receiver assembly.

All channels are sampled using a Licel digital transient recorder with a record time of 0.1 μs which corresponds to a vertical resolution of 15 m. The high and low gain Raleigh channels are electronically gated at 22 km and 12 km, respectively, to avoid damaging the photomultipliers with large signal returns. Further details can be found in (Keckhut et al., 1993) and (Khaykin et al., 2017) .

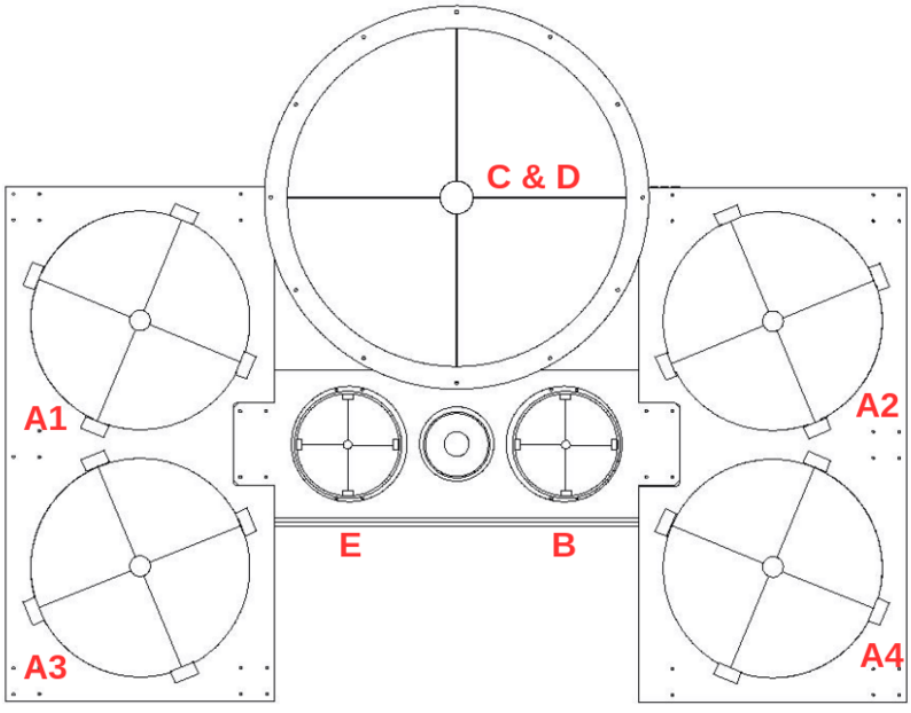

Figure 1: Mirrors A1, A2, A3, A4 (50 cm) are combined for the high gain Rayleigh channel. B (20 cm) is low gain Rayleigh channel. Mirror C&D (80 cm) is the Raman channel for water vapour and molecular nitrogen. E (20 cm) is the Mie channel. The beam expander for the transmitted laser source is between mirrors E and B.

2.2 DIAL Ozone System (LiO_3S)

The OHP Differential Absorption Lidar (DIAL), also referred to as Lidar Ozone Stratosphère (LiO_3S), uses two lasers to make a measurement of the vertical ozone profile using the differential absorption by ozone at two different wavelengths. The first laser is an XeCl eximer laser used to produce a 308 nm laser source with a maximum power of 10 W. The beam is passed through a 3X beam expander and has a final divergence of less than 0.1 mrad. The second laser is a tripled Nd:YAG which is used to produce a 355 nm laser source with a maximum power of 2.5 W. The beam is passed through a 2.5X beam expander and has a final divergence of less than 0.2 mrad. Both transmitted beams have a repetition rate of 50 Hz, and a 7 ns pulse width.

The receiver assembly consists of four ~~530 mm~~ 53 cm mirrors each having a focal length of 1500 mm, a field of view of 0.67 mrad, and an average parallax of 3100 mm. Each of these four telescopes are focused onto an actuator-mounted 1 mm diameter fibre optic. The outgoing signals are bundled before being passed through a mechanical signal chopper to block low altitude returns below 8 km which would saturate the photon counting electronics. The combined signal is split using a Horiba Jobin Yvon holographic grating with 3600 grooves/mm and a dispersion of 0.3 ~~mm/mm~~ mm/nm. The light from the grating is projected directly onto the photomultipliers for a high (92%) and low gain (8%) Rayleigh channel at 308 nm, a high gain (92%) and low gain (8%) Rayleigh channel at 355 nm, and two Raman channels at 331.8 nm and 386.7 nm for molecular nitrogen. The spectral resolution of the light incident on the photo cathode is on the order of 1 nm. Figure 2 shows a schematic of the OHP DIAL system.

All channels are sampled using a Licel digital transient recorder with a record time of 0.25 μs which corresponds to a vertical resolution of 75 m. Further details can be found in (Godin-Beekmann et al., 2003).

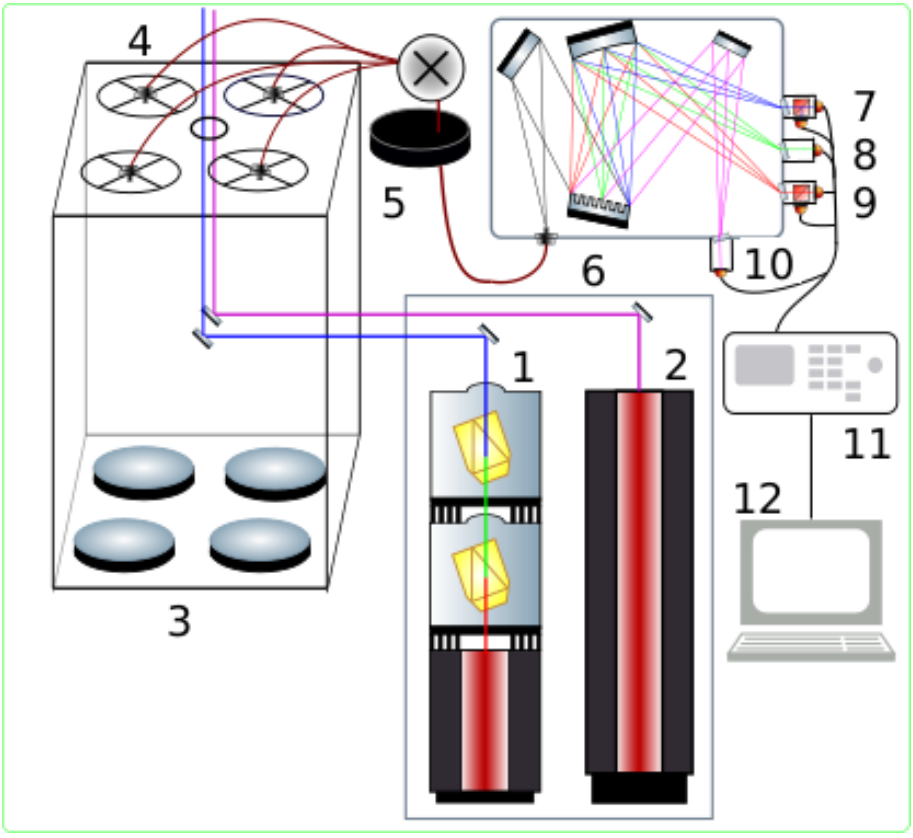

Figure 2: LiO_3S DIAL system. **1** 355 nm laser source, **2** 308 nm laser source, **3** four 530 mm mirrors, **4** four actuator mounted fibre optic cables, **5** mechanical chopper, **6** Horiba Jobin Yvon holographic grating, **7** 308 nm high and low gain photomultipliers, **8** 331.8 nm photomultiplier, **9** 355 nm high and low gain photomultipliers, **10** 386.7 nm photomultiplier, **11** Licel transient signal recorder, **12** Signal processing and analysis computer.

3 Methods

In this section we will set forth rigorous and well defined procedures for the retrieval of lidar temperatures in the middle atmosphere which will minimize the uncertainties at the upper limit of the lidar altitude range.

3.1 Rayleigh Lidar Equation

To calculate absolute temperature profiles from relative density profiles we exploit the gradient of the measured profile of back-scattered photons collected by the receiver. From classical lidar theory (Hauchecorne and Chanin, 1980), we know that the number of photons received is a simple product of transmitted laser power, atmospheric transmission, telescope geometry, and receiver efficiencies. This quantity can be expressed numerically in Eq. (1):

$$N(z) = \xi_{sys} \cdot \tau_{emitted}(z,\lambda) \cdot \tau_{return}(z,\lambda) \cdot O(z) \cdot P_{laser} \cdot \beta \frac{\lambda_{laser}}{h \cdot c} \cdot \sigma_{cross} \cdot n(z) \cdot \frac{A}{4\pi z^2} \cdot \Delta t \cdot \Delta z + B \quad (1)$$

N is the count rate of returned photons per time integration per altitude bin

z is the altitude above the detector

ξ_{sys} is the system specific receiver efficiency

$\tau_{emitted}(z,\lambda)$ is the transmittance of the photons through the atmosphere

$\tau_{return}(z,\lambda)$ is the return transmittance of the photons through the atmosphere

$O(z)$ is the overlap function of the receiver field of view

P_{laser} is the laser power at a given wavelength

~~β_{cross}~~ σ_{cross} is the backscattering cross section of the target molecule

$n(z)$ is the number density of scatterers in the atmosphere

$\frac{A}{4\pi z^2}$ is the effective area of the primary telescope

Δt is the temporal integration for data collection

Δz is the spatial range over which photons in a bin are integrated

B is the background count rate.

There are four simple assumptions we make when Eq. (1) is used. First, we assume that each photon we count only scatters once. While this is almost certainly not the case, we can say that it is approximately true. Visual wavelength photons have a very low probability of scattering in the atmosphere and with a multiple-scatter process we must square that very small probability. Of these multiply scattered photons, only those with a scatter angle towards the lidar receiver assembly will be seen, with the vast majority scattering outside ~~out~~ of the field of view. Further, the tenuous nature of the UMLT means that the small probability of detecting a photon which has scattered more than once becomes exponentially negligible with increasing altitude.

Second, we assume that the atmospheric density is directly proportional to the number of returned photons incident on the receiver assembly. In the case of high signal returns from the lower atmosphere, when the number of returned photons can saturate the photon counting electronics, the measured photon count rate will diverge from the received photon count rate. Multiple detection channels, at different sensitivities, are used to compensate for this effect. In this work we are primarily concerned with the UMLT, a region where lidars operate at very low count rates, so for the purposes of this work we can safely make this assumption. A correction for saturation in the lower stratosphere is described in Sect. 3.5.1.

Third, we assume that the atmosphere is in local hydrostatic equilibrium as well as local thermodynamic equilibrium (LTE) and obeys the ideal gas law. This assumption is potentially problematic at high altitudes where non-LTE processes can affect gravity wave dynamics and temperature profiles (Apruzese et al., 1984). ~~In this work we are unable to relax~~ However, given that a single lidar

profile is acquired every 2.8 minutes and a nightly average temperature is generated every 4 hours, we can have some confidence in this assumption.

Fourth, we assume that the atmosphere at mid-latitudes is generally free of aerosols above 30 km ~~and the lidar returns above this height are solely due to Rayleigh scattering processes (Hauchecorne and Chanin, 1980)~~ when there are no active volcanic or fire events (Hauchecorne and Chanin, 1980). During less severe background aerosol conditions (aerosol scattering ratio < 1.02), (Gross et al., 1997) suggests lidar temperature cold biases due to Mie scattering are less than 0.5 K at 20 km.

In the UMLT the signal to noise ratio and the model derived a priori assumptions for pressure and density are the main sources of error for the lidar temperature retrieval method. This paper lays out a rigorous method for reducing the noise in this region of the lidar signal with the goal of producing more robust mesospheric temperatures.

3.2 The Raw Counts Lidar Signal

When backscattered photons are incident on the lidar receiver they are ~~co-added~~ integrated for a set period of time in the counting electronics. This ensures that the recorded signals are based on a similar number of transmitted photons. In the case of LTA a photon count profile, as a function of arrival time, is generated for every 5000 laser shots. Similarly for LiO_3S a photon counts profile is produced for every 8000 laser shots. These measurements can be further ~~co-added~~ integrated for the entire night to increase the signal to noise ratio at the upper limit of the measurement range. We use the speed of light to convert our profiles of photon count rate per second as a function of arrival time at the detector to total photon count rate per second as a function of altitude.

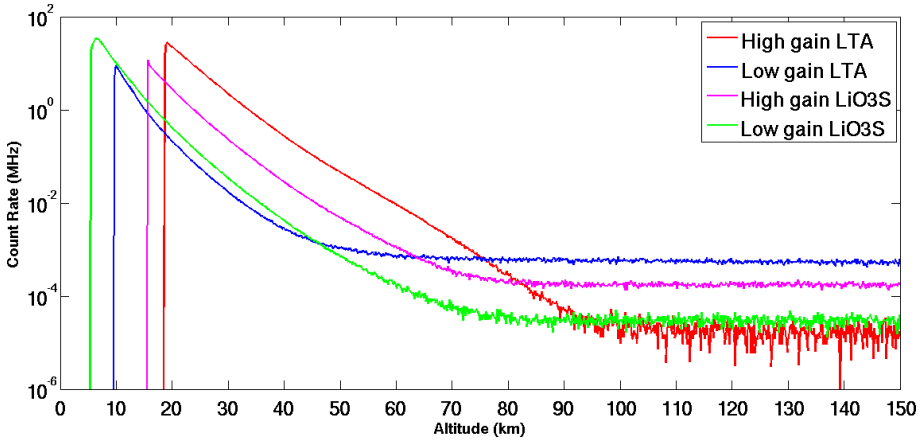

Figure 3: Nightly ~~co-added scans~~ integrated profiles for high and low gain Rayleigh signals for LTA and LiO_3S. The background for LTA extends to 246.23 km and for LiO_3S extends to 154.13 km. A single lidar ~~scan~~ profile for both LTA and LiO_3S has a temporal resolution of roughly 2 minutes and 45 seconds and a vertical resolution of ~~7.5~~ 75 m.

Figure 3 shows four nightly ~~co-added~~ integrated OHP lidar count rate profiles as a function of altitude. Both lidar systems employ a high gain and a low gain channel to extend the measurements over a greater altitude range. The lower altitudes (corresponding to the fastest signal return times) of each channel are either blocked by a mechanical chopper or electronically blanked. This is done to avoid saturation of the receiver assembly from very large signals in the lower atmosphere. Additionally, each channel has a set of optics designed to minimize the noise, with greater care being given to the high gain channels. These optics are fully described in the instruments Sect. 2.

3.3 Identifying Outliers, Signal Spikes, Signal Induced Noise, and Transient Electronic Interference

When retrieving lidar temperature profiles in the UMLT it is necessary to take extra precautions to carefully remove outliers, spikes, and electronic contamination from each profile in both the background region and the signal regions. Any contamination of the signal in the background region will be of the same order of magnitude as the true signal and thus, have a disproportionate effect on the temperature. An overestimation of the ~~noise~~ background due to localized signal contaminations will result in the removal of true photons, a lower estimated density, and by the ideal gas law, a ~~warmer~~ higher temperature. The ~~opposite holds true for an underestimation of the background (produces a colder profile). The~~ shape of the temperature profile itself will be distorted if there is a non-constant background. If it is not possible to fully correct the issue it is highly recommended to exclude the entire ~~scan~~ profile from the nightly analysis.

3.3.1 Spikes

Spikes in fast integration photon counting data are not always easy to spot but can be defined as anomalously large, isolated, signal rates which occur in only one altitude bin without affecting adjacent data. If not properly identified and extracted from the data they can contribute to false temperature features and inaccurate background estimations. The spikes can have many potential origins (thermal or electronic imperfection in the photomultiplier, small charges in the Licel digital recorder, interaction of the photocathode substrate with a cosmic ray, or dozens of different kinds of electronic 'cross-talk' between all the instruments at the observatory station) and are therefore impossible, in practical terms, to completely prevent in the lidar data set, and completely impossible to prevent in measurements which have already been made. Therefore, it is necessary to address this problem using software during the analysis. It is particularly challenging to separate small amplitude spikes when the signal to noise ratio approaches 1. It is therefore necessary to establish a consistent criterion to determine which data points belong to the the population of real lidar returns and which points are likely contamination spikes. We have chosen to employ a straight forward Tukey Quartile test (Tukey, 1949) on the difference between consecutively binned lidar returns as this statistic is relatively insensitive to signal drift during the course of the night. The quartile technique is equally

useful in both regions of high signal returns as well as the background regions and shows stability and consistency in identifying outliers. Figure 4 is a plot of photon count rate as a function of binned arrival time and shows an example of several photon count acquisitions plotted as a stack plot with the black line representing the 2σ limit on the population of lidar returns. Data points above the black line are considered as signal contamination and are removed from the analysis.

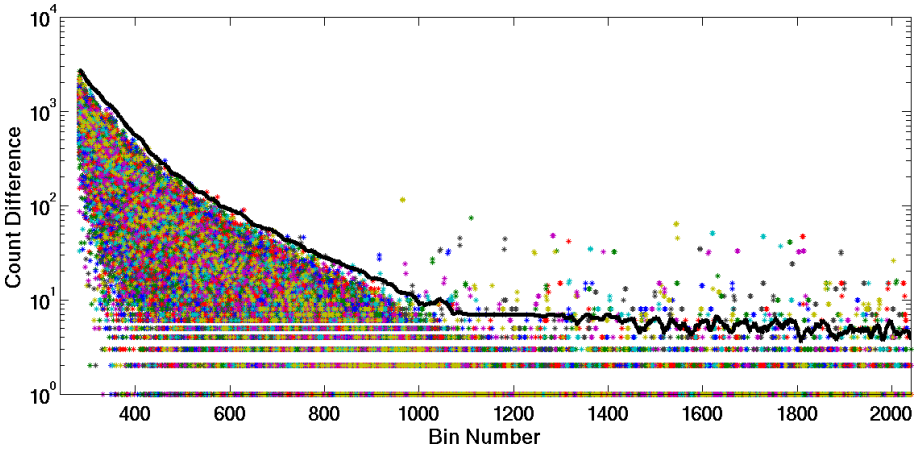

Figure 4: Tukey Quartile spike identification based on the signal difference between consecutive lidar time bins for short integration lidar returns. An entire night of lidar ~~scans~~ profiles is over-plotted in the stack plot. The black line is the 2 sigma limit and points above this line are removed.

3.3.2 Transient Electronic Signals

Transient Electronic Signals (TES) are short lived bursts in the lidar acquisition chain and may be internal to the system or related to nearby electronic interference. Possible sources for these transients include photomultiplier ringing from signal saturation, voltage fluctuations in the power supply, ambient RF signals, and ground loops between lidar electronics and Ethernet switches with metal sheathed cables. While these events are rare they can drastically alter the background and resulting temperature profile by inducing wavelike structures into the data.

Unlike simple spikes these features have an amplitude, a duration, and an effect on the ~~downstream counting rate~~ counting rate in bins subsequent to the TES burst. In the example shown in Fig. 5 ~~(top)~~ is a surface plot of counts differences between consecutive altitude bins for the first 100 altitude bins of lidar data. Each bin is 0.1 µs wide. This plot shows ~~scans~~ profiles for a night of lidar data with each ~~scan~~ profile accounting for roughly 1.6 minutes of lidar data. We can see that the 22nd and 46th ~~scans~~ profiles are contaminated by a TES with a duration of about 0.5 μs. These signals cannot be detected using the Tukey Quartile test as the time derivative of the photon return signal may not be sufficiently far from the nightly population median. However, a 2-D kurtosis test will consistently detect this type of signal contamination as a TES will induce a large skew in the photon count rate population distribution. The kurtosis test is done in the time dimension as well as with altitude to

exclude false positives in the photon count rate skew which may be due to clouds or aerosols. Figure 5 (bottom) shows a plot of the kurtosis in the population of photon counts in each lidar ~~scan~~ profile and the red line shows the 2σ estimation of total lidar ~~scan~~ profile skew. Isolated ~~scans~~ profiles with a total kurtosis above this limit are excluded.

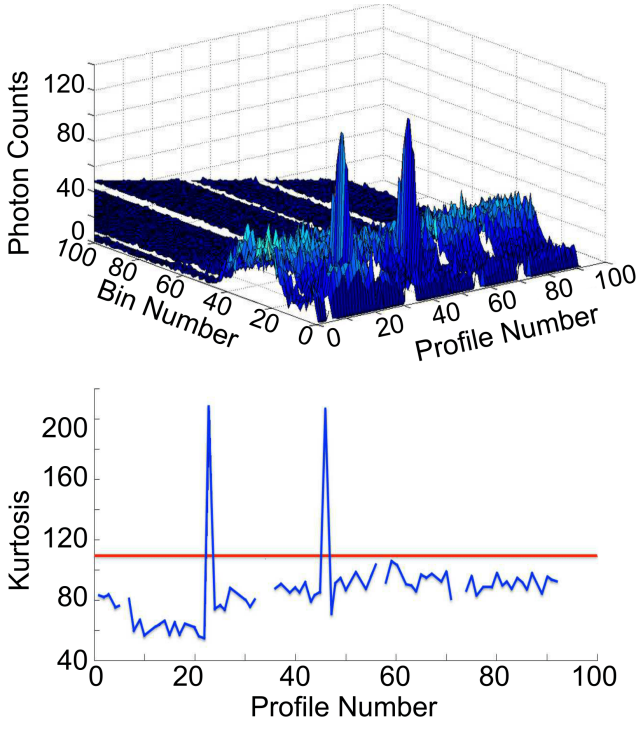

Figure 5: Upper panel is a surface plot of lidar returns as a function of time bin and ~~scan~~ profile number. For clarity, only the first 100 bins are shown in this plot. The test is carried out using all bins of each profile. Two instances of TES can be seen as anomalous peaks in the photon count rate. Lower panel is a summation of the fourth statistical moment (kurtosis/skew) for each scan. The red line indicates a ~~2σ~~ 2σ limit on the skew of the population. Points above the limit are excluded.

3.3.3 Bad ~~Scans~~Profiles

After the removal of lidar ~~scans~~ profiles which suffer from clear signal contamination, there may still be ~~scans~~ profiles which ought not be included in a lidar temperature analysis because they are outliers of poor quality compared to other profiles within the same night. Conceptually, 'bad ~~scans' are lidar scans~~ profiles' are lidar profiles with a high background and/or a low signal strength ~~. These scans~~ as compared to profiles measured shortly before or after the profile in question. These profiles need to be positively identified as not belonging to the general population of nightly lidar ~~scans and excluded~~profiles. Quantitatively, identifying a 'bad ~~scan~~profile' is a challenge as both the background and the signal can change abruptly over the night as the laser power drops or sky conditions change (see Fig. 6 for an example). In the top panel of the figure we see the evolution

of the background for a night of lidar data. ~~Intuitively, we~~ We might suggest that ~~scans~~ profiles 1 through 23 and ~~scans~~ profiles 36 through 46 might belong to one population and the rest (excluding ~~scan~~ profile 69) belong to a second population. However, when we look at the panel representing the signal~~our intuition becomes a bit more subjective. There are clearly four groupsof~~ , it is equally reasonable to, instead, interpret the plot as containing four groups. Each of these groups has similar signals which match fairly well with the changes in the backgrounds shown in the panels above ~~however,~~ (profiles 1-23, profiles 24-35, profiles 36 - 48 and profiles 49 - 92) . However, whether these four groups of signals ~~represent~~ should be treated in analysis as two, three, or four distinct populations is ~~somewhat dependent on which statistics the author chooses to use.~~

~~We have shown~~ open to interpretation. Therefore, we seek an objective programmatic solution for identifying bad profiles. We now show two approaches for attempting to address the issue of changing signal quality. ~~Both have advantages and points for concern and illustrate just how subtle and challenging this aspect of lidar science can be.~~ In Fig. 6 the green margin is an attempt to identify 'bad ~~scans~~profiles' based on a moving average approach however, this method cannot accommodate quick transitions in signal strength and results in false positives when signal quality changes abruptly. ~~The blue line is an attempt to use Matlab Neural Network software to estimate the number of lidar signal-to-noise populations for a given night. This approach was abandoned as the training process for the software requires an exhaustive set list of example 'bad profiles' which we cannot supply. Additionally, we found that estimating the number of local medians for each sub-population of lidar scans in a given night was too highly dependent on the number of degrees of freedom specified in the Matlab tool.~~

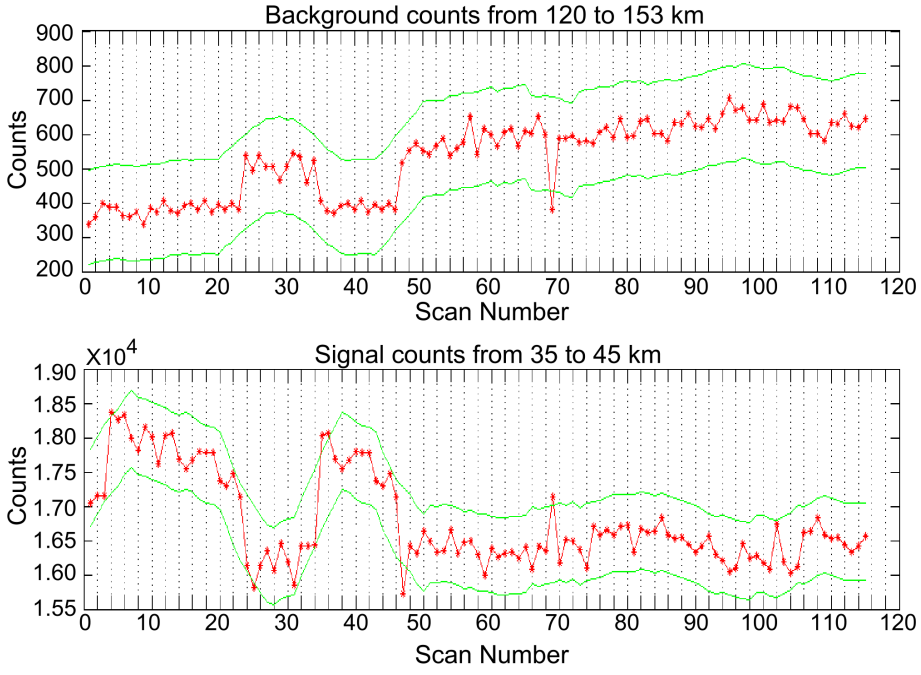

Figure 6: Example of lidar signal and noise during a night of measurements. Top panel shows the total background counts summed from 120 km to 153 km and the bottom panel shows the total signal summed between 35 km and 40 km. Green bounds are calculated based on a ~~simple moving average~~ smoothed 2σ error estimation of the summed photon counts~~(red) and the blue line is an attempt to estimate local population medians using the Matlab Neural Network tool~~.

The simple reality of ground based observation means that lidar signals clearly detect changes in the viewing conditions such as moonrise, thin cirrus clouds, optically thick clouds, changing light pollution, as well as changes in signal quality. Systematically identifying outlier signals is further complicated as there can be multiple signal to noise population medians during the course of the night. To properly characterize the non-Gaussian distribution of ~~scans~~ profiles and determine which should be excluded ~~requires~~ we require a non-parametric statistic. We use a one sided non-parametric Mann-Whitney-Wilcoxon rank-sum test (Mann and Whitney, 1947) to identify lidar ~~scans~~ profiles which do not belong to the nightly population or subpopulations of lidar ~~scans~~profiles.

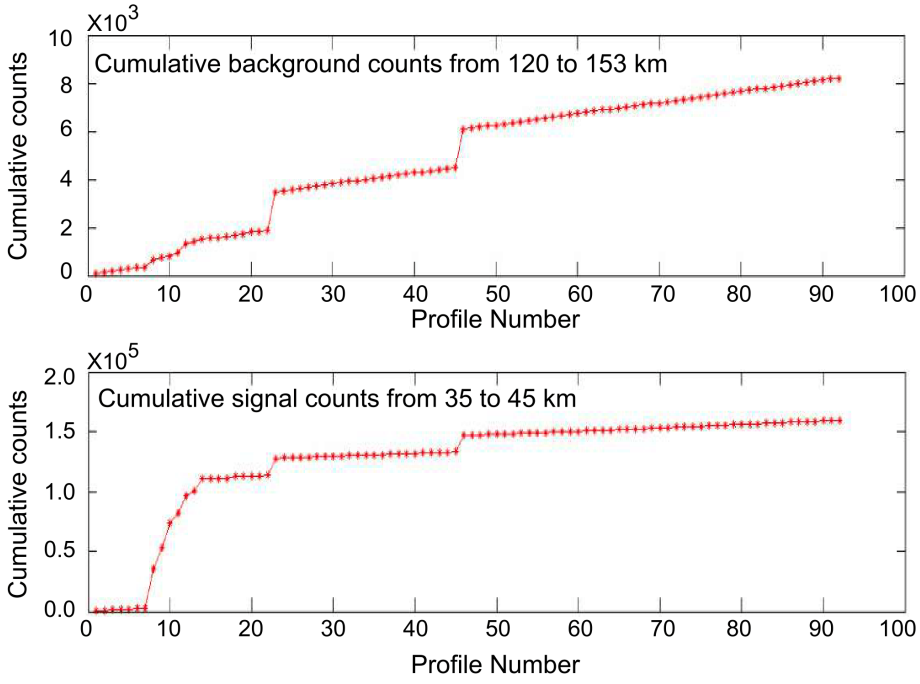

Figure 7: Rank sum plots for a night of lidar data. Top panel is the cumulative background count and the bottom panel is the cumulative signal count. The signal to noise ratio of the rank summed photon counts in each profile is evaluated using a Mann-Whitney-Wilcoxon rank-sum test to determine if an individual lidar ~~scan~~ profile belongs to the nightly population of lidar ~~scans~~profiles.

Figure 7 shows the ranked sum of the background (noise) and signal counts for a night of lidar data. We do not exclude the profiles which fail the test for having high quality. The benefit of using this metric is that it allows us to have a standardized definition of a 'bad ~~scan~~profile' which takes into account the nightly median without the assumption that the quality of lidar ~~scans~~ profiles is normally distributed. In this example the first 13 ~~scans~~ profiles fail the rank-sum test and are discarded.

3.3.4 Good ~~Scans~~Profiles

Given that our objective is to calculate accurate temperature profiles at the highest possible altitudes we must quality test each ~~scan~~ profile that we choose to include in the nightly average. It is possible to include partial ~~scans~~ profiles but that is not done in this work. The conceptual difference between a 'bad ~~scan~~profile' and a 'good ~~scan~~profile' is that bad ~~scans~~ profiles are positively identified as outliers to the general population whereas good ~~scans~~ profiles represent the portion of the population of ~~scans~~ profiles which contribute more information than noise to the nightly average at a given altitude. Consider that a poor quality lidar ~~scan~~ profile which has a signal to noise ratio of 1 at 70 km contributes more information ~~than noise~~ from the signal than from the (background + noise) at 60 km, but ~~more noise than signal~~ less information from the signal than from the (background

+ noise) at 80 km. Thus, we need a flexible metric to determine signal quality over a diagnostic altitude which reflects the general signal quality of the night.

Quantitatively, we express this with a signal, S, to noise, N, inequality in Eq. (2). The ~~noise is always evaluated~~ background (noise) of an individual profile, N_i, is expressed as the summation of photon counts in bins which fall between 120 km and 155 km and the ~~altitude range for the evaluating the signal is defined as the scale height below the point~~ nightly background, N_{sum} is the summation of all N_i for the night. To determine a metric for the nightly average lidar signal, S_{sum}, we first calculate a quick density profile and determine the lowest altitude where the signal to noise ~~equals one in the density profile. Each individual scan has a value representing the signal, S_i, and a noise, N_i. The scan values are compared to the nightly sum of the signal,~~ ratio equals 1. We chose a cutoff value of SNR=1 because it is the least strict value we could use which ensures that we have more information than noise (or, specifically, more information than noise plus background counts), at the altitude within the density profile where we begin the downward temperature integration. Had we chosen a criterion which was less strict (SNR$\ll$1), we would expect to see more statistical variability in the top altitudes of the temperature retrieval as a result of starting the temperature integration in a region which contains more noise than signal. Conversely, choosing a criterion which is too strict (SNR$\gg$1) limits the maximum altitude of the temperature retrieval as discussed in Sect. 3.6.1. The SNR = 1 point forms the upper bound of the altitude range from which we derive the representative signal for the profile. The lower bound of this representative signal range is defined to be one density scale height ($\sim$8 km) below the upper bound. The lidar range bins which correspond to this altitude range are then summed to yield S_{sum} ~~and the nightly sum of the noise, N_{sum}~~. A similar calculation, using the same range bins as in the nightly average calculation, is done to determine the signal of a single profile, S_i. If a ~~scan~~ profile fails the inequality test then it is not included in further nightly analysis.

$$\sqrt{\frac{S_{sum}+N_{sum}}{S_{sum}}} < \sqrt{\frac{(S_{sum}-S_i)+(N_{sum}-N_i)}{S_{sum}-S_i}} \tag{2}$$

3.4 Noise Reduction

Statistical uncertainty in photon counting can be described by a Poisson distribution based on the square root of the number of photons received. Systematic uncertainties in the photon counts are introduced by ambient background light (light pollution, moonlight etc.), thermal excitation in the photomultipliers (so-called dark current), and signal induced noise. The first two sources of error are minimized by using narrow filters in the optical receiver chain and by cooling the photomultipliers. The signal induced noise can be very difficult to correct experimentally and is usually estimated in data processing. This type of noise can occur if the photomultipliers have become saturated at any point in the signal acquisition process and often manifest as non-linear artifacts superimposed upon the true photon count profile.

Figure 8 shows the reduction in the background noise due to recent hardware improvements. The first drop corresponds improvements made to the photomultiplier cooling system which reduces the number of thermally excited electrons detected at the photo cathode of the photomultiplier in the absence of signal from the sky. The second drop in background counts results from replacing the Hamamatsu R7600U-20 multi-alkali photomultiplier with the improved Hamamatsu R9880U-110 photomultiplier having a super bi-alkali photo-cathode. The third and final drop in background counts is a result of replacing a 532 nm optical filter which has a width of 1 nm with a newer filter having a bandwidth of 0.3 nm. These experimental modifications result in a 100 fold decrease in the background noise and allows us greater confidence in our UMLT temperature retrievals. The regular monthly variations in the signal which become apparent at lower noise levels are due to the phase of the moon.

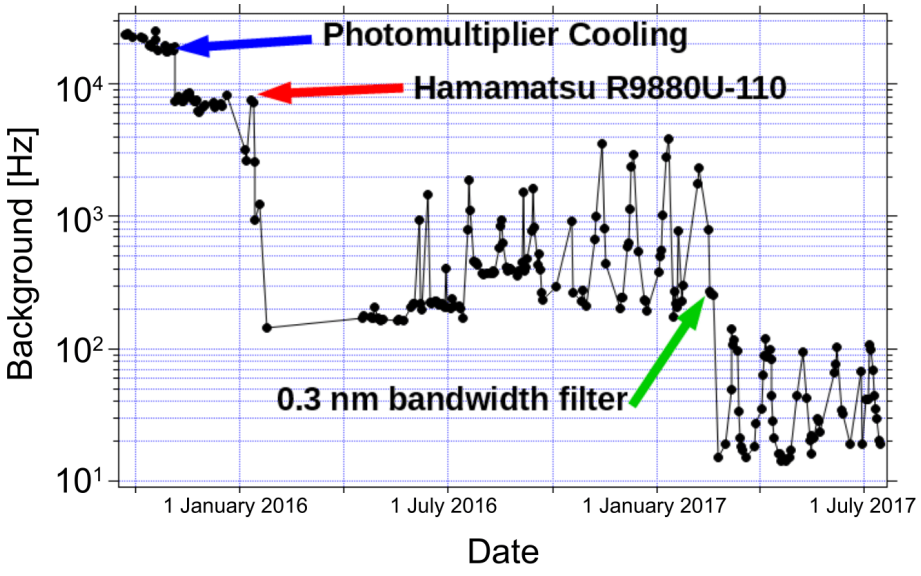

Figure 8: This figure shows the improvements in the background count rate due to photomultiplier cooling, new photomultipliers, and new optical filters. Note the logarithmic y-axis and the total reduction of background counts by more than 2 orders of magnitude.

3.5 Corrections Applied Before Temperature Calculation

In the previous subsection we detailed the process for removing bad ~~scans~~ data points and profiles from our nightly lidar measurement. In this subsection we will detail several corrections to our remaining photon ~~counts~~ count profiles which correct for signal saturation, atmospheric transmission, and background estimation.

3.5.1 Deadtime Correction

The OHP lidars measure photons using photomultipliers and a digitizing signal counter. This system is highly efficient at detecting low signals and is optimized for single photon returns in the UMLT. However, given that the returned lidar signal directly follows the exponential density of the atmosphere, the photomultipliers and counting systems are susceptible to missing photons at lower altitudes due to high count rates. To correct for this saturation effect we can estimate a correction coefficient, τ, also referred to as a deadtime.

The background theory and derivation of Eq. (3) is well described by (Donovan et al., 1993), where ~~N is the photon count rate and Δt is the~~ $N_{received}$ is the number of photons incident on the PMT per measurement time interval and $N_{counted}$ is the number of photons per measurement time interval ~~.~~ which are actually counted by the system. In general, $N_{counted} < N_{received}$ due to effects of the system deadtime. This deadtime correction can be calculated based on factory specification of the counting electronics, a theoretically derived deadtime, or it can be measured directly using a low gain lidar channel. The OHP lidars measure the deadtime directly and correct for saturation in the high gain channels with information from the low gain channels. If the low gain channel is not available a theoretical correction of 7 ns is applied to pre-2013 data and 4 ns is applied to more recent data following the installation of a Licel digital recorder.

In order to measure the deadtime experimentally, we assume that the low gain channel, because it has low photon count rates, will always operate in the linear response regime and will never suffer from deadtime effects. Thus, it represents a value proportional to the 'true' rate for returned photons for each altitude. Once scaled by a constant (e.g. using MSIS or another model), we can use this count rate as $N_{received}$.

The high gain channel, conversely, measures higher photon count rates at every altitude than the low gain channel does. Similarly to the low gain channel, at the low end of its dynamic range, the high gain channel operates linearly, and therefore represents a value proportional to the 'true' rate for returned photons for each altitude. The constant of proportionality is different for low and high gain channels. At low count rates, the scaled counts measured by the high gain and low gain channels are equal. As photon count rates move into the higher end of the high gain channel's dynamic range, deadtime begins to have an effect: The high gain channel will measure too few photons compared to the 'true' rate; the number of photons which are returned to the lidar. Therefore, we call the scaled high gain count rate $N_{uncorrected}$ in Eq. (3); it has not yet been dead time corrected. We will refer to the deadtime corrected scaled high gain count rate as N_{dtc}. Equation (3) is used several times. First, we use data only from altitudes for which the low gain and high gain channels both have measurements (nominally 40 to 60 km). We iterate through various values of τ, calculating a N_{dtc} for each $N_{uncorrected}$ value. This is carried out until the difference between $N_{corrected}$ (from the high gain channel) and $N_{received}$ (from the low gain channel) is minimized. This determines the dead time of the system, τ. Next, Eq. (3) is used again, using the measured nightly value for τ, to

calculate N_{dtc} for all $N_{uncorrected}$ high gain channel measurements. This allows us to correct the high gain measurements for the entire profile.

$$N_{counted_{dtc}} = N_{received_{uncorrected}} * exp(\frac{\tau * N_{received}}{\Delta t} \frac{\tau * N_{uncorrected}}{\Delta t}) \quad (3)$$

3.5.2 Atmospheric Transmission Correction

To correct for Rayleigh extinction we use MSIS-90 model (Picone et al., 2002) to generate a vertical profile of ozone, molecular oxygen, oxygen radical, molecular nitrogen, and argon, and then apply the correct Rayleigh cross-section to each species. This method is adapted from (Argall, 2007) and is important for accurate retrievals of density and neutral temperature in the UMLT. Correction for aerosols is not done in this work as we assume that the atmosphere is generally clean above 30 km (Hauchecorne and Chanin, 1980).

3.5.3 Defining the Background

Normally, we assume that the rate of counted photons per laser shot is constant in the background region during the signal acquisition time and can therefore be approximated by a simple Poisson distribution. We further assume that in this background region we are not measuring returned photons from the laser signal but instead are measuring ambient sky light. However, if there is ~~a~~ non-linear signal induced noise in the photon counting chain, the number of counted photons is not constant with time during the acquisition period of a single laser shot. When this occurs we cannot assume that the variation in the background is a strictly Poisson distribution around a constant expected value.

If left uncorrected, we risk overestimating the number of 'true' photons returned from the upper atmosphere and the result is an artificially dense and cold UMLT. Erring on the side of caution we fit three backgrounds (constant, linear, and quadratic) to each nightly summed profile, in a standard diagnostic region, and choose the function with the best Chi-squared goodness of fit as our estimate of signal induced noise. The best background function is subtracted from the raw photon counts profile. Shown in Fig. 9 is an example of a night where the low gain Rayleigh channel (blue) experienced signal induced noise which was best approximated by a quadratic function; the high gain Rayleigh channel (red) had a background best estimated by a small negative linear function; and the nitrogen Raman channel (green) had no apparent signal induced noise and was fit with a constant background. The optimal solution for non-linear signal induced noise is to determine the contribution of both the signal and the noise using exponential fits however, we have found that method to be extremely sensitive to the choice of background diagnostic region and was less stable than the simple quadratic approximation.

We have some confidence that the quadratic background correction to the low gain channel correctly approximates the moderate non-linear signal induced error because we can compare the

corrected low gain channels to the high gain channel. In the overlap region we have two channels making coincident measurements and we can safely assume that the response rate for the high gain channel is linear. Therefore, a correction for signal induced noise in the low gain channel which brings the resulting low gain count rates into the closest agreement with the high gain channel count rates at the same altitudes will be the optimal choice for the correction. In some cases, the quadratic correction for signal induced noise in the low gain channel yields better agreement than the constant or linear corrections, in which case it is employed. The best individual choice (constant, linear, quadratic) is used for each profile. We believe these empirical corrections to be sufficient, because (a) the resulting agreement with the high gain channel improves as compared to the uncorrected profile, (b) the resulting corrected low gain count profiles are generally equal to the high gain count profiles to within statistical uncertainty, and (c) for the few cases in which the empirical correction ultimately fails, this will be apparent by the corrected signal retaining poor SNR values. The melding procedures of Section 3.6 weight the combined high and low gain Rayleigh channels according to SNR, and so in these cases, the poorly-corrected low gain contributions to the final melded counts profile will be negligible, and all information will be obtained from the high gain channel.

For the quadratic case, as soon as there is signal induced noise the profiles no longer represent Poisson distributions as the count rate in each lidar bin is no longer fully independent of the count rates in the bins on either side of it. Therefore, precise calculations of the SNR would require the addition in quadrature of real noise (from sky background and signal photon counts) and contamination noise (from signal induced noise). Here, however, we make the assumption that the signal induced noise is able to be completely removed from the raw profiles with the subtraction of the quadratic function. We therefore interpret the background subtracted profiles to obey approximately Poisson distributions, thereby approximating the total noise in the profile to the noise of only the real photons, which can be treated as uncorrelated. Our standard altitude range for background selection is 120 km to 155 km but this number is system and channel specific. To illustrate this point we compare the background regions of the high gain Rayleigh channel (red) and the nitrogen Raman channel (green) in Fig. 9. The nitrogen Raman channel background could be calculated from 50 ~~km~~ to 155 km or 120 ~~km~~ to 155 km and yield the same result.

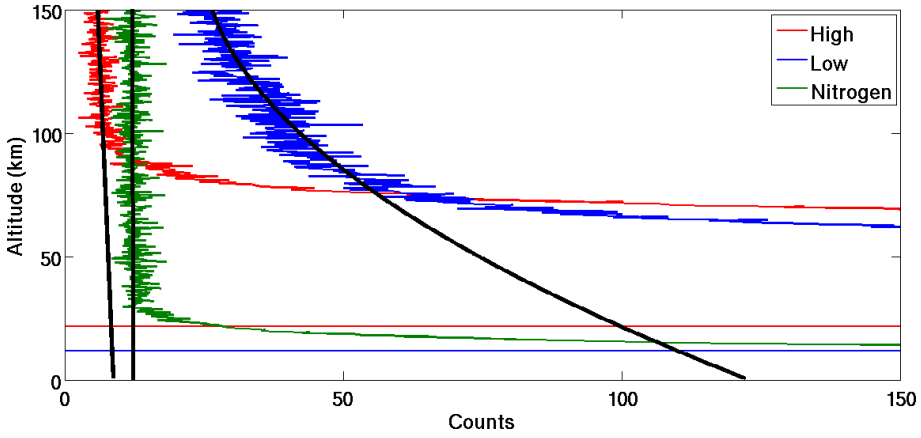

Figure 9: An example of ~~a~~ non-linear signal induced noise in the low gain Rayleigh channel best estimated by a quadratic background. Also shown is the high gain Rayleigh channel (red) with a background best fit by a negative linear function and the nitrogen Raman channel (green) with no apparent signal induced noise and a constant background.

3.6 Temperature Inversion Equation

The standard NDACC algorithm for Rayleigh temperature retrieval is the Hauchecorne-Chanin (HC) method (Hauchecorne and Chanin, 1980) which makes a scalar normalisation of the photon-count profile to an in-situ density measurement or to a density calculated from a model like CIRA-72, SPARC-80, or MSIS-90. From a density gradient profile we calculate a pressure gradient profile Eq. (4) and using the ideal gas law, Eq. (5), we can arrive at an expression for pressure, Eq. (6). Here P is pressure, z is altitude above the lidar station, ρ is density, g is the latitude dependent acceleration due to gravity for an ellipsoid Earth given by the Somigliana formula, R is the ideal gas constant, T is the temperature, and M is the molecular mass.

$$dP(z) = -\rho(z)g(z)dz \tag{4}$$

$$P(z) = \frac{R\rho(z)T(z)}{M} \tag{5}$$

$$\frac{dP(z)}{P(z)} = -\frac{Mg(z)}{RT(z)}dz = d(\log(P(z))) \tag{6}$$

The crux of the challenge for initializing the lidar equation lies in the non-linear nature of Eq. (6) which will necessitate the introduction of an a priori estimate of pressure at the top of the atmosphere followed by an iterative approach to retrieving the profile at lower attitudes. A full theoretical description of this problem was well laid out by (Khanna et al., 2012). In this work we have chosen to take our initial a priori seed pressure value, $P(z_1)$, from the MSIS-90 model. We now arrive at an iterative expression for the generation of the pressure profile as a function of altitude Eq. (7).

$$\frac{P(z_i) - \frac{\Delta z}{2}}{P(z_i) + \frac{\Delta z}{2}} = \exp\frac{Mg(z_i)}{RT(z_i)}\Delta z \tag{7}$$

Given our iteratively generated pressure profile we can do an inverse calculation to map our pressures to a set of temperatures using Eq. (8) and Eq. (9). This iteration starts at the top of the atmosphere, in ~~an area~~ a region of low signal to noise and thus of large relative uncertainty, and proceeds downwards in altitude and becomes exponentially less uncertain with each step as signal quality improves with increasing atmospheric pressure. As we iterate downward the influence of our choice of a priori pressure becomes less significant and the calculated temperature profile becomes entirely data driven.

$$X_i = \frac{\rho(z_i) g(z_i) \Delta z}{P(z_i) + \frac{\Delta z}{2}} \tag{8}$$

$$T(z_i) = \frac{M g(z_i)}{R \log(1 + X_i)} \Delta z \tag{9}$$

In order to calculate a single temperature profile from 5 km to above 80 km we meld the photon counts from the high and low gain Rayleigh channels together with the counts from the N_2 Raman channel. The slope of the logarithm of each of the three photon counts profiles is compared to a synthetic lidar counts profile generated based on the nightly average MSIS-90 density profile. The comparison gives us a first estimation of the linearity and alignment of the lidar data. We then select a clear linear region of each ~~scan~~ profile to use in calculating a MSIS derived scaling factor for each profile. This procedure allows the top of the nitrogen Raman profile to be melded to the bottom of the low gain Rayleigh profile and the top of the low gain Rayleigh profile to be melded to the bottom of the high gain Rayleigh profile. The melding calculation is conducted over a signal-to-noise defined altitude range and is a straightforward weighted average. The resulting melded density and pressure profiles are used to generate a single nightly average temperature profile like the one shown in Fig. 10. The use of MSIS-90 as a scalar density reference for the synthetic lidar profile does not affect the final lidar temperature profile which depends only on the relative density and not the absolute value. We follow similar procedures to those described by (Alpers et al., 2004).

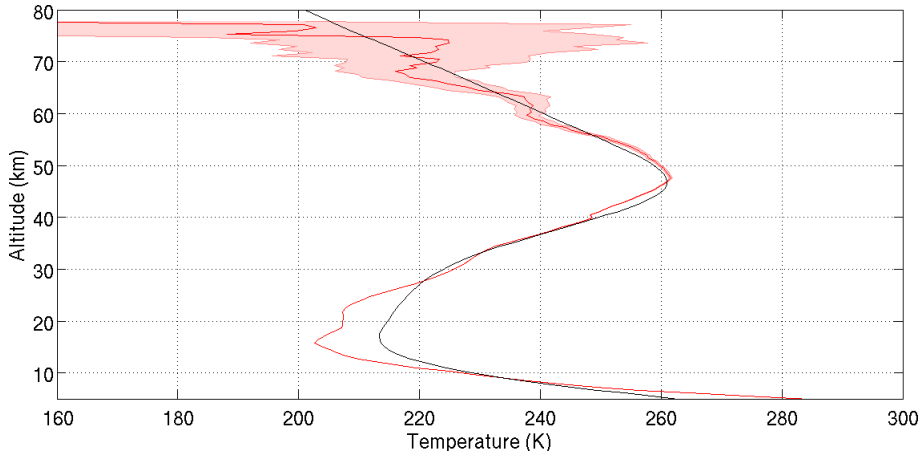

Figure 10: An example of a nightly average melded temperature profile from two Rayleigh channels and one Raman channel. The profile is calculated at 300 m vertical resolution from a single combined photon count profile and has a maximum relative error near 80 km of 30%. Black line is the MSIS-90 temperature profile which corresponds to the MSIS-90 pressure and density information we used as an a priori.

3.6.1 Where to start the inversion

As can be seen in Eq. (8) and Eq. (9) the calculation of lidar temperature requires an a priori guess of pressure at the top of the atmosphere and a relative density gradient. Given that the signal to noise in the UMLT can be very low, the choice of a priori as well as the uncertainties in the density gradient can have a very large effect on the temperature profile (Khanna et al., 2011). As a result, it is prudent to remove the top 15 km of the retrieval to minimize the contribution of the a priori (Leblanc et al., 1998b).

In our treatment the a priori pressure is selected at the altitude where the signal to noise ratio in a smoothed photon counts profile is 1. The resulting temperature profile is subsequently cut when the relative error exceeds 30 percent. This treatment is not the optimal solution for the retrieval altitude as a fully Bayesian algorithm is required to properly characterize the influence of the a priori choice (Sica and Haefele, 2015). However, we believe that our signal to noise metric is sufficiently rigorous, and more importantly reproducible.

4 Net result of temperature algorithm modifications

~~By implementing the changes from the previous section to both raw data processing and lidar temperature retrieval described in this section~~ The NDACC algorithm contains such corrections as deadtime, background, and transmission. The new algorithm improves upon the background correction and identification of bad profiles, and introduces corrections for: signal spikes, TES,

identification of good profiles, and noise reduction, all which have not previously been addressed by the NDACC algorithm.

The LTA data is recorded and saved at 75 m resolution. The spike and TES corrections described in sect.3.3.1 and 3.3.2 are carried out at this resolution. Then the profiles are integrated to 300 m, at which point the remainder of the corrections in Section 3 are applied.

Temperature profiles using the new algorithm are calculated at 300 m resolution for LTA, and are plotted as the green line in Fig. 11. This is higher resolution than the standard NDACC temperature resolution, which is 1 km, smoothed to 2 km effective vertical resolution. The LTA NDACC-calculated temperatures (black line in Fig. 11) are plotted at 2 km effective resolution. By implementing the new algorithm, we have cooled the UMLT lidar temperature retrievals with respect to the standard NDACC temperature algorithm. ~~This cooling reduces the lidar-satellite warm bias which was noted in the introduction.~~ The modifications cool the mesospheric retrievals by approximately 5 ~~K~~ K near 85 ~~km~~ km and 20 ~~K~~ K by 90 ~~km~~km. There is no significant ~~change to the lidar temperatures~~ difference between the new and the NDACC algorithms for LTA below 70 ~~km~~km.

~~Figure 11 shows the ensemble median difference between the temperatures produced using the standard NDACC temperature algorithm on LTA data (black), with the modified algorithm (green), the temperatures produced by~~ Temperature profiles calculated for LiO_3S are all carried out using the NDACC algorithm at an effective vertical resolution of 2 km, and these are shown as the orange line in Fig. 11. Whereas the LTA NDACC algorithm results are warmer than the LiO_3S ~~(orange), the~~ NDACC algorithm results above 70 km, we now see that the LTA new algorithm results are cooled sufficiently that they more closely match the LiO_3S measurements up to 78 km. Therefore the corrections for LTA proposed in the new algorithm represent a significant improvement over the LTA NDACC algorithm for altitudes above 70 km.

A comparison with temperature retrievals from the satellites MLS (red line in Fig. 11) and SABER (blue with shaded ensemble variance), and with the MSIS-90 model (magenta ~~)~~line in Fig. 11), also shows an improvement in the LTA temperatures retrieved using the new algorithm as compared to the LTA NDACC algorithm. By implementing the techniques described in the sections above we can account for nearly half of the temperature difference between the lidar and the satellites at 90 ~~km~~km. The character change in the difference functions above and below 84 ~~km~~ km is in part due to the increasing contributions of the species specific Rayleigh backscattering correction and the corrections to the gravity vector. The remaining temperature difference between the improved lidar temperatures (green) and the satellites and model may be in part due to distortions in the satellite a priori for the geopotential vector. This possibility is explored further in the companion paper, and all coincidence criteria for the satellite comparisons are available therein (Wing et al., 2018b).

It is important to note that additional complications exist when comparing temperatures derived from ground based lidars to temperatures derived from satellite data which have their own calibration concerns. We explore the issues of lidar-satellite comparison in part B of this paper (Wing et al., 2018b)

. A co-located ground-based resonance Doppler or Boltzmann lidar would provide a better comparison data set as resonance lidars have high signal to noise ratios above 75 km (Alpers et al., 2004).

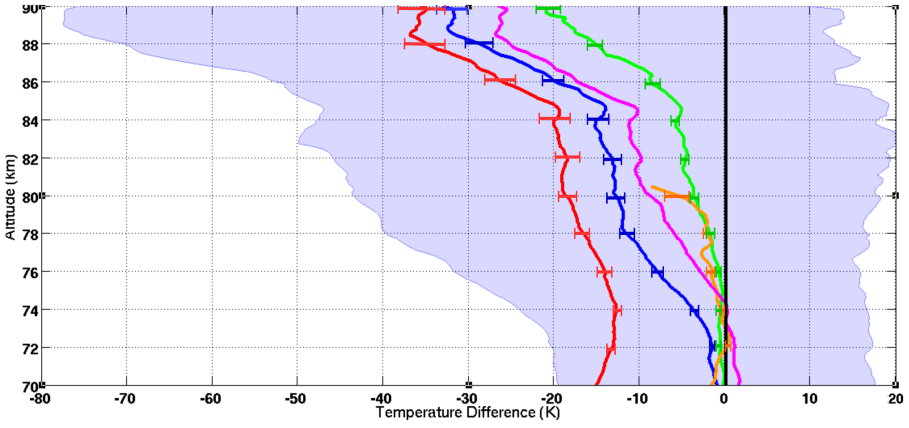

Figure 11: Ensemble temperature differences from NDACC standard LTA Rayleigh temperatures (black). MLS (red), SABER (blue with shaded ensemble variance), MSIS-90 (magenta), LiO_3S (orange), and LTA Rayleigh temperatures with corrections given in this work (green).

5 20 Year Comparison of OHP Lidar Temperatures

Conducting systematic inter-comparisons between independent lidar systems is essential for assuring data quality and is a requirement for NDACC certified instruments. Most comparisons are conducted on a campaign basis where two or more lidar systems are co-located and make coincident measurements. A good example of this type of work was the stratospheric lidar and Upper Atmospheric Research Satellite (UARS) validation campaign (Singh et al., 1996). ~~This~~ The present study proposes a completely novel type of inter-lidar study on the ~~long-term~~ long term stability of the Rayleigh lidar technique. The first step in our analysis is to compare the temperature profiles from the LTA and LiO_3S systems. LTA temperatures were calculated using the OHP NDACC temperature code and LiO_3S temperatures were calculated using a modified version of the same code. There are very few significant differences between these two codes. The most important difference involves the choice of parameters for melding the high and low gain channels for the two systems. Given the differences in the relative gain between the four lidar channels being considered, the melding of LiO_3S often occurs at a lower altitude than LTA. ~~This~~ The present study considers temperatures ~~in~~ between 35 km and 75 km to ensure that we are well above any contamination from aerosols and below any significant initialization errors. From Fig. 11 we can see that there is no significant difference in the temperature outputs of these two algorithms (black baseline and orange) or with the improved algorithm (green) below 75 km.

We selected the data from 1993 to 2013 for the comparison as both instruments operated regularly and without significant design changes during this time. Since the ~~two~~ lidars are co-located and are operated by the same technicians they often make measurements simultaneously. Figure 12 shows the average number of measurements per month made by the LTA and LiO_3S which were included in ~~in~~ this study as well as the average number of common measurements per month. We defined common measurement times based on more than 80% temporal overlap, good quality ~~scans~~ profiles in both systems, and good internal alignment of both lidars. Of the 2482 nights of LTA data and 3194 nights of LiO_3S, 1496 nights met our criteria for coincidence.

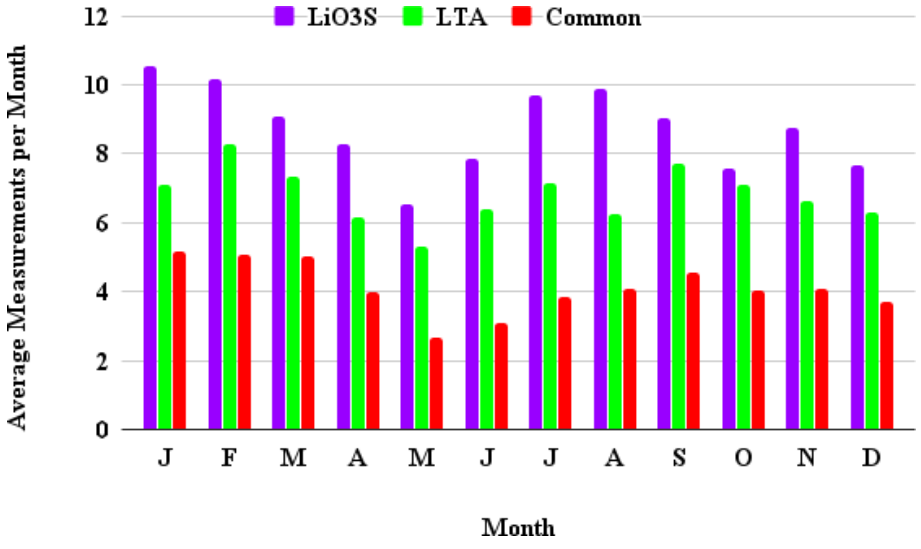

Figure 12: Average number of OHP lidar temperature measurements per month during the period of 1993-2013.

Figure 13 shows the nightly temperature differences between the two lidar systems. The 20 year data set contains 1496 coincident measurements lasting longer than four hours. Black vertical rectangles indicate some of the time periods where the high or low gain channels were mis-aligned in one or the other lidar. Internal misalignments happen when one or more of the five mirrors in LTA or four mirrors in LiO3S are not properly aligned with the laser or the fibre optic is not centered on the focal point of the mirror. A few of these time periods can be associated with minor system modifications. Misaligned lidar signals were identified by comparing the slopes of the density profiles in the high (generally above 50 km) and low (below ~50 km) gain channels of each system. A simple chi-squared test was used to detect these nights and exclude them from the rest of the analysis. It is possible that the criteria described above for identifying periods of misalignment is not yet stringent enough. Therefore, one limitation of the OHP measurements in terms of accuracy, and depending on time scale, also precision, is the influence of periods of misalignment that have not been programmatically identified. An ideal solution would be to have an independent method of

monitoring mirror alignment during atmospheric measurements (e.g. installation of a small sighting telescope to measure the alignment coupled with an automatic fiber optic alignment system). With the existing data set from OHP extending back two decades, we unfortunately cannot retrospectively address such a hardware goal, but there may be opportunities in future to look into the effects of choosing different criteria to identify periods of misalignment.

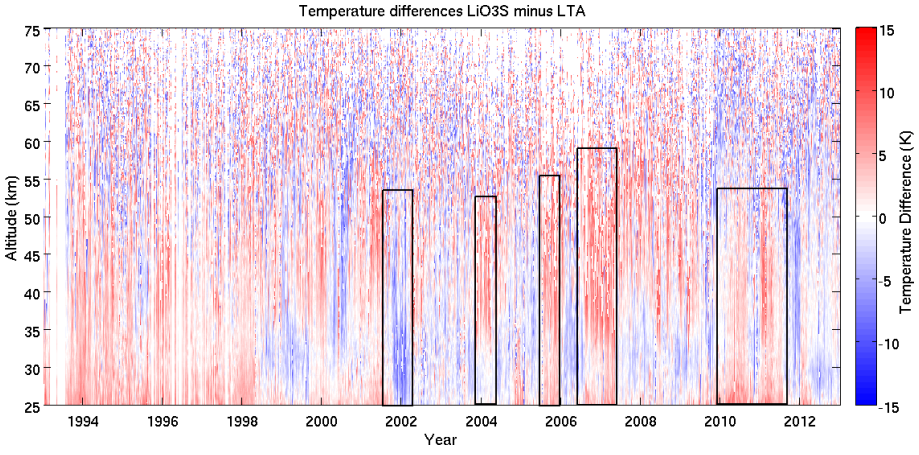

Figure 13: Temperature differences between LTA and LiO_3S OHP lidars for a 20 year period between 1993 and 2013. There are 1496 nights of comparison in this plot. Red indicates that LiO_3S was warmer than LTA and blue that it was colder. The black boxes highlight periods where the two lidars were out of alignment with respect to each other.

Figure 14 shows four curves depicting the average temperature differences as a function of altitude and year. The red curve is the average temperature difference between 65 km and 75 km with an average standard deviation of 6.6 K; the green curve is the average temperature difference between 55 km and 65 km with an average standard deviation of 4.5 K; the blue curve is the average temperature difference between 45 km and 55 km with an average standard deviation of 2.7 K; and the magenta curve is the average temperature difference between 35 km and 45 km with an average standard deviation of 1.6 K. A 30 day averaging window is applied to each of the four curves.

For reference, a typical LTA temperature profile with an effective vertical resolution of 2 km has an uncertainty due to statistical error of 0.2 K at 40 km; 0.4 K at 50 km; 0.6 K at 60 km; 0.7 K at 70 km; 1.8 K at 80 km; and 6 K at 90 km. For reference, a typical LiO3S temperature profile with an effective vertical resolution of 2 km has an uncertainty due to statistical error of 0.3 K at 40 km; 0.5 K at 50 km; 1.0 K at 60 km; 2.7 K at 70 km; and 10 K at 80 km.

Examining the time evolution of the average temperature differences between LTA and LiO_3S at four altitude levels gives us confidence that both measurements are stable in both time and altitude. ~~Without excluding~~ Using all data, including misaligned periods (example: winter 2006-2007 in Fig. 13 and Fig. 14) none of the lidar temperature differences are ~~not significant as a function of~~

~~altitude or year at the 2 sigma level.~~ significant at the 2-sigma level, although certain periods do have temperature differences which are detectable at the 1-sigma level. This can be seen where the blue shaded region (2005 - 2008) and the magenta shaded region (in 2007) are entirely above the zero line. If the misaligned periods are disregarded, no temperature differences are significant, even at the 1-sigma level. Therefore, we conclude that the results from the lidars, when well-aligned, are stable in time, over the 20-year period studied.

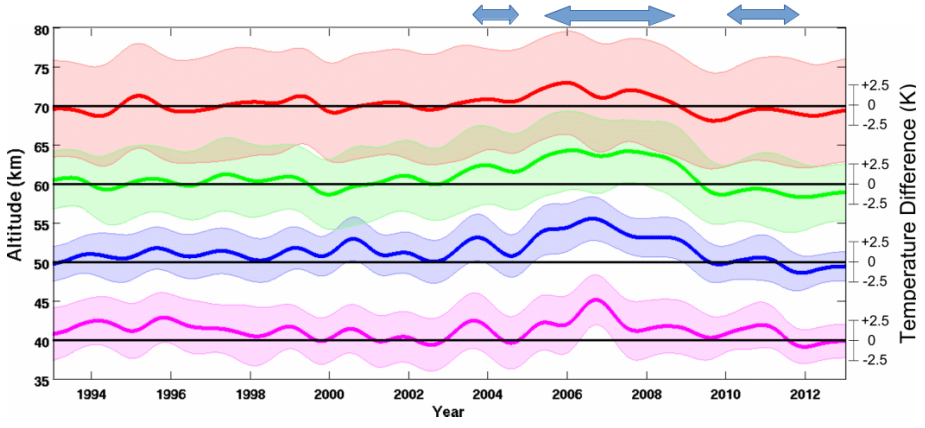

Figure 14: Average temperature differences between LTA and LiO_3S OHP lidars for a 20 year period between 1993 and 2013 at four altitude levels: 65-75 km (red), 55-65 km (green), 45-55 km (blue), and 35-45 km (magenta). Shaded uncertainties are shown at 1 sigma for clarity and the black lines are zero temperature difference displaced to 40, 50, 60 and 70 km. All measurements, including periods of lidar misalignment, are included in this plot. The apparent anomalies (~~e.g. between 2005 and 2009~~blue arrows) occur only during times where the lidars were often misaligned, as indicated in Fig. 13.

After removing comparisons between mis-aligned instruments we can calculate the ensemble median difference between the two systems. The ensemble median difference in Fig. 15 shows very good agreement between the two co-located lidar instruments. The temperatures produced by LTA and LiO_3S are statistically equal above 45 km for the 20 year period between 1993 and 2013. There is a small –0.6 K systematic difference which reaches a maximum near 40 km. We believe this slight cold bias is due to small differences in the signal melding technique between the high and low gain channels in both systems. On a typical night, the LTA low gain channel starts to significantly contribute to the combined signal near 50 km. If the photon count rate in the low gain channel is too large at these altitudes (due to residual noise contributions or from a slight misalignment with the high channel) the counts will be artificially higher than expected, resulting in a ~~colder~~ lower temperature. The converse holds true when the low gain channel is misaligned in the opposite sense, resulting in a slight warming due to underestimation of the counts.

The effect of these small temperature perturbations is so small that they can't be seen in single nightly temperature comparisons and were not detected before this study. It is important to note that the 2σ distribution about our ensemble at 40 km has a magnitude of approximately 0.45 K while the statistical error for a single night of lidar measurements near 40 km at 300 m vertical resolution can be on the order of 2 K. Detecting and resolving this small disagreement will be extremely challenging and will not be accomplished in this work.

Given that the primary interest of this work is the upper middle atmosphere (nominally above 50 km), we will focus on the upper portions of Fig. 15 where the two lidars are in statistically perfect agreement. To our knowledge, this is the first ever ~~long-term~~ long term study of the temperatures produced by co-located temperature lidars operating at 532 nm and 355 nm. The excellent agreement between these two independent measurements gives us confidence that A) there is no vertical misalignment between the lidars, B) there are no unaccounted for optical transmission effects which influence our temperatures, C) the lidar measurements are ~~accurate~~ reasonable and reproducible, D) we can now proceed with some confidence that our ground based lidar measurements can be useful as a calibration source for the space based satellite measurements.

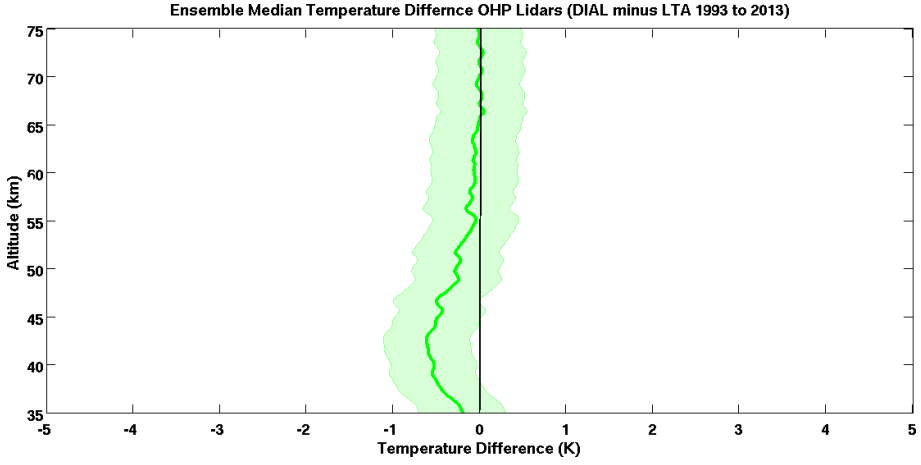

Figure 15: Ensemble of median temperature differences between LTA and LiO_3S based on temperature measurements between 1993 and 2013. Shaded error is the two sigma distribution about the ensemble.

6 Summary and Discussion

6.1 Changes to Lidar Temperature Algorithm

In this work we have attempted to minimize the systematic temperature bias at the top of the lidar temperature retrieval which has been noted previously by several studies cited in the introduction. We have done this by clearly and carefully outlining a rigorous, and complete algorithm for the

calculation of lidar temperatures in the UMLT. We have presented techniques for the detection of signal contamination, the selection of the best data for inclusion in the calculation, criteria for where to initialized the inversion when assuming an a priori pressure at the top of the atmosphere, and have demonstrated the benefit of photomultiplier cooling and narrow band pass filters to reduce lidar backgrounds.

After applying our techniques we have seen a systematic cooling of the high altitude lidar temperatures which brings them into better agreement with the temperatures measured by both MLS and SABER (Fig. 11). It is also important to note the large variance associated with these ensemble differences can partially be attributed to the lack of control exerted on the error contribution from the choice of a priori initial pressure for lidar data and a priori contribution and non-LTE effects for satellite data. Part of the difference may also be due to altitude offsets and coarse vertical resolution.

Having applied these new data filtering techniques we have produced an improved lidar temperature data set which is exploited in the companion paper (Wing et al., 2018b) in an effort to validate satellite temperatures.

6.2 OHP Lidar 20 Year Comparison

We have conducted the first ever decadal temperature inter comparison between a co-located 532 nm Rayleigh lidar and an ozone DIAL system calculating temperatures from a 355 nm line. We have shown that:

1) Rayleigh lidar temperatures calculated from ozone DIAL non-absorbing 355 nm line are statistically equal to temperatures from a traditional 532 nm Rayleigh temperature lidar over a large altitude range. This finding is of particular interest for the NDACC lidar temperature database as temperatures from ozone lidars may also be available for validation and inclusion.

2) Further theoretical work must be done on algorithms for melding data from high and low gain photon counting channels. The current techniques produce statistically identical nightly temperature profiles however, a -0.6 K bias near 40 km becomes apparent when multiple years of data are compared. It is doubtful that current data processing techniques can be easily adapted to address this problem. However, an iterative, cost minimizing, Bayesian approach such as the one proposed by (Sica and Haefele, 2015) would be able to produce a single melded temperature profile with the accompanying averaging kernels and an estimate of the error due to the photon count melding. As a lidar development note, Fig. 13 demonstrates the need move towards the use of automated nightly alignment of lidar system optics. Manual alignment by operators appears to lack consistency over the time frame of multiple decades.

3) The two independent lidars show no evidence of significant instrument drift over a 20 year period. This means that ground based lidars are the ideal choice of instrument for detecting small calibration drifts in satellite remote measurements over long time scales. We rely on this finding to

justify the use of lidars as a reference data set for satellite validation in the companion paper Wing et al. (2018b).

4) There is no evidence of a relative vertical offset between the two independently calibrated lidar systems which would be seen as an 'S' shaped temperature bias in Fig. 15 due to the sign change in temperature vertical gradient at the stratopause (Leblanc et al., 1998a). Based on personal communication, recent July-August 2017 and March 2018 NDACC Ozone validation campaign at OHP (LAVANDE) revealed no vertical shifts between either OHP lidar and the NASA STROZ mobile validation lidar (McGee et al., 1995).

Acknowledgements. The data used in this paper were obtained as part of the Network for the Detection of Atmospheric Composition Change (NDACC) and are publicly available (see http://www.ndacc.org, http://~~eds-espri~~cdsespri.ipsl.fr/NDACC) as well as from the SABER (see ftp://saber.gats-inc.com) and MLS (see https://mls.jpl.nasa.gov) data centres for ~~the public accessvia their websites~~public access. This work is supported by the ~~project~~ Atmospheric dynamics Research InfraStructure Project (ARISE 2) ~~funded by funded by~~ which is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. ~~653980~~ 653980. French NDACC activities are supported by Institut National des Sciences de l~~'~~'Univers/Centre National de la Recherche Scientifique (INSU/CNRS), Université de Versailles Saint-Quentin-en-Yvelines (UVSQ), and Centre National d~~'~~'Études Spatiales (CNES). The authors would also like to thank the C. Wing for graphics support, and the technicians at La Station Géophysique Gérard Mégie at OHP.

References

NDACC Lidar, http://ndacc-lidar.org/.

Alpers, M., Eixmann, R., Fricke-Begemann, C., Gerding, M., and Höffner, J.: Temperature lidar measurements from 1 to 105 km altitude using resonance, Rayleigh, and Rotational Raman scattering, Atmospheric Chemistry and Physics, 4, 793–800, doi:10.5194/acp-4-793-2004, https://www.atmos-chem-phys.net/4/793/2004/, 2004.

Apruzese, J. P., Strobel, D. F., and Schoeberl, M. R.: Parameterization of IR cooling in a Middle Atmosphere Dynamics Model: 2. Non-LTE radiative transfer and the globally averaged temperature of the mesosphere and lower thermosphere, Journal of Geophysical Research: Atmospheres, 89, 4917–4926, doi:10.1029/JD089iD03p04917, http://dx.doi.org/10.1029/JD089iD03p04917, 1984.

Argall, P.: Upper altitude limit for Rayleigh lidar, Annales Geophysicae, 25, 19–25, doi:10.5194/angeo-25-19-2007, 2007.

Donovan, D. P., Whiteway, J. A., and Carswell, A. I.: Correction for nonlinear photon-counting effects in lidar systems, Appl. Opt., 32, 6742–6753, doi:10.1364/AO.32.006742, http://ao.osa.org/abstract.cfm?URI=ao-32-33-6742, 1993.

Dou, X., Li, T., Xu, J., Liu, H.-L., Xue, X., Wang, S., Leblanc, T., McDermid, I. S., Hauchecorne, A., Keckhut, P., Bencherif, H., Heinselman, C., Steinbrecht, W., Mlynczak, M. G., and Russell, J. M.: Seasonal oscillations of middle atmosphere temperature observed by Rayleigh lidars and their comparisons with TIMED/SABER observations, Journal of Geophysical Research: Atmospheres, 114, n/a–n/a, doi:10.1029/2008JD011654, http://dx.doi.org/10.1029/2008JD011654, d20103, 2009.

García-Comas, M., Funke, B., Gardini, A., López-Puertas, M., Jurado-Navarro, A., von Clarmann, T., Stiller, G., Kiefer, M., Boone, C. D., Leblanc, T., Marshall, B. T., Schwartz, M. J., and Sheese, P. E.: MIPAS temperature from the stratosphere to the lower thermosphere: Comparison of vM21 with ACE-FTS, MLS, OSIRIS, SABER, SOFIE and lidar measurements, Atmospheric Measurement Techniques, 7, 3633–3651, doi:10.5194/amt-7-3633-2014, https://www.atmos-meas-tech.net/7/3633/2014/, 2014.

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, pp. 57–67, doi:10.1039/B205880D, 2003.

Gross, M. R., McGee, T. J., Ferrare, R. A., Singh, U. N., and Kimvilakani, P.: Temperature measurements made with a combined Rayleigh–Mie and Raman lidar, Appl. Opt., 36, 5987–5995, doi:10.1364/AO.36.005987, http://ao.osa.org/abstract.cfm?URI=ao-36-24-5987, 1997.

Hauchecorne, A. and Chanin, M.-L.: Density and temperature profiles obtained by lidar between 35 and 70 km, Geophysical Research Letters, 7, 565–568, doi:10.1029/GL007i008p00565, http://dx.doi.org/10.1029/GL007i008p00565, 1980.

Keckhut, P., Hauchecorne, A., and Chanin, M.: A critical review of the database acquired for the long-term surveillance of the middle atmosphere by the French Rayleigh lidars, Journal of Atmospheric and Oceanic Technology, 10, doi:10.1175/1520-0426(1993)010<0850:ACROTD>2.0.CO;2, 1993.

Keckhut, P., McDermid, S., Swart, D., McGee, T., Godin-Beekmann, S., Adriani, A., Barnes, J., Baray, J.-L., Bencherif, H., Claude, H., di Sarra, A. G., Fiocco, G., Hansen, G., Hauchecorne, A., Leblanc, T., Lee, C. H., Pal, S., Megie, G., Nakane, H., Neuber, R., Steinbrecht, W., and Thayer, J.: Review of ozone and temperature

lidar validations performed within the framework of the Network for the Detection of Stratospheric Change, J. Environ. Monit., 6, 721–733, doi:10.1039/B404256E, http://dx.doi.org/10.1039/B404256E, 2004.

Khanna, J., Sica, R. J., and McElroy, C. T.: Atmospheric temperature retrievals from lidar measurements using techniques of non-linear mathematical inversion, AGU Fall Meeting Abstracts, 2011.

Khanna, J., Bandoro, J., Sica, R. J., and McElroy, C. T.: New technique for retrieval of atmospheric temperature profiles from Rayleigh-scatter lidar measurements using nonlinear inversion, Appl. Opt., 51, 7945–7952, doi:10.1364/AO.51.007945, http://ao.osa.org/abstract.cfm?URI=ao-51-33-7945, 2012.

Khaykin, S. M., Godin-Beekmann, S., Keckhut, P., Hauchecorne, A., Jumelet, J., Vernier, J.-P., Bourassa, A., Degenstein, D. A., Rieger, L. A., Bingen, C., Vanhellemont, F., Robert, C., DeLand, M., and Bhartia, P. K.: Variability and evolution of the midlatitude stratospheric aerosol budget from 22 years of ground-based lidar and satellite observations, Atmospheric Chemistry and Physics, 17, 1829–1845, doi:10.5194/acp-17-1829-2017, https://www.atmos-chem-phys.net/17/1829/2017/, 2017.

Kumar, V. S., Rao, P. B., and Krishnaiah, M.: Lidar measurements of stratosphere-mesosphere thermal structure at a low latitude: Comparison with satellite data and models, Journal of Geophysical Research: Atmospheres, 108, doi:10.1029/2002JD003029, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JD003029, 2003.

Leblanc, T., McDermid, I. S., Hauchecorne, A., and Keckhut, P.: Evaluation of optimization of lidar temperature analysis algorithms using simulated data, Journal of Geophysical Research: Atmospheres, 103, 6177–6187, doi:10.1029/97JD03494, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/97JD03494, 1998a.

Leblanc, T., McDermid, I. S., Keckhut, P., Hauchecorne, A., She, C. Y., and Krueger, D. A.: Temperature climatology of the middle atmosphere from long-term lidar measurements at middle and low latitudes, Journal of Geophysical Research: Atmospheres, 103, 17 191–17 204, doi:10.1029/98JD01347, http://dx.doi.org/10.1029/98JD01347, 1998b.

Mann, H. B. and Whitney, D. R.: On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, Ann. Math. Statist., 18, 50–60, doi:10.1214/aoms/1177730491, https://doi.org/10.1214/aoms/1177730491, 1947.

McGee, T. J., Ferrare, R. A., Whiteman, D. N., Butler, J. J., Burris, J. F., and Owens, M. A.: Lidar measurements of stratospheric ozone during the STOIC campaign, Journal of Geophysical Research: Atmospheres, 100, 9255–9262, doi:10.1029/94JD02390, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JD02390, 1995.

Picone, J. M., Hedin, A. E., Drob, D. P., and Aikin, A. C.: NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues, Journal of Geophysical Research: Space Physics, 107, SIA 15–1–SIA 15–16, doi:10.1029/2002JA009430, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JA009430, 2002.

Remsberg, E. E., Marshall, B. T., Garcia-Comas, M., Krueger, D., Lingenfelser, G. S., Martin-Torres, J., Mlynczak, M. G., Russell, J. M., Smith, A. K., Zhao, Y., Brown, C., Gordley, L. L., Lopez-Gonzalez, M. J., Lopez-Puertas, M., She, C.-Y., Taylor, M. J., and Thompson, R. E.: Assessment of the quality of the Version 1.07 temperature-versus-pressure profiles of the middle atmosphere from TIMED/SABER, Journal of Geophysical Research: Atmospheres, 113, n/a–n/a, doi:10.1029/2008JD010013, http://dx.doi.org/10.1029/2008JD010013, d17101, 2008.

Sica, R. and Haefele, A.: Retrieval of temperature from a multiple-channel Rayleigh-scatter lidar using an optimal estimation method, Appl. Opt., 54, 1872–1889, doi:10.1364/AO.54.001872, http://ao.osa.org/abstract.cfm?URI=ao-54-8-1872, 2015.

Singh, U. N., Keckhut, P., McGee, T. J., Gross, M. R., Hauchecorne, A., Fishbein, E. F., Waters, J. W., Gille, J. C., Roche, A. E., and Russell, J. M.: Stratospheric temperature measurements by two collocated NDSC lidars during UARS validation campaign, Journal of Geophysical Research: Atmospheres, 101, 10 287–10 297, doi:10.1029/96JD00516, http://dx.doi.org/10.1029/96JD00516, 1996.

Sivakumar, V., Prasanth, V. P., Kishore, P., Benchérif, H., and Keckhut, P.: Rayleigh LIDAR and satellite (HALOE, SABER, CHAMP and COSMIC) measurements of stratosphere-mesosphere temperature over a southern sub-tropical site, Reunion (20.8° S; 55.5° E): climatology and comparison study, Annales Geophysicae, 29, 649–662, doi:10.5194/angeo-29-649-2011, https://hal.archives-ouvertes.fr/hal-00586264, 2011.

Taori, A., Jayaraman, A., Raghunath, K., and Kamalakar, V.: A new method to derive middle atmospheric temperature profiles using a combination of Rayleigh lidar and O_2 airglow temperatures measurements, Annales Geophysicae, 30, 27–32, doi:10.5194/angeo-30-27-2012, 2012.

Taori, A., Kamalakar, V., Raghunath, K., Rao, S., and Russell, J.: Simultaneous Rayleigh lidar and airglow measurements of middle atmospheric waves over low latitudes in India, Journal of Atmospheric and Solar-Terrestrial Physics, 78-79, 62–69, doi:10.1016/j.jastp.2011.06.012, 2012.

Tukey, J. W.: Comparing Individual Means in the Analysis of Variance, Biometrics, 5, 99–114, 1949.

Wing, R., Hauchecorne, A., Godin-Beekman, S., Khaykin, S., and McCullough, E. M.: Lidar temperature series in the middle atmosphere as a reference data set. Part B: Assessment of temperature observations from MLS/Aura and SABER/TIMED satellites, Atmospheric Measurement Techniques, Submitted, 2018b.

Yue, C., Yang, G., Wang, J., Guan, S., Du, L., Cheng, X., and Yang, Y.: Lidar observations of the middle atmospheric thermal structure over north China and comparisons with TIMED/SABER, Journal of Atmospheric and Solar-Terrestrial Physics, 120, 80–87, doi:10.1016/j.jastp.2014.08.017, 2014.