Dear Referee,
Thank you very much for your helpful comments and suggestions. I have attempted to address each of your concerns to the best of my ability. If you would like me to implement further changes or iterations on a point please let me know. I appreciate your efforts to help me improve this paper.

The questions you raise about error estimation after removal of signal induced noise contributions are in my opinion very critical. I think that as a lidar community we really need to push the envelope on our data retrieval techniques and error estimates - particularly if we want to do new work in the Mesosphere/Thermosphere. The work in this paper is not perfect but it is an improvement to the commonly used (Hauchecorne/Chanin 1980) lidar temperature inversion. It's my belief that as a community we should continue investigating improvements to our techniques. A fully Bayesian approach such as the Optimal Estimation Technique presented by (Sica/Haefele 2015) might be a profitable endeavor. As an added benefit a Bayesian Technique produces full averaging kernels which would make lidar data much more attractive for assimilation to people in the satellite and reanalysis communities .
****

Response *Lidar temperature series in the middle atmosphere as a reference data set. Part A: Improved retrievals and a 20 year cross-validation of two co-located French lidars*: Referee #1

---

1) Page 6: Lidar equation (1) has dimension mismatch on the left and right hand sides. The first part on the right side has a dimension of energy, but the left side N(z) is claimed to be count rate per time integration per altitude bin. This equation is not acceptable for publication. Furthermore, beta (β) is commonly used to represent volume backscatter coefficient, not backscattering cross-section, as the cross-section symbol is usually sigma (σ). Authors are suggested to consult with a commonly referenced class lecture at the following link, and use the more commonly accepted lidar equations and symbols.

Good catch thank you. I've changed divided the right hand side by the photon energy hc/lambda and changed β_cross to σ_cross

---

2) Page 9: Please provide a reference to Turkey Quartile test, as this isn't a common practice for most lidar people. BTW, it should be "when the signal to noise ratio approaches 1".

Cited Tukey(1949)
"signal to noise" changed to "signal to noise ratio"

---

3) Page 12: Please provide a reference to the "one sided non-parametric MannWhitney-Wilcoxon rank-sum test" as it isn't common for the lidar field. BTW, what does "a scan" mean in Figure 6? Did you mean one profile?

Cited Mann and Whitney (1947)

I've always called a single level_0 or level_1 photon count time series a 'scan' and used the word 'profile' for level_2 things like density, pressure, temperature. Reviewer #2 and Reviewer #3 made the same point so perhaps it's a personal idiosyncrasy. In any case, I've changed all occurrences of 'scan' to 'profile'

---

4) Page 14, how are Si and Ni determined? Please provide a bit more details. Do you do this (equation (2)) for every altitude bin?

Line 268 - 272
The noise is always evaluated between 120 km and 155 km and the altitude range for the evaluating the signal is defined as the scale height below the point where the signal to noise equals one in the density profile. Each individual profile has a value representing the signal, S_{i}, and a noise, N_{i}. The profile values are compared to the nightly sum of the signal, S_{sum} and the nightly sum of the noise, N_{sum}.
Changed to:
***T*he 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 noise, 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 ratio equals 1. Then we calculate the altitude that is one density scale height (~8 km) below this point. The lidar range bins which correspond to this altitude range are then summed to yield S_{sum}. A similar calculation, using the same range bins as in the nightly average calculation, is done to determine the signal of single profile, S_{i}.**
---

5) Page 16, notations are needed for equation (3).

N, τ, and Δt are described in lines 309-310. We have now replaced the definition of N with separate definitions for $N_{counted}$ and $N_{received}$, as they appear in Equation (3):
Replacement text:

The background theory and derivation of Eq. (3) is well described by (Donovan et al., 1993), where $N_{received}$ 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.

---

6) Page 16, after the quadratic fit to the background, how do you handle such background and data? Did you mean to subtract the quadratic fitted background from the raw data? In this case, how do you handle the noise term in calculating SNR? Are photon counts still obey Poisson distribution? Please clarify in the manuscript.

Yes, in the case of a quadratic background I subtract the quadratic function from the entire photon counts profile in exactly the same way I would treat a constant or linear background.

As you correctly point out, as soon as there is signal induced noise the profile is no longer Poisson 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. The Total counts are some combination of 'Real counts' and 'Contamination counts ' (T = R + C) with a common shot noise dT = 1/sqrt(T) with some contribution dC = 1/sqrt(C) coming from the Signal Induced Noise portion and dR = 1/sqrt(R) representing the noise from all other sources. When I'm using the linear or quadratic backgrounds I am making an assumption that I'm completely removing the signal induced noise, C and I no longer have to add dR and dC in quadrature. I'm approximating dN ~= dR and that the photon count profiles are now approximately Poisson.

On page 16 line 334, we have added a new sentence to the manuscript:
"...as our estimate of signal induced noise. The best background function is subtracted from the raw photon counts profile."

On page 16 line 341 we have added a new sentence to clarify about SNR:
"...than the simple quadratic approximation. 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."

---

7) Page 20, Figure 11, it's necessary to point out in the manuscript that satellite data aren't the real references as various satellites have their own calibration issues. Rayleigh temperatures around 90 km should be compared with ground-based resonance Doppler or Boltzmann lidar temperatures as these resonance lidars have much better signal to noise ratios at these altitudes.

Line 414 inserted text: **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 85 km (Alpers, 2004).**

---

8) Page 22-23, what do you mean by "misaligned"? A lidar beam was misaligned relative to its own receiver's field of view, or else? How were two lidars misaligned? Authors' writings here are confusing.

In both lidar systems the high gain Rayleigh channel has 4 mirrors, each of which needs to be aligned independently with respect to the laser in the sky and also the fibre optic with respect to the primary focus of the mirror. In LTA the low gain channel is a single independent mirror. So a total of 9 mirrors need to be aligned every night to make Rayleigh measurements.

Line 456 inserted text: **Internal misalignments happen when one or more of the five mirrors in LTA or four mirrors in LiO3S is not properly aligned with the laser or the fibre optic is not centered on the focal point of the mirror.**

Minor comments on English writing: As this is a very long paper, I strongly encourage authors go over the manuscript carefully to correct grammar and typo issues. For example, on page 24, near line 495, it should be "to initialize the inversion", not "initialized". The paper title doesn't have good English grammar, for which I suggest to change "a 20 year cross-validation" to "a 20-year cross validation

I will have an anglophone colleague read over the paper for grammatical errors.

Dear Referee,
Thank you very much for your helpful comments and suggestions. I have attempted to address each of your concerns to the best of my ability. If you would like me to implement further changes or iterations on a point please let me know. I appreciate your efforts to help me improve this paper.

Thank you for raising the point about aerosols. It forced me to do quite a bit of reading on the subject. I think this is a very important question for lidars that measure both Rayleigh temperatures and Mie scatter from aerosols. (Hauchecorne and Chanin 1980) is given very often in lidar papers to justify the 30 km assumption of a clean atmosphere. I think this is valid most of the time but I can see problems in some of the old temperature profiles after the 1982 El Chichon eruption. It really underscores the need for a Raman channel.

Thanks also for helping me to clarify my language when describing the statistics and my discussion of LTE. I think this work is more clear after removing some of my idiosyncratic language.

****

Response *Lidar temperature series in the middle atmosphere as a reference data set. Part A: Improved retrievals and a 20 year cross-validation of two co-located French lidars*: Referee #2

Specific comments:
P6-7L140-146: In this section saturation is neglected, but in Section 3.5.1 the correction is described and in Figures 10, 13, 14, 15 stratospheric data is shown. I suggest not to neglect saturation throughout the manuscript.

Text added line 146: **A correction for saturation in the lower stratosphere is described in Sect. \ref{Deadtime Correction}**

P7L147-150: I have not found any number on the integration time for the temperature profiles used here. I assume that it is long enough to at least partly overcome the issue of non-LTE. If not, the potential errors by assuming LTE need to be described. The statement "unable to relax" would not be sufficient, if differences between data sets are examined and "standards" are defined.

Changed
In this work we are unable to relax this assumption.
To

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.

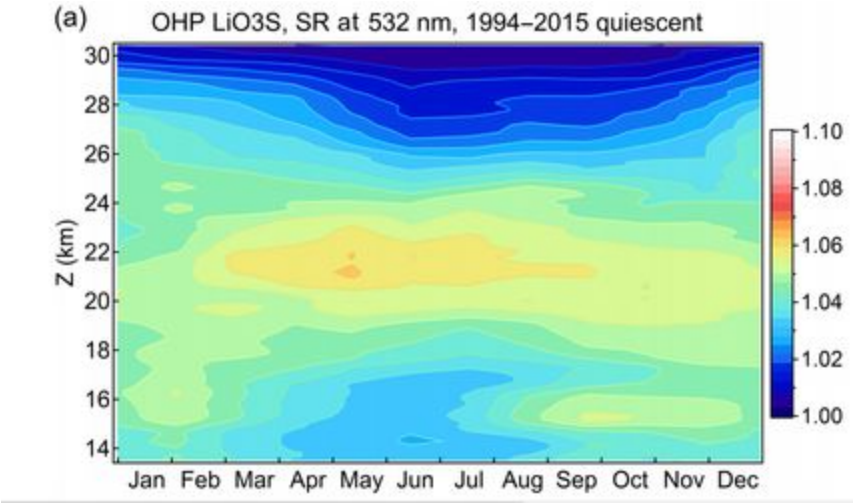
P7L151: This assumption is problematic as there are different studies showing aerosols up to at least 35 km.

You are correct. In the presence of significant aerosol loading (volcanos and fires) we can see a cold bias in our temperatures above 30 km. However, in times when the aerosol loading is less pronounced the Rayleigh lidar temperature cold bias is relatively small and can generally be corrected by using the Raman lidar channel. I've weakened my assumption and provided two justifications for my assumptions.

Changed assumption 4 and citation to read:

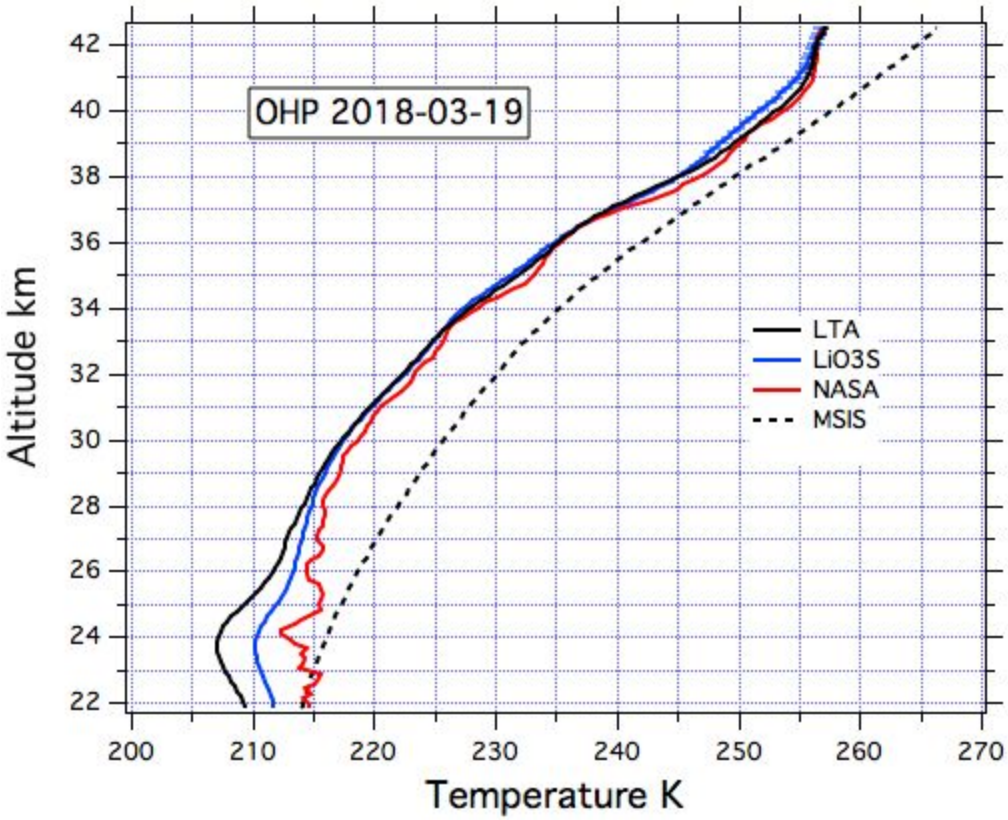
Fourth, we assume that the atmosphere at mid-latitudes is generally free of aerosols above 30 km when there are no active volcanic or fire events (Hauchecorne and Chanin, 1980). During less severe background aerosol conditions (aerosol scattering ratio < 1.02), (Gross et al. 1997) suggests lidar temperature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

(Khaykin et al. 2017) published a 22 year stratospheric aerosol climatology for OHP
https://www.atmos-chem-phys.net/17/1829/2017/acp-17-1829-2017.html
They found that during quiescent times (no major eruptions or fires) that the scattering ratio was near one at 30 km



Here is a plot of temperature from our recent ozone blind NDACC validation campaign at OHP. The two OHP temperatures from LTA (532 nm) and LiO3S (355 nm) are in good agreement with the NASA mobile validation lidar (355 nm) above 30 km. N.B. that this particular LTA profile doesn't have a 607 nm Raman correction as the Raman channel experienced some difficulties during this period. But even without a Raman correction

the temperatures converge by 30 km.



P8Figure3: The count rates are comparatively high and saturation is likely to become a problem (see above). I assume a typo in the vertical resolution of 7.5 m.

Typo corrected to 75 m

P9Sec3.3.1: Please explain the (potential) origin of these spikes. Fig. 4 shows that they easily reach 10-100 counts, i.e. they are quite substantial. I wonder whether it would be useful to work on the origin of the spikes instead of only removing the resulting counts. Do you remove only the spiky bins or the whole profile? I guess, the first would result in too low counts rates in the altitude of the spikes after integration of several profiles. Please make clear.

The spikes can be expected to occur in any Poisson counting process, could be induced by some 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. On any given night there are three lidars, several ozone instruments, OH spectrometers, and probably several other instruments at various times in simultaneous operation at the Geophysical Observatory building at OHP.

I would genuinely love to track down every little bug and glitch in the lidar system but unfortunately, I'm in my 3rd year of a 3 year PhD in Paris. The observatory is in Haute Provence in the south of France and only CNRS technicians are permitted to make changes to the experiment. I don't have the time or access required to address this problem at the experimental level. Additionally, this project takes advantage of measurements taken as many as 20 years ago. While increasing the quality of all future lidar data is indeed a positive goal for the lidar group, all existing data requires some software treatment in any case. Therefore, I've done my best to address the spikes with software - and this approach appears to be both adequate and successful.

Individual spiky bins are removed from the profiles. When averaging multiple profiles together, it is possible to do so in a manner which accounts for bins with "not a number" (i.e. spiky bins whose data is totally removed) separately from bins which have "zero counts" (bins which have zero photons, but which are still valid data). For example, the nanmean function in Matlab. The overall SNR may decrease slightly at the altitude of the spiky bin (since we're adding bins from fewer profiles into the average, which is equivalent to taking fewer measurements at that altitude), but the averaged count rate will not be skewed too low.

Caption Fig04 has been clarified to say "Tukey Quartile spike identification based on the signal difference between consecutive lidar time bins for short integration lidar returns. An entire night of lidar profiles is over-plotted in the stack plot. The black line is the 2 sigma limit and points above this line are removed.".

We have added a few sentences to page 9 line 189:
"...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."

P10L207: Please explain "downstream counting rate".

“downstream counting rate” changed to “counting rate in bins subsequent to the TES burst”

P10L207-219: Is the Kurtosis test always only done on the first 100 bins? If yes, how do you detect TES that may appear above that range? If no, how does the exponentially decreasing (i.e. non-Gaussian) signal influence the test

Changed the Fig5 caption to read, with 2 extra sentences for clarity:

Figure 5: Upper panel is a surface plot of lidar returns as a function of time bin and scan 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 2\sigma limit on the skew of the population. Points above the limit are excluded.

P11L230-233: I do not see four groups of signals. Essentially it is either high background and low signal or low background and high signal. Please explain. Why does the number of groups depend on the statistics “the authors choose to use”? Which statistics? I am generally missing an explanation of the strategy or method. Why not simply defining a signal-noise-limit to separate good profiles and noisy profiles?

This is what I’m trying to show. You say 2 groups of signals and that is very reasonable. I say that the median of the first period of low signal and high noise is significantly different from the second and they are in fact two different populations. Whether that depends on laser power, sky transmission, background or something else - I don’t know. That’s why I’m trying to develop an automated tool for data quality assessment.

We have changed the sentences on line 229-233 to read:
However, when we look at the panel representing the signal, 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 (profiles 1-23, profiles 24-35, profiles 36 - 48 and profiles 49 - 92) . However, whether these four groups of signals should be treated in analysis as two, three, or four distinct populations is open to interpretation. Therefore, we seek an objective programmatic solution for identifying bad scans.

In essence I have defined a signal to noise cut off, as you suggest, with equation (2), in section 3.3.4 about "Good Scans". This is indeed useful for identifying and rejecting scans which contribute more noise than information to the nightly average at a given altitude. Therefore, it is the final quality control step used.

However, before we get to that point, we address in section 3.3.3 about "Bad Scans" the separate question of rejecting scans which do not conform to the general population, and are therefore outliers. We point out on line 246 that there can be multiple signal to noise population medians during the course of the night, which makes setting a constant minimum SNR criterion for the whole night inappropriate as a sole means of judging good vs. bad scans. The one sided non-parametric Mann-Whitney-Wilcoxon rank-sum test, is the solution to the subjective interpretation problems presented by Figure 6. It is not the final step in quality control - but it is a useful intermediate step.

P11L236 and Fig.6: The green line is not only a running average but contains some offset. Please explain.

Fig06 Caption has been clarified:

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 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.

P11L238-243: It remains open how the blue line is derived. It is the result of a blackbox-software and the results are discarded by the authors. I suggest removing this section and the blue line in Fig. 6.

I've had several discussions at NDACC lidar meetings and at the last IRLC about using machine learning and using MatLab's neural network toolbox to estimate lidar profile backgrounds. I think that is plot will be interesting to several people. Using machine learning to process lidar data is to my knowledge a wide open field to be explored. Your point is well taken about blackbox-software.

P12L249-256 and Fig. 7: Please explain in more detail how this test works. Please use for Figure 7 the same data set as for Figure 6. Otherwise the reader can hardly comprehend the method. If I understand the test correctly, it only removes the worst profiles of a particular night. If the whole night has a bad signal, the data will not be removed. Correct? In line 253 you do "not exclude" the bad profiles, but in line 256 13 bad profiles are identified (how??) and "discarded". I do see a contradiction between the two sentences.

Thanks for catching the mistake between Fig 6 and Fig 7. Figure 6 has been remade over the same range as Figure 7.

Cited Mann and Whitney (1947) for details of the statistic
This Mann-Whitney-Wilcoxon test only removes profiles which are "very different" compared to others nearby on the same night, and you are correct that it will not remove all profiles if the whole night has bad signal. The latter is not its purpose.

The Mann-Whitney-Wilcoxon test has two ways in which scans are determined to be outliers: (a) SNR is too low compared to that of nearby scans and (b) SNR is too high compared to that of nearby scans. To apply this test to lidar, we want to reject from our analysis scans which fail for reason (a; scan is low quality), but not those which fail for reason (b; scan is high quality). Therefore the last sentences are not contradictory: We in fact have not rejected any scans on the basis of high quality ("failure reason (b)"), but have rejected 13 scans on the basis of low quality ("failure reason (a)").

To address data for which the whole night has very bad signal:
First, OHP operators monitor the lidar measurements as they are made throughout the night, and attempt to either correct the issue or stop the measurement if the sky clouds in. The first 'quality filter' is the judgement of the OHP technicians.

Second, I have an arbitrary condition that an LTA temperature profile must reach 80 km in 4 hours integration at 2 km effective vertical resolution. This catches the few remaining 'bad nights' where the lidar acquisition was too short or the observing conditions were too cloudy. The OHP operators are very good at maintaining data integrity.

P13Sec3.3.4: I am sorry, but I do not understand this section. Why not simply considering only data up to altitude z by defining a criterion like SNR(z) > Threshold ?

Because there is no flexibility in that kind of SNR definition. I used 5,676 nights of lidar data from two instruments over 20 years. I needed something that could be adaptable to changing signal levels as transmitter power changes over decades. As well I wanted to use the data as efficiently as possible. On a clear night I can get temperatures up to 90 km but there's no point in wasting a night of data with light cirrus where I only get temperatures up to 80 km.

P14L284-294: The noise reduction is interesting. To allow the reader evaluating the technical progress, it would be helpful to learn a) whether these are the most important changes in background count rate for the whole 20 y data set and b) what are the benefits for the temperature calculation if the background is reduced to 1/100 (e.g., range extended by .. km).

I agree. This is an area of lidar science that is waiting to be developed with the aid of modern computers and new models. As mentioned previously I have been in communication with colleagues looking to use Bayesian statistics and machine learning to look at lidar backgrounds and noise. This paper is already very long and I'm in the 3rd year of a 3 year PhD. Perhaps this work could be done in a different article?

P16Sec3.5.3: It would be helpful for the interpretation of the results (also of the companion paper) to have a quantitative description of the influence of a wrong background shape on the temperature calculation. Additionally, the SIN of the low channel in Fig. 9 is extremely high and the choice of the shape of the SIN profile is essential. Why quadratic?? I suggest validating the resulting temperature profile with independent information.

I think a full quantitative description of background is going to require another article. Combined with your previous point I see the outline of a very interesting project. Thanks for the great questions more work is definitely required in this area.

I tried both exponential fits and splines to model the SIN but neither were very stable solutions. Given small changes in my background selection or fitting parameters the exponential changed too drastically. I used a quadratic because it is better than a linear fit, gives me stable and reproducible results (which were important for processing so many nights of data), and removes most of the SIN in the region where the signal to noise ratio is close to one. This is not a perfect solution but, I think it is an incremental improvement.

To my knowledge there are no independent validation sources that are appropriate. I think that two co-located lidars are the best we can hope for. Satellites have their own calibration issues as I point out in part B.

P19Fig10: From my point of view the upper range of the temperature is somewhat optimistic. There seem to be superadiabatic gradients at 75 and 80 km. 30% relative error is ~70 K, i.e. the content of information is rather low. Which altitude is chosen for initialization? How is the signal smoothed for the choice of the initialization altitude (L395)? The melding of the signals should be visible in the uncertainties, but is not in Fig. 10. Please explain.

Please note that this example temperature profile was calculated at 300 m vertical resolution. I was simply demonstrating a troposphere to mesosphere temperature profile at high vertical resolution. The relative error drops significantly at 1 km vertical resolution and we generally get 30% error above 90 km.

I use a 3rd order Savitzky-Golay filter with a small 11 point window. This filtering is not passed though into the data product it is only applied to the photon counts profile for the purpose of determining where the lowest altitude where the signal to noise ratio is equal to one. This altitude is different every night and depends on the transmitter power, nightly integration time, and sky conditions.

I use a relative error weighting function to minimize the total uncertainty. This ensures that I'm not adding extra noise to my photon counts and makes the transition in the temperature profile as smooth as possible.

P20Fig11: I suggest showing the error of the mean instead of the variance.

Error on the median added. Caption and text updated.

P22L450: How many nights are excluded here?

I set a 2 sided p-value of 0.05. I didn't think to record the number of nights excluded just my confidence interval.

P22L452: Please mention the averaging window.

Added to text: 'A 30 day averaging window is applied to each of the four curves.'

P22L460: This conclusion cannot be proven without acknowledgement of the temperature uncertainty. The shaded area in Fig. 14 seems to show geophysical variability rather than measurement uncertainties. Fig. 13 shows persistent red or blue patches, indicating systematic differences between the lidars.

Good point! Thanks. 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 0.3 K at 40 km; 0.7 K at 50 km; 1.5 K at 60 km; and 4.6 K at 70 km.'

The two lidars are measuring the same air at the same time, so it can't be geophysical variability.

Yes. There are time periods where there appears to be internal misalignment in one or the other lidar. This is most definitely a systematic error. That's why I tried to identify and remove these time periods using the Chi squared method before plotting Fig15. I think this is a very good technical argument to be made for investing in automated alignment systems in lidars. To my knowledge this is the first time that anyone has looked at comparing co-located lidar signals over such a long period. I think it shows that perhaps the lidar community needs to do further work on testing signal linearity and overlap corrections.

P22Fig15: I am surprised about the small differences. Averaging the purple and blue (40 and 50 km) line in Fig. 14, I would guess the difference is ~1K. At 70 km the difference is close to 0 K, but ~1 K in Fig. 11 (green and orange line). Is there any mis-interpretation from my side?

Going back to my response to your comment on the variance in Fig11 this is an example of the error on the median. It looks really small given the large amount of data being considered. So no misunderstanding on your part - given all 20 years of data there is a (un)remarkable degree of consistency between two co-located lidars.

Minor comments:

P2L24-26: Please check this sentence (grammar).

Inserted "and"

P2L30-35: Please clean up the brackets, making this section easier to read.

Changed to square braces

P2L54: Remove "of two co-located lidar systems" and similar repetitions.

I have removed some of the repeated redundant wording. However, there are 3 to 5 independant co-located lidars at OHP (depending on how you count them). This work only uses two of the OHP lidars.

P3L56-61: I do not see this section relevant for the paper.

Motivation for other DIAL systems to submit validated temperature profiles to NDACC. I think that the people are hesitant to put forward 355 temperatures for validation as NDACC data products when the main focus of their system is ozone.

P5L99: I assume a dispersion of 0.3 mm/nm. Correct?

Typo corrected

P6L136: "multiple scattered photons"

We mean photons which have each scattered multiple times. We do not mean multiple photons which have each been scattered. We could change to "multiply-scattered photons". Please advise.

I will check with an anglophone. But I think this is correct.

P6L137: "outside of the field of view"

Done

P7L164-166: Example for textbook knowledge that can be removed.

This is intuitive to lidar scientists familiar with remote sensing but the prompt can be useful for modelers, satellite scientists, and pure geophysicists as an orientation point.

P8L167-173: This section is partly redundant and should be shortened or removed.

The authors felt it was important to show a co-added lidar signal as this may not be obvious outside our community.

P11L220: I suggest using "profile" instead of "scan".

Done

P11L230: The intuition is always subjective. Please rephrase.

We have addressed this comment in our changes to the previous comment for P11L230-233, in which this and adjacent sentences have been reworked.

P11L235-236: I suggest deleting this sentence.

Done

P13L260: Please explain "partial scan".

Changed to "partial profile". Using a partial profile entails only using the linear portions of the photon count time series and cutting out instead of correcting saturation, spikes, and other data problems.

P18L367: “in an area of low signal”

Changed “area” to “region”

P20L423: I suggest writing “The present study” instead of “This study”.

Done

P23L471: “colder” should read “lower”

Done

P26L540-544: Sentences are mixed up. Please correct.

We have edited this section so far as possible, given the limitations of proper names, in two languages, of the various funding agencies.

This section has been corrected to read:
"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://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 public access. This work is supported by the Atmospheric dynamics Research InfraStructure Project (ARISE 2) which is funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 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 technicians at La Station Géophysique Gérard Mégie at OHP.

Response *Lidar temperature series in the middle atmosphere as a reference data set. Part A: Improved retrievals and a 20 year cross-validation of two co-located French lidars*: Referee #3

Dear Referee,
Thank you very much for your helpful comments and suggestions. I have attempted to address each of your concerns to the best of my ability. If you would like me to implement further changes or iterations on a point please let me know. I appreciate your efforts to help me improve this paper.

I appreciate the push back you have given on Signal Induced Noise corrections. These are exactly the kind of discussions we need to have when designing new lidar experiments. As well we need to be aware of these kinds of problems in older lidar systems as we can't change the past. The scientific value of the OHP lidar database is often forgotten. This year marks the 40th year of continuous lidar measurements of temperature between 30 km and 85 km. This data record is longer and more stable than any satellite or rocketsonde record. Creating good software tools to deal with noise is essential for getting the most out of this exceptional resource.

You're point about lidar temperature 'accuracy' was also particularly thought provoking. I think that temperatures in the middle atmosphere are often dismissed as a 'solved problem' however, Part B of this paper shows fundamental disagreements between the lidar and satellites on the 'simple' question of stratopause heights. There may still be some unresolved questions surrounding the 'true state' of the atmosphere and how well we can know it though different measurement techniques.

****

Specific comments:

Please state the temporal resolution of your retrieved temperature profiles.

Inserted P6_L170:
' However, given that a single lidar profile is acquired every 2.8 minutes and a nightly average temperature is generated every 4 hours'

What does LiO3S stand for? In the paper you use different terms for the Ozone DIAL, e.g. line 86: OHP Differential Absorption Lidar, line 104: OHP DIAL, caption of Figure 2: LiO3S DIAL, line 162: LiO3S. Please use a single term to avoid confusion.

Added P4_L86
“also referred to as Lidar Ozone Stratosphère (LiO3S)”

Figure 2: How are the 4 fibers combined before the chopper? Is the light coming off the four fibers coupled into a single fiber which relays the light to the receiver? What is the diameter of that fiber?

They’re arranged in a linear fashion using a commercial fibre optic bundler. Diameter of the bundle is 1600 um.

Line 153: You may add that this assumption excludes observations at mid to high latitudes in summer where NLC can occur.

Sightings of NLCs over OHP seem to be extremely rare. (Pérot et al. 2010) First climatology of polar mesospheric clouds from GOMOS/ENVISAT stellar occultation instrument https://www.atmos-chem-phys.net/10/2723/2010/ has seen only a few NLCs over decades of measurements.

Section 2.3.2: Maybe more efforts should be spent on finding and eliminating the root cause of these transients. I believe there are lidar systems in operation which do not suffer from these problems.

I didn’t know these phenomena existed before looking at a full 20 years of lidar profiles. It is equally plausible that other lidar systems also have these problems but are unaware of them. I’m a PhD student in my third and final year and I can’t justify a weeks long trip to the south of France to investigate these issues. I have tried to correct the issue to the best of my ability in the software. I spoke at some length with the representatives from Licel at the IRLC2018 conference and they showed some level of concern over these TES. However, given that they do not occur frequently or with regularity it is difficult to track down.

Line 201: “We can see that the 22nd and 46th scans are contaminated by a TES with a duration of about 0.5 µs“ I can’t see that. The plot you are referring to is labeled with bins rather than time. What is a “scan”?

Each bin is 0.1 μs and the FWHM is about 5 bins. The exact width is less important than the fact that there is a temporal duration to this signal. This is what differentiates a TES from a spike which occurs in a single isolated bin without affecting data in adjacent bins.

Changed 'scan' to 'profile' everywhere in the text. These profiles are 1.6 minutes long, each as indicated on line 210.

Changed the sentence on page 10 line 208 in the text to read:
"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 . "

Changed first sentence of Fig 5 caption to read:
"Figure 5: Upper panel is a surface plot of lidar returns as a function of altitude bin and profile number...."

Line 215: "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." Well, a cirrus cloud drifting through the lidar beam might actually look like peaks in Figure 5.

Yes. That's why I do it in the profiles as well. I use a simple test to see if a potential TES shows up at the same altitude for multiple profiles. If it does it's a cloud not a TES.

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.

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.

Line 234: "We have shown two approaches for attempting to address the issue. . ." Which approaches are you referring to? Do you mean the two approaches you explain below?

Tense is changed and sentence clarified

Lines 244-246: “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.” What do you mean by “changes in signal quality”? I believe all aspects you listed, e.g. cirrus clouds, impact signal quality.

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 due to operational instrumental factors including laser power, data acquisition noise, cleanliness of optics, and issues with signal saturation.

Figure 7: Why did you chose a different data set and not use the same data set shown in Figure 6? In order for the reader to evaluate the different algorithms, it would be beneficial to show results based on the same data set.

Thanks for catching the mistake. I’ve remade Fig06 over the same range as Fig07.

Linens 267-269: What is the reason for choosing “the point where the signal to noise equals one in the density profile”?

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.

In equation (2), which profile do you use for determining the altitude of this point, the summed profile or the individual profile?

Nightly sum.

Figure 8: I am not sure about the unit on the y-axis. Shouldn’t that be just Hz?

Changed

Section 3.5.1: What is the maximum count rate at which gating (or the chopper) cuts the profile off? Have you checked the validity of equation (3) within that range? One possibility would be plotting the correction as function of count rate using equation (3) and the actual measurements (ratio of high gain and low gain channels).

The gating is based on an adjustable time delay not maximum count rate. The High gain Raleigh channel is currently blanked at 22 km and the low gain Rayleigh channel is blanked at 12 km.

We have added the following text on page 15 after line 312 to clarify.
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 equation 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 X to X km). We iterate through various values of \tau, increasing by XXX each time, 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, \tau.

Next, equation 3 is used again, using the measured nightly value for \tau, 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.

Variable names in Eq(3) have also been changed for clarity

No I haven't investigated the validity of Eq(3). I'm relying on the work of (Donovan 1993). However, if I iterate through N_recieved and N_counted I get a reasonable convergence

No. We use the high and low gain Rayleigh channels as well as the nitrogen Raman channel to directly measure the correction. The set values are only used if the data is unavailable. If the data is unavailable we have no reference to compare to.

Lines 318-319: You can actually check the validity of this assumption using the 532 nm and 607 nm channels.

Yes more recent data uses 607 nm channel to make the correction. Unfortunately, the N2 Raman channel was not present for the entire period from 1993.

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?

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.

Figure 10: The superadiabatic gradient at approximately 75 km altitude looks suspicious to me. I assume the upper part of the profile is dominated by noise a sition from the upper to the lower channels happens. Please explain why this is not the case.

Please note that this example temperature profile was calculated at 300 m vertical resolution. I was simply demonstrating a troposphere to mesosphere temperature profile at high vertical resolution. The relative error drops significantly at 1 km vertical resolution and we generally get 30% error above 90 km.

No. The low gain Rayleigh channel contributes nothing that high up. When I meld the two channels I weight the addition by the relative error.

Figure 11: What is the shaded area?

Now noted in caption it's the ensemble variance.

Line 446: What do you mean by "mis-aligned"? Please explain.

In both lidar systems the high gain Rayleigh channel has 4 mirrors each of which needs to be aligned independently. In LTA the low gain channel is a single independent mirror. So a total of 9 mirrors need to be aligned every night to make a Rayleigh measurement.

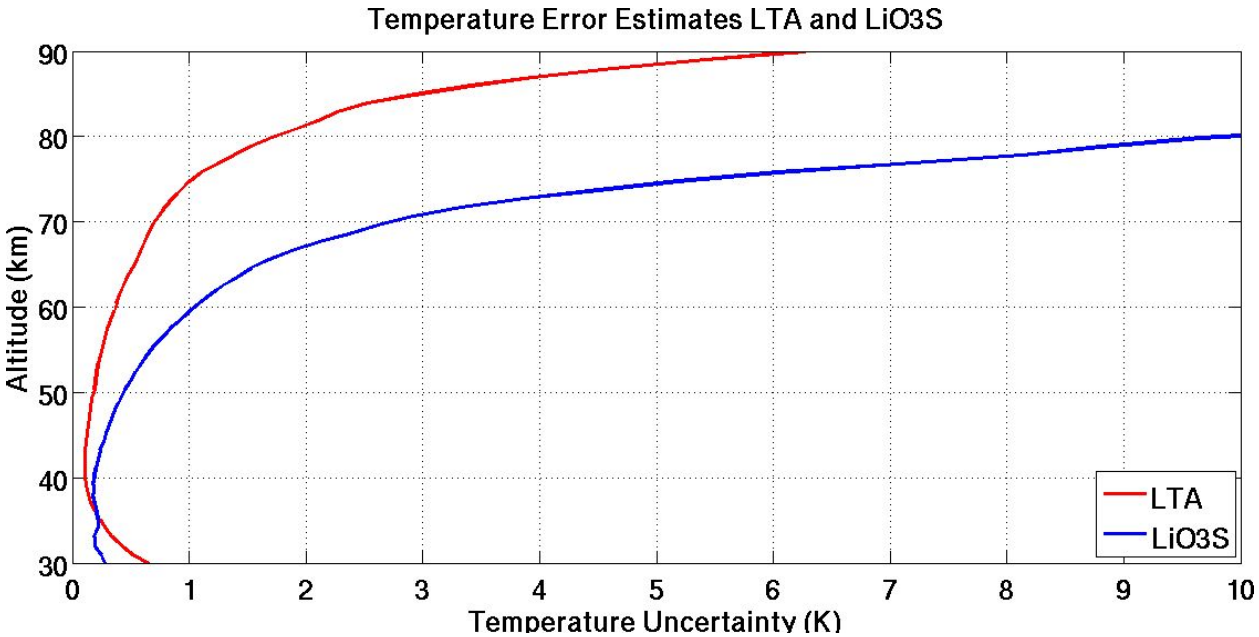
Line 456 inserted text: **Internal misalignments happen when one or more of the five mirrors in LTA or four mirrors in LiO3S is not properly aligned with the laser or the fibre optic is not centered on the focal point of the mirror.**

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.

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



Line 441: For clarification, the observation period of one lidar could be up to 20% longer compared to the other lidar? Why not just make both observation periods equal in length by cutting the longer observation?

4 hours is the standard OHP temperature measurement. This criterion excludes the few cases when there is a significant temporal offset between the two lidars. Maybe one lidar was being temperamental and took an extra hour to start up. I didn't look into why some measurements were not coincident I just excluded them if they were too different.

Line 442: What is meant by "good internal alignment"? Figure 14: Can you please mark periods of misalignment similar to Figure 13.

In both lidar systems the high gain Rayleigh channel has 4 mirrors each of which needs to be aligned independently. In LTA the low gain channel is a single independent mirror. So a total of 9 mirrors need to be aligned every night to make a Rayleigh measurement.

Line 456 inserted text: **Internal misalignments happen when one or more of the five mirrors in LTA or four mirrors in LiO3S is not properly aligned with the laser or the fibre optic is not centered on the focal point of the mirror.**

Fig14 is marked

Lines 459-461: "Without excluding misaligned periods the lidar temperature differences are not significant as a function of altitude or year at the 2 sigma level". I am not sure if I understood that sentence correctly. Is the implication removal of misaligned periods causes the differences become significant intended?

Thanks for picking up on that implication - indeed it is not what we mean. We have rephrased the sentence.

Replace lines 459 through 461 with this text:
Using all data, including misaligned periods (example: winter 2006-2007 in Fig. 13 and Fig. 14) none of the lidar temperature differences are 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.

Lines 462-463: "After removing comparisons between mis-aligned instruments we can calculate the ensemble median difference between the two systems." I do not understand that sentence. What was removed? The data affected by misalignment?

Yes. A chi-squared test was used to detect these nights and exclude them from the rest of the analysis.

Line 486: "lidar measurements are accurate" I do not think you have sufficiently backed up this claim. Maybe it depends on what we understand by "accurate". I agree that the long-term average (20 years) appears to be accurate, however according to Figure 13 nightly means obtained by the two co-located lidars can differ by more than 10 K. What is the reason for these large differences? Are these large differences expected from an SNR point of view, or are there other maybe unknown error sources which average out on long time scales?

I completely agree the differences cannot be explained from a SNR point of view. I think that alignment is a major unaddressed problem in lidar science. With a single lidar we often assume that our transmitter and receiver are well aligned if we can maximise the

count rates at some reference altitude. Before conducting the comparison between the two lidar systems I did not fully realize how great of an effect slight misalignment can have on the resulting temperature profile.

When we make temperature comparisons with other techniques like radiosondes, satellites, or other sensors we can easily dismiss small deviations based on claims geophysical variability, sampling slightly different air masses, averaging effects, or sampling error. However, in this study we have two active remote sensors, making measurements in the same building, at the same resolution, operated by the same technicians, designed by the same optical and electrical engineers, and compared over 2 decades. They should be the same. But since they are not I think it is legitimate to entertain the possibility that manual nightly alignment of lidars is not as robust or repeatable as we like to assume.

The Pandora's box that is lidar alignment is really horrifying when you sit and try to imagine all the possible sources of alignment drift: operator change blindness (not noticing small changes over a long period of time), thermal changes shifting optics, optical degradation, hysteresis in optical mounts, angular sensitivity of optics which exceeds manufacturer specifications. Etc.

Your point about 'accurate' is well taken. Changed to 'reasonable'

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

Lidar temperature series in the middle atmosphere as a reference data set. Part A: Improved retrievals and a 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 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 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 and is a 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) [5-10 K relative to HALOE, 6-10 K relative to COSMIC/CHAMP, 10-16 K relative to SABER]; (Yue et al., 2014) [13 K at 75 km relative to SABER]; (García-Comas et al., 2014) [3-4 K at 60 km relative to SABER and MIPAS]; (Yue et al., 2014) [13 K at 75 km relative to SABER]; (Dou et al., 2009) [4 K at 60 km relative to SABER]; (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 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 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 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 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 (mm)	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).

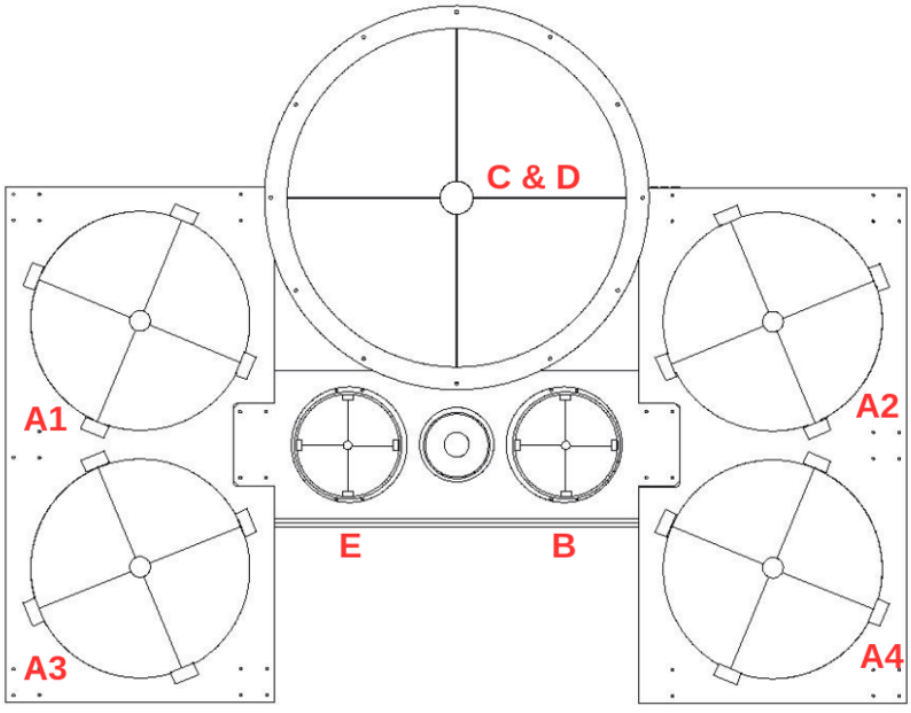


Figure 1: Mirrors A1, A2, A3, A4 (50 cm) are combined for the high gain Rayleigh channel. B (20 cm) is low gain Rayleigh channel. Mirror C&D (80 cm) is the Raman channel for water vapour and molecular nitrogen. E (20 cm) is the Mie channel. The beam expander for the transmitted laser source is between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

2.2 DIAL Ozone System (LiO_3S)

The OHP Differential Absorption Lidar (DIAL), also referred to as Lidar Ozone Stratosphère (LiO3S), 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 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/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).

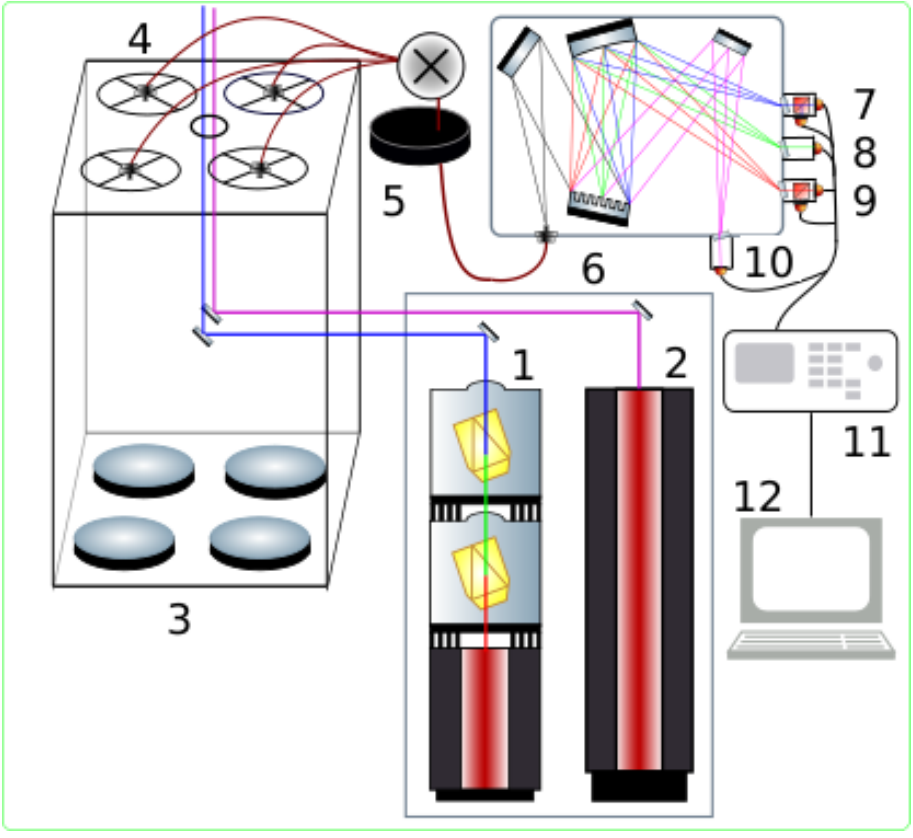


Figure 2: LiO_3S DIAL system. **1** 355 nm laser source, **2** 308 nm laser source, **3** four 530 mm mirrors, **4** four actuator mounted fibre optic cables, **5** mechanical chopper, **6** Horiba Jobin Yvon holographic grating, **7** 308 nm high and low gain photomultipliers, **8** 331.8 nm photomultiplier, **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):

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

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

z is 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

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 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). 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 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 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 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.

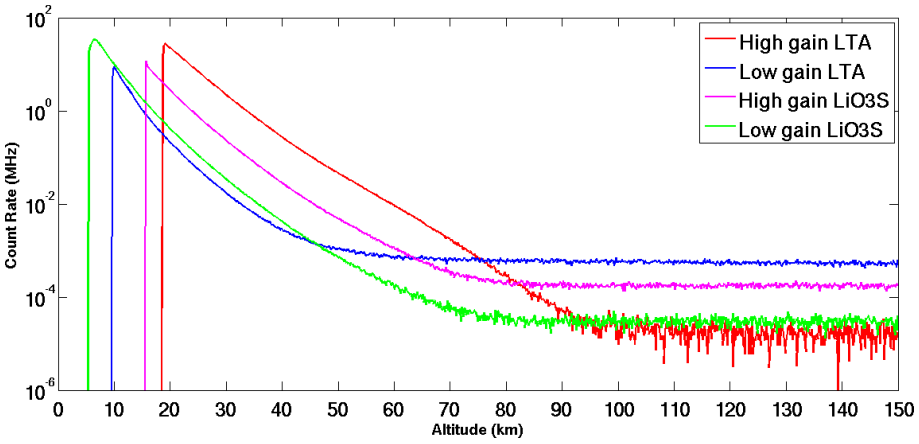


Figure 3: Nightly co-added profiles for high and low gain Rayleigh signals for LTA and LiO_3S. The background for LTA extends to 246.23 km and for LiO_3S extends to 154.13 km. A single lidar profile for both LTA and LiO_3S has a temporal resolution of roughly 2 minutes and 45 seconds and a vertical resolution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

Figure 3 shows four nightly co-added 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 will result in the removal of true photons, a lower estimated density, and by the ideal gas law, a warmer 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 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.

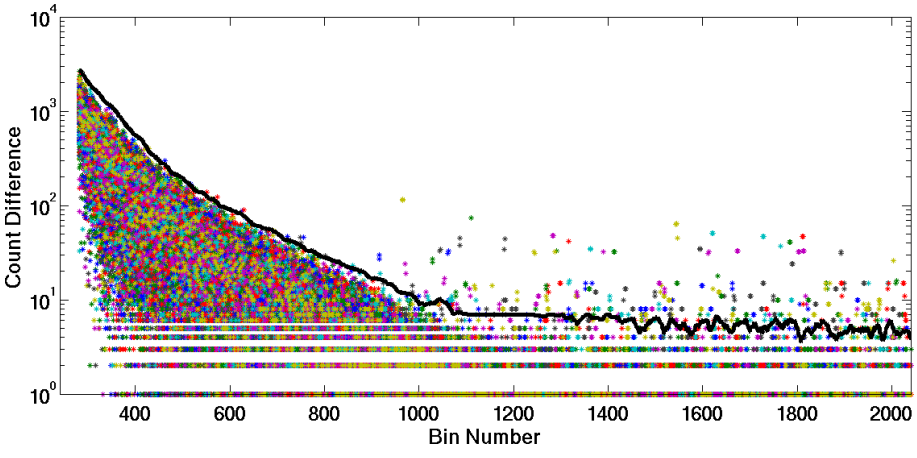


Figure 4: Tukey Quartile spike identification based on the signal difference between consecutive lidar time bins for short integration lidar returns. An entire night of lidar profiles is over-plotted in the stack plot. The black line is the 2 sigma limit and points above this line are removed.

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 counting rate in bins subsequent to the TES burst. In the example shown in Fig. 5 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 profiles for a night of lidar data with each profile accounting for roughly 1.6 minutes of lidar data. We can see that the 22nd and 46th 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 profile and the red line shows the 2σ estimation of total lidar profile skew. Isolated profiles with a total kurtosis above this limit are excluded.

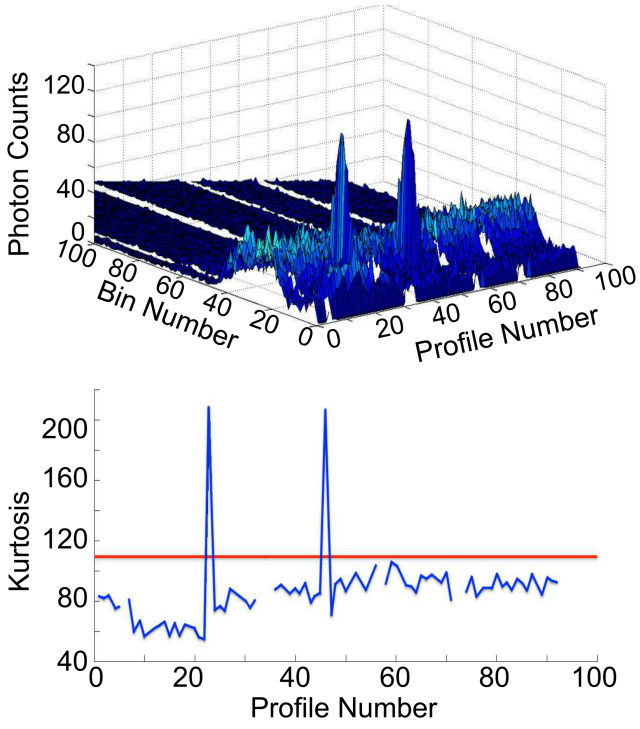


Figure 5: Upper panel is a surface plot of lidar returns as a function of time bin and profile number. For clarity, only the first 100 bins are shown in this plot. The test is carried out using all bins of each profile. Two instances of TES can be seen as anomalous peaks in the photon count rate. Lower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

3.3.3 Bad Profiles

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. Conceptually, 'bad profiles' are lidar profiles with a high background and/or a low signal strength. These profiles need to be positively identified as not belonging to the general population of nightly lidar profiles and excluded. Quantitatively, identifying a 'bad 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. We might suggest that profiles 1 through 23 and profiles 36 through 46 might belong to one population and the rest (excluding profile 69) belong to a second population. However, when we look at the panel representing the signal, 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 (profiles 1-23, profiles 24-35, profiles 36 - 48 and profiles 49 - 92) . However, whether these four groups of signals should be treated in analysis

as two, three, or four distinct populations is open to interpretation. Therefore, we seek an objective programmatic solution for identifying bad scans. We now show two approaches for attempting to address the issue of changing signal quality. In Fig. 6 the green margin is an attempt to identify 'bad 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 profiles in a given night was too highly dependent on the number of degrees of freedom specified in the Matlab tool.

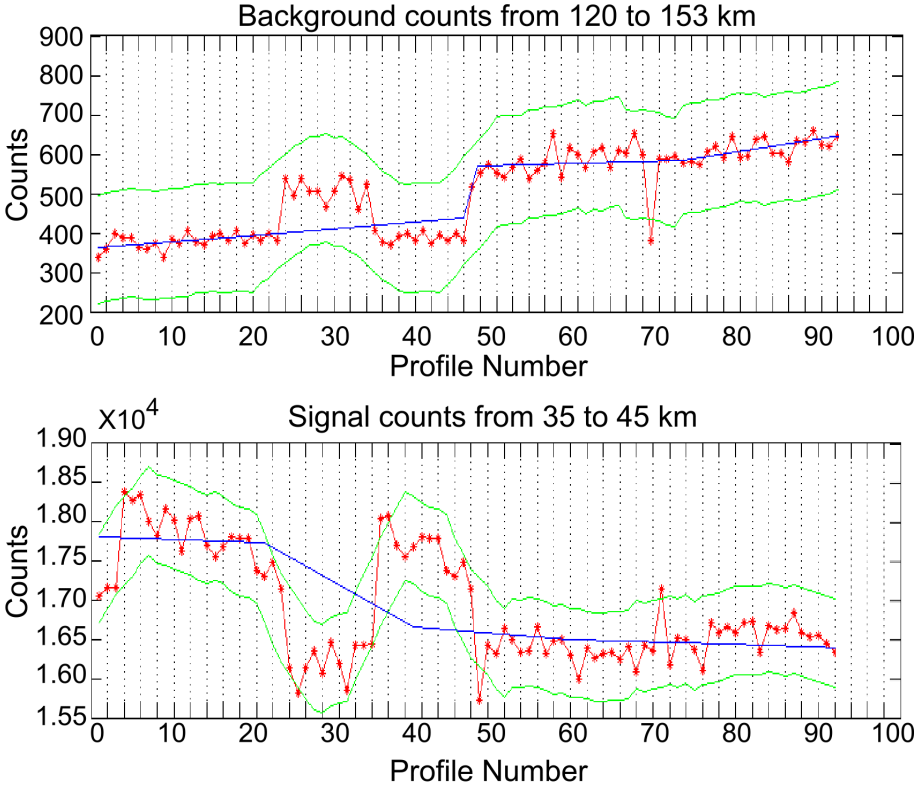


Figure 6: Example of lidar signal and noise during a night of measurements. Top panel shows the total background counts summed from 120 km to 153 km and the bottom panel shows the total signal summed between 35 km and 40 km. Green bounds are calculated based on a smoothed 2σ error estimation of the summed 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 profiles and determine which should be excluded requires a non-parametric statistic. We use a one sided non-parametric Mann-Whitney-Wilcoxon rank-sum test (Mann and Whitney, 1947) to identify lidar profiles which do not belong to the nightly population or subpopulations of lidar profiles.

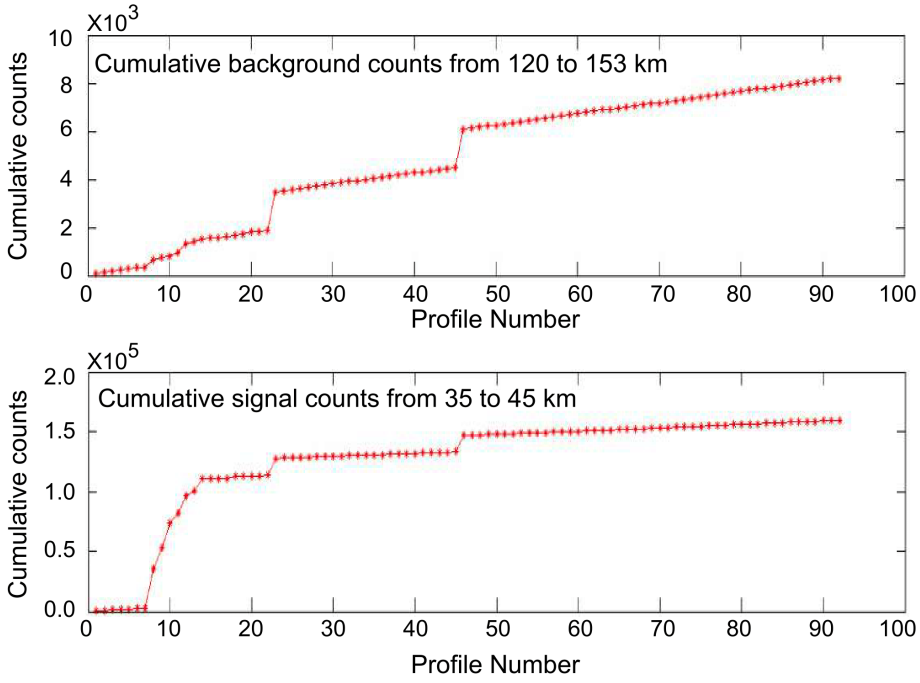


Figure 7: Rank sum plots for a night of lidar data. Top panel is the cumulative background count and the bottom panel is the cumulative signal count. The signal to noise ratio of the rank summed photon counts in each profile is evaluated using a Mann-Whitney-Wilcoxon rank-sum test to determine if an individual lidar profile belongs to the nightly population of lidar 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 profile' which takes into account the nightly median without the assumption that the quality of lidar profiles is normally distributed. In this example the first 13 profiles fail the rank-sum test and are discarded.

3.3.4 Good Profiles

Given that our objective is to calculate accurate temperature profiles at the highest possible altitudes we must quality test each profile that we choose to include in the nightly average. It is possible to include partial profiles but that is not done in this work. The conceptual difference between a 'bad

profile' and a 'good profile' is that bad profiles are positively identified as outliers to the general population whereas good profiles represent the portion of the population of profiles which contribute more information than noise to the nightly average at a given altitude. Consider that a poor quality lidar profile which has a signal to noise ratio of 1 at 70 km contributes more information than noise at 60 km, but more noise than signal 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 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 noise, 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 ratio equals 1. Then we calculate the altitude that is one density scale height ($\sim$ 8 km) below this point. The lidar range bins which correspond to this altitude range are then summed to yield S_{sum}. A similar calculation, using the same range bins as in the nightly average calculation, is done to determine the signal of single profile, S_i. If a 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.

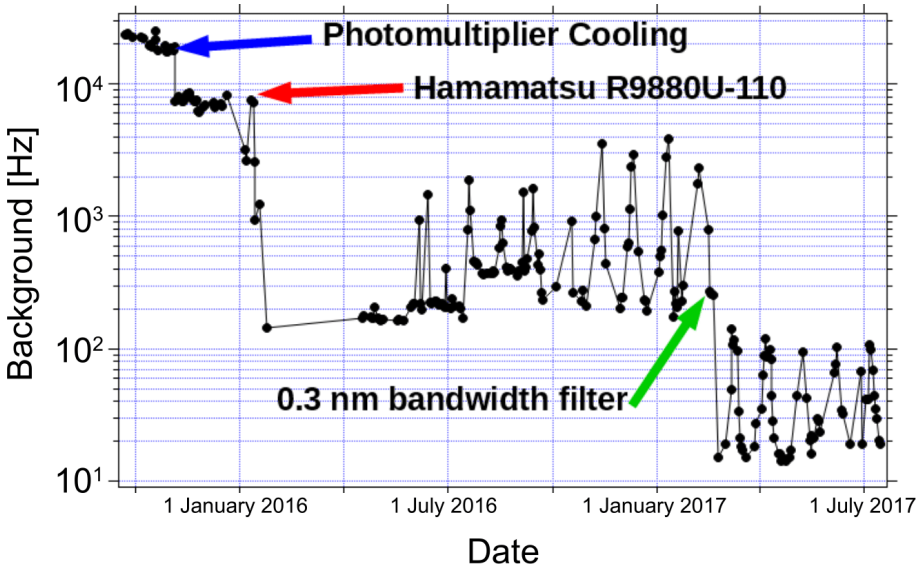


Figure 8: This figure shows the improvements in the background count rate due to photomultiplier cooling, new photomultipliers, and new optical filters. Note the logarithmic y-axis and the total reduction of background counts by more than 2 orders of magnitude.

3.5 Corrections Applied Before Temperature Calculation

In the previous subsection we detailed the process for removing bad profiles from our nightly lidar measurement. In this subsection we will detail several corrections to our remaining photon counts 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 $N_{received}$ 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 X to X 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.

$$N_{dtc} = N_{uncorrected} * exp(\frac{\tau * N_{uncorrected}}{\Delta t}) \tag{3}$$

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. 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.

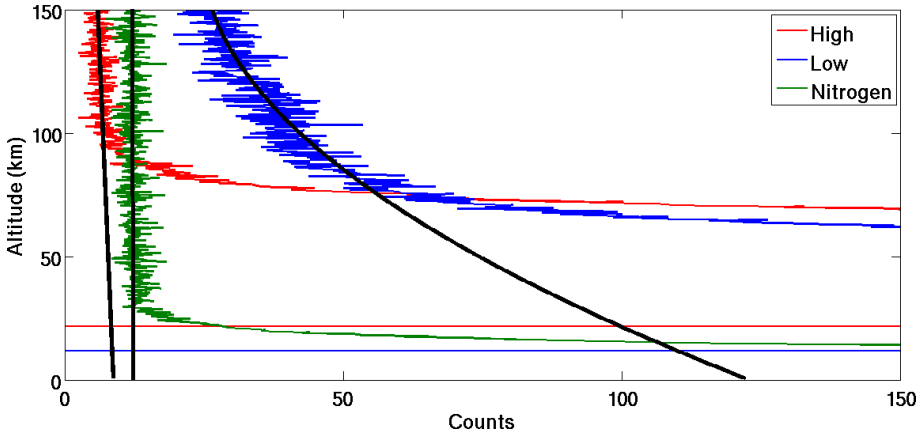


Figure 9: An example of a non-linear signal induced noise in the low gain Rayleigh channel best estimated by a quadratic background. Also shown is the high gain Rayleigh channel (red) with a background best fit by a negative linear function and the nitrogen Raman channel (green) with no apparent signal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

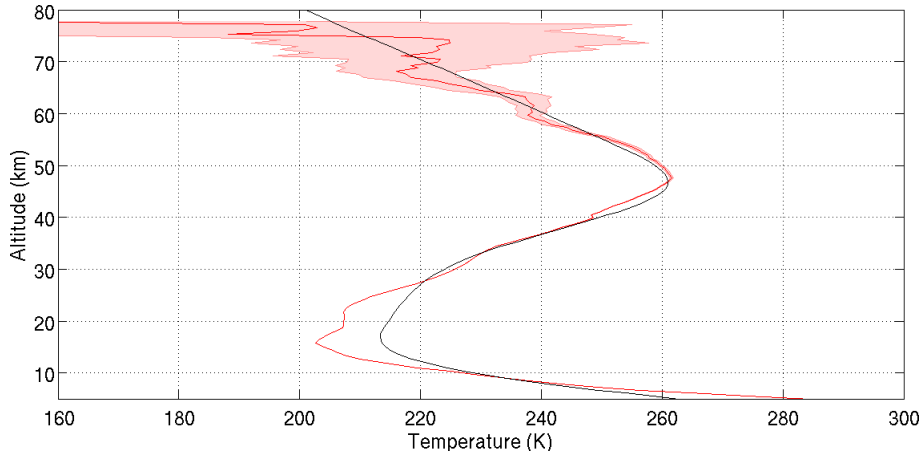


Figure 10: An example of a melded temperature profile from two Rayleigh channels and one Raman channel. The profile is calculated at 300 m vertical resolution from a single combined photon count profile and has a maximum relative error near 80 km of 30%. Black line is the MSIS-90 temperature profile 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 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 near 85 km and 20 K by 90 km. There is no significant change to the lidar temperatures below 70 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 LiO_3S (orange), the satellites MLS (red) and SABER (blue with median error and shaded ensemble variance), and the MSIS-90 model (magenta). 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).

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 (Wing et al., 2018b).

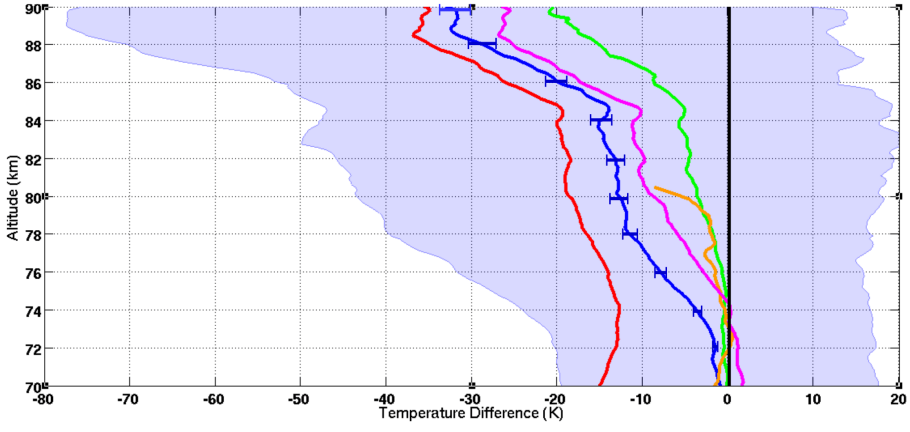


Figure 11: Ensemble temperature differences from NDACC standard LTA Rayleigh temperatures (black). MLS (red), SABER (blue with median error and shaded ensemble variance), MSIS-90 (magenta), LiO_3S (orange), and LTA Rayleigh temperatures with corrections given in this work (green).

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). The present study proposes a completely novel type of inter-lidar study on the 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. 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 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 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.

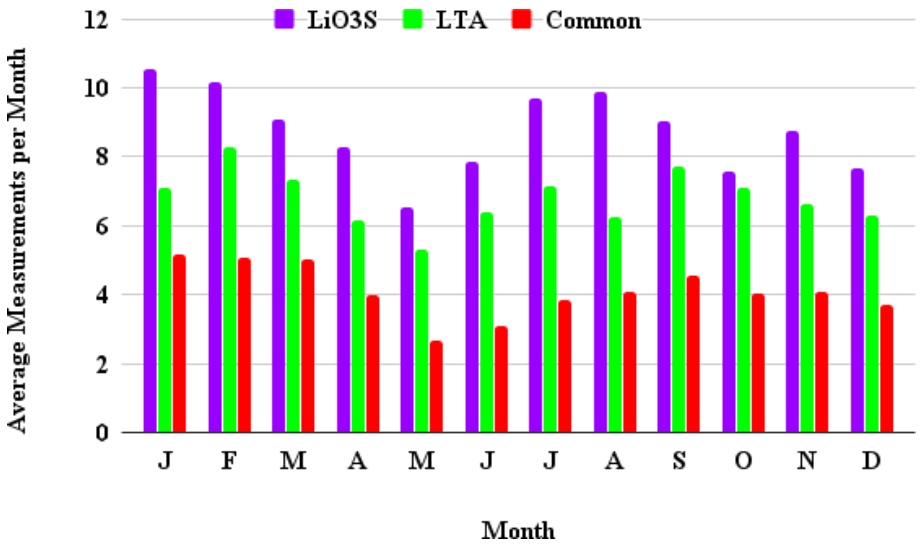


Figure 12: Average number of OHP lidar temperature measurements per month during the period of 1993-2013.

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.

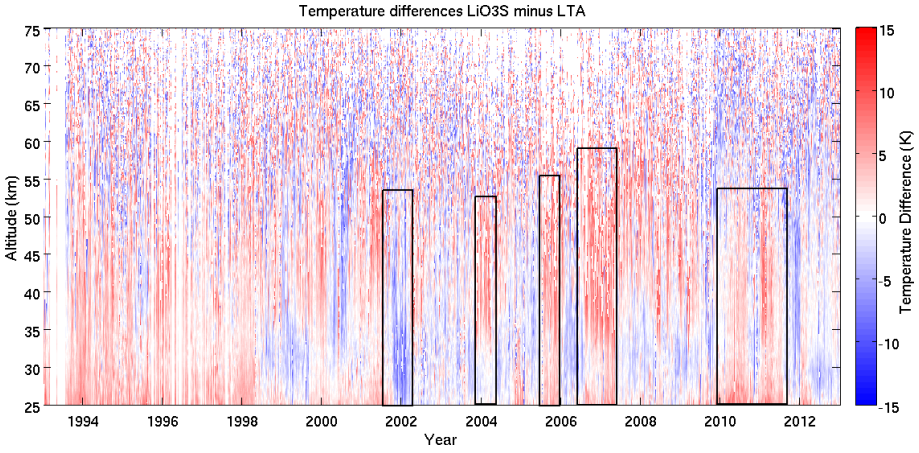


Figure 13: Temperature differences between LTA and LiO_3S OHP lidars for a 20 year period between 1993 and 2013. There are 1496 nights of comparison in this plot. Red indicates that LiO_3S was warmer than LTA and blue that it was colder. The black boxes highlight periods where the two lidars were out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

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. Using all data, including misaligned periods (example: winter 2006-2007 in Fig. 13 and Fig. 14) none of the lidar temperature differences are 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.

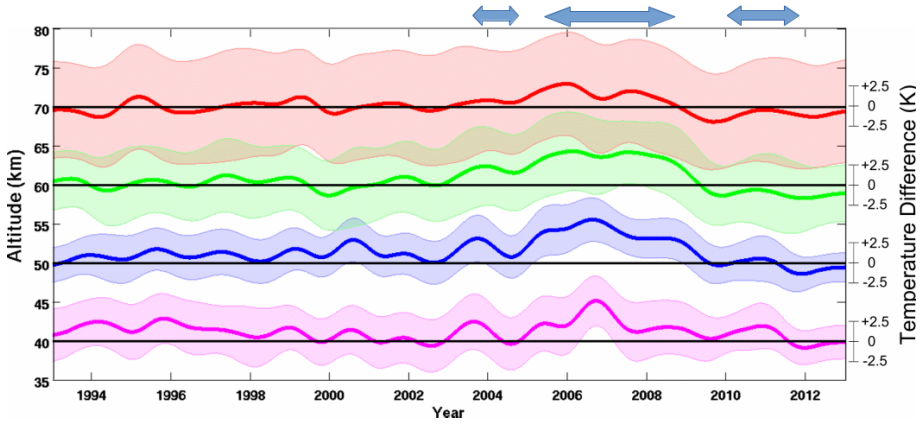


Figure 14: Average temperature differences between LTA and LiO_3S OHP lidars for a 20 year period between 1993 and 2013 at four altitude levels: 65-75 km (red), 55-65 km (green), 45-55 km (blue), and 35-45 km (magenta). Shaded uncertainties are shown at 1 sigma for clarity and the black lines are zero 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 (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 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 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 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.

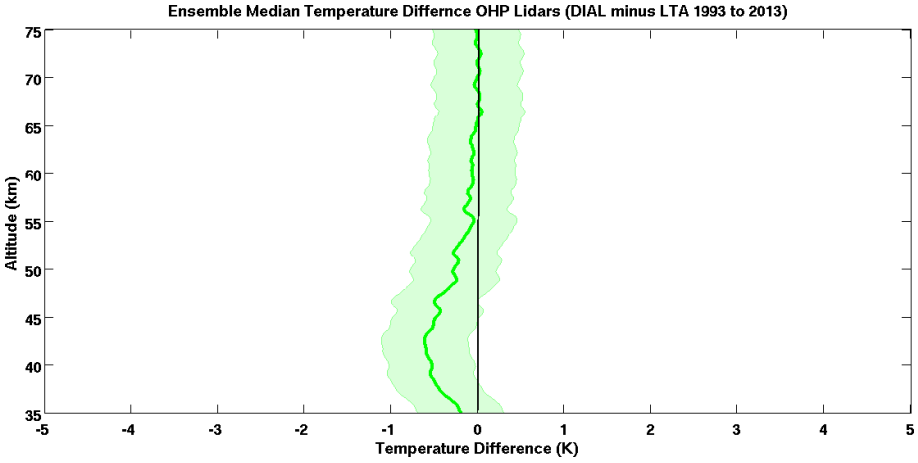


Figure 15: Ensemble of median temperature differences between LTA and LiO_3S based on temperature measurements between 1993 and 2013. Shaded error is the two sigma distribution about the ensemble.

6 Summary and Discussion

6.1 Changes to Lidar Temperature Algorithm

In this work we have attempted to minimize 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://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 public access. This work is supported by the Atmospheric dynamics Research InfraStructure Project (ARISE 2) which is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 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 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.

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.