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.

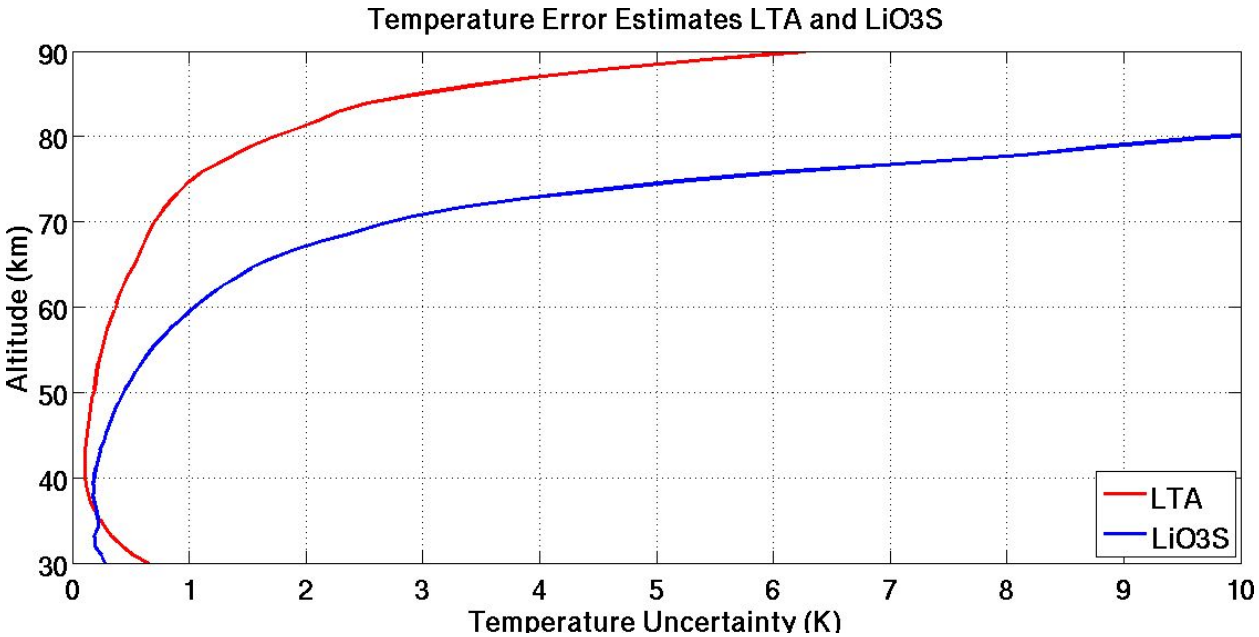
Line 456 inserted text: **Internal misalignments happen when one or more of the five mirrors in LTA or four mirrors in LiO3S is not properly aligned with the laser or the fibre optic is not centered on the focal point of the mirror.**

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'