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.

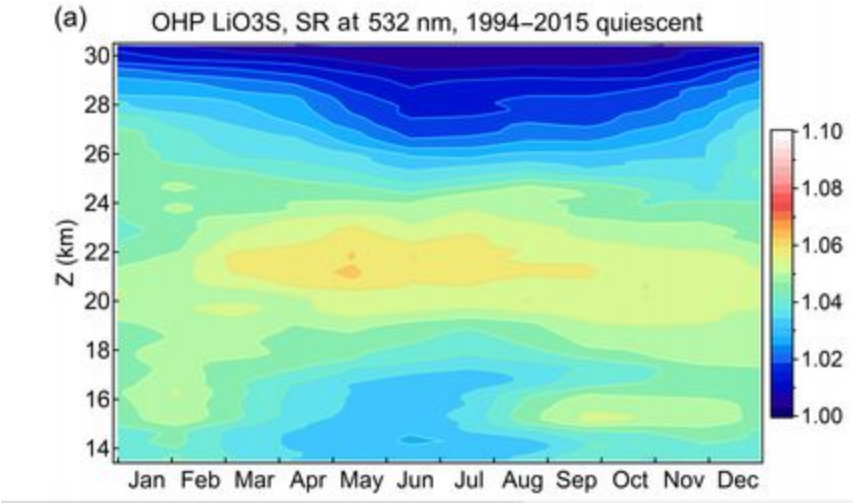
P7L151: This assumption is problematic as there are different studies showing aerosols up to at least 35 km.

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:

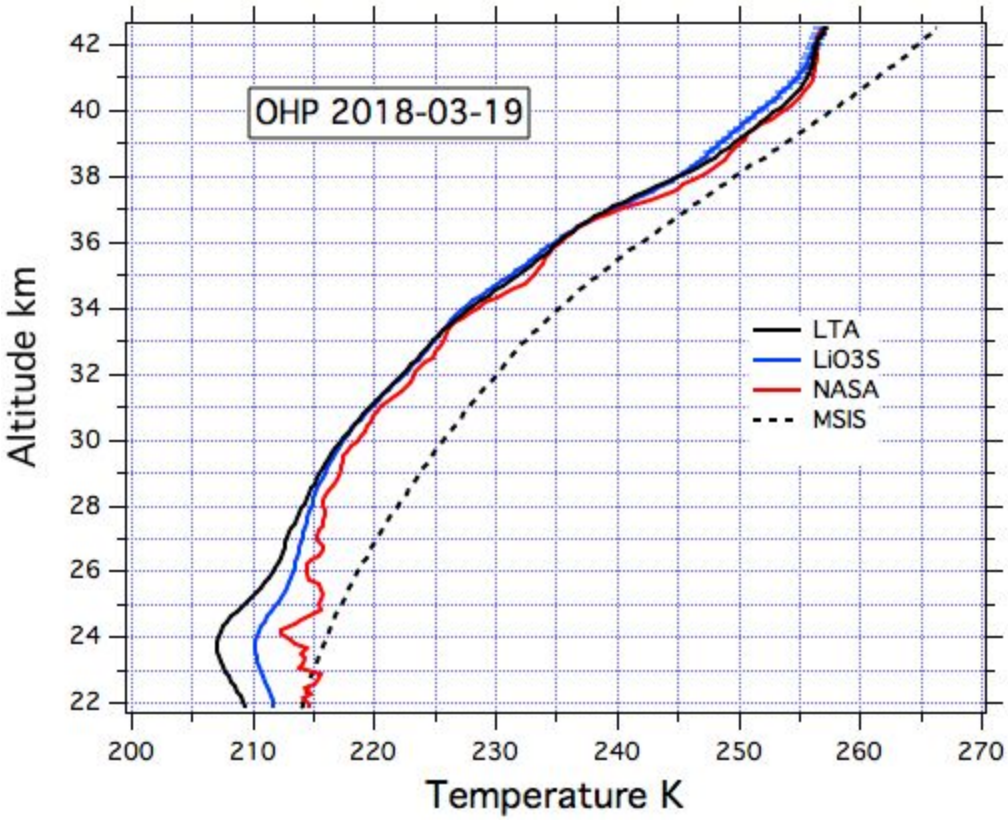
Fourth, we assume that the atmosphere at mid-latitudes is generally free of aerosols above 30 km when there are no active volcanic or fire events (Hauchecorne and Chanin, 1980). During less severe background aerosol conditions (aerosol scattering ratio < 1.02), (Gross et al. 1997) suggests lidar temperature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