Atmospheric Measurement Techniques Discussion
Response to Short Comments by Tyler Thorsen – September 2019

"Cloud-Aerosol Transport System (CATS) 1064 nm Calibration and Validation" - Pauly, R., J.E. Yorks, D.L. Hlavka, M.J. McGill, V. Amiridis, S.P. Palm, S.D. Rodier, M.A. Vaughan, P. Selmer, A.W. Kupchock, H. Baars, A. Gialitaki

We received short comments from Tyler Thorsen (tyler.thorsen@nasa.gov) of NASA Langley Research Center. Some of his comments are also raised by the two referees of the manuscript. Where appropriate, we will echo those responses here in this document. Our responses appear below in red.

Comment on "Cloud Aerosol Transport System (CATS) 1064 nm Calibration and Validation" by Pauly et al.

Tyler J. Thorsen
NASA Langley Research Center; Hampton, VA
tyler.thorsen@nasa.gov

With several recent CATS science-related publications, this is a timely manuscript on the inner working of CATS and its retrievals. The wide-reach of CATS-related science applications certainly makes this manuscript appropriate for publication in AMT, eventually. However, in its current form, serious revisions are needed as there is a general lack of rigor and substantial ambiguity in the writing and analyses.

Show the full variability of the coefficients
The objective of this paper is to calculate calibration coefficients and their uncertainties, yet only a select 120 days of nighttime coefficients are shown in Fig. 3. I strongly encourage the authors to showcase the result of all their hard work: plot the entire 2.5+ years of calibration coefficients (night and day) along with their uncertainties. Just showing 120 nighttime coefficients over some unknown time period leaves several open questions in a reader's mind. How stable is the calibration? Are then trends/drifts in the calibrations and their uncertainties that a user should be aware of? Does the relationship with the cold plate temperature hold over the entire mission?

One of our referees also asked us to expand on the evolution of the CATS nighttime and daytime calibration coefficients and what specifically causes the fluctuations. We echo our response here.

The time evolution of the CATS calibration coefficients is correlated to the thermal stability of the cooling loop on the ISS, which in turn is attributed to the changing of the sun's angle with respect to the ISS orbital plane, known as its beta angle. The CATS nighttime calibration coefficients oscillate from 4×10^8 to 1.4×10^9 $km^3 sr\ J^{-1}$ counts with a period of roughly 30-40 days. This oscillation is a result of changes in the CATS laser properties (i.e. wavelength, alignment, energy) due to thermal instability of the cooling loop. The thermal instability of the cooling loop and instrument was monitored by the cold plate temperature. Text has been added to state all of these changes on page 6, lines 30-38. Also, Fig. 3 has been updated to include the entire mode 7.2 dataset (top, April 2015 – October 2017) as well as a subset from January- April 2016 (bottom). The daytime calibration coefficients for each month have been added as red dots. A discussion of the daytime calibration values, variability, and comparison to nighttime calibration coefficients has been added starting at pages 7, line 36. Unfortunately, the funding for producing CATS data products has expired. But, if we were to ever receive funding to create another version of the CATS data products, we would make more rapid estimates of the daytime calibration coefficients than the current monthly estimates.

In addition to a time series of the coefficients/uncertainties, compositing these values by local time would be extremely useful information for the science community. The sampling throughout the diurnal cycle is one of CATS' most unique aspects that has, and will continue to, attract interest from those investigating diurnal cycles. However, any potential conclusions from these analyses needs to be tempered by the increased calibration uncertainty (and other uncertainties) during the daytime. Plotting and discussing the dependence of the calibration coefficients and their uncertainties with respect to local time would provide valuable context for those wishing to use CATS to study the diurnal cycle. Showing how the calibration uncertainties vary is essential as they propagate into every aspect of the downstream science data products. The authors tend to focus on the effect of calibration uncertainties on the in-aerosol / in-cloud attenuated backscatter. But, before they are relevant there, they impact the detection thresholds. Somewhere, the authors should comment on how the calibration uncertainties and their diurnal variability impacts their feature detection. For example, presumably the increase in daytime calibration uncertainty necessitates more conservative detection thresholds to avoid any false positives. Have estimates been made of how many features may go undetected from this? This could be sussed out by imposing an artificial increase in the nighttime detection thresholds to see what features go undetected.

Since the CATS daytime calibration coefficients are computed on a monthly basis and the CATS nighttime calibration coefficients oscillate with a period of roughly 30-40 days, there is little change, if any (daytime), in the calibration values throughout the diurnal cycle on the Earth’s surface for a given day. There are day/night differences, and these can be seen in Figure 3. We agree that calibration uncertainties propagate into every aspect of the downstream science data products. This is undoubtedly an interesting topic, but beyond the scope of this paper (it is a paper in itself). Based on a referee comment, we have added some brief text to the conclusion to express the relationship between calibration and data products such as optical depth and extinction.

Proper CALIOP comparisons

The CATS and CALIOP comparisons in Section 2.1 and Fig. 1 need to be refined and expanded upon. First, the altitude range in Figure 1 should be extended to include both the CATS and

CALIOP calibration range (i.e. up to 39km): the SNRs at these high altitudes are the relevant ones for calibration.

We did not address this comment in the latest version of the paper, for the following reasons:

a) Since CALIOP does not use the molecular normalization technique to calibrate its 1064 nm channel, computing molecular backscatter SNR at high altitudes still will not yield an apples-to-apples comparison. The reason that CALIOP does not use the molecular normalization technique at 1064 nm is that its high-altitude molecular SNR is far too poor. The intent of Figure 1 is to demonstrate that the CATS nighttime SNR is substantially larger than CALIOP's, and hence the atmospheric normalization technique is indeed viable *at night*.

b) CALIOP does not make 1064 nm measurements above 30 km.

c) The 8.2-to-20.2 km comparison region was chosen to illustrate SNR differences simply because the vertical and horizontal resolutions of the CATS and CALIOP level 1 data products are almost identical there (~60 m vertical for both instruments; ~350 m horizontally for CATS vs. ~335 m for CALIOP).

In addition to comparing SNR in the entire column, as is done in Section 2.1, the text should be expanded to include a discussion of the SNR difference in each lidar's respective calibration regions. Additionally, since CALIOP performs its calibration at 532 nm, the SNR profiles for CALIOP at 532 nm should be added in as well and compared with CATS.

On this point we disagree. What's relevant to this paper is assessing the accuracy of the CATS 1064 nm calibration. The 532 nm SNR in the CALIOP calibration regions is relevant to this exercise only insofar as it contributes to the error budget of the CALIOP 1064 nm calibration coefficients (e.g., see section 4.3 in Vaughan et al., 2019).

Despite the wavelength difference, comparing CATS 1064 nm to CALIOP 532 nm is more of an apples to apples comparison since they are the respective workhorse wavelengths for each lidar. These are the wavelengths for each lidar that are calibration, where feature detection is performed and the most accurate optical properties are available for. Therefore, comparing CATS 1064 nm to CALIOP 532 nm is the most relevant comparison to those using the data products.

To repeat a response to a referee comment, "the propagation of calibration errors in the solution of the lidar equation is both nonlinear and non-trivial, hence a more complete discussion of the link between calibration uncertainty and Level 2 data product uncertainties lies well beyond the scope of this paper. A complete mathematical description of calibration error propagation for elastic backscatter lidar measurements is given by Young et al., 2013 and Young et al., 2016." We note that the development given in the Young papers is wavelength agnostic, and thus would apply in equal measure to CALIOP at 532 nm and CATS at 1064 nm.

However, as the authors do discuss, it is important to also point out CATS' superior SNR at 1064 nm during nighttime for those whose particular investigations would benefit from this.

Error analysis

Section 3 is very unclear. Words like "overall" and "typical" are used to describe the various numbers and ranges given without any explanation of what they correspond to. Please be more precise when giving these numbers. Are these the average calibration uncertainties? Are the ranges interquartile ranges? Minimums and Maximums? Standard deviations? What is “var” in Eq. (13)? The authors seem to refer to this as “variability", do they mean variance? Please also indicate at what significance level the uncertainties presented here and in the data products are given for.

The words “overall’ and “typical” have been removed from Section 3 and Equation 13 has been updated as suggested.

The uncertainty in the assumed backscatter color ratio does not appear to be included in the error analysis.

Equation 10 has been updated and now includes a color ratio term. Regarding the evaluation of this term we will repeat a response to one of the manuscript’s referees.

To the authors’ knowledge, the value or variability of the stratospheric aerosol backscatter color ratio is not documented in the literature. For the mean value, we follow Hair *et al.* (2008), so χ_P =0.40 is taken as a constant for the aerosol loading in the upper troposphere/ lower stratosphere. This value is originally derived from backscatter data shown in Spinhirne et al. (1997). Given that sulfate aerosols are potentially the largest contributor to the stratospheric aerosol loading, this value is also consistent with lower tropospheric measurements of sulfate aerosols. Text has been added to page 5, lines 6-9 that now states this.

Its value can vary substantially for (e.g. Burton et al., 2012), which should be accounted for in the error budget.

We disagree with this statement. For aerosols in the lower troposphere, as discussed in Burton et al. (2012), backscatter color ratios can vary substantially because of shorter lifetimes and plumes that are more heterogenous. For aerosols in the stratosphere, as discussed here in our AMTD paper, backscatter color ratios are not expected to vary much (unless a large volcanic eruption occurs that injects aerosols into the stratosphere – and no eruptions injected aerosols into the CATS calibration altitude region during the 33 months of operation) because aerosols in the stratosphere have longer lifetimes and plumes are more homogenous. A nice general overview of stratospheric aerosol lifetimes and properties is given by Kremser et al. (2016). To estimate error (or variability) in the backscatter color ratio, we performed an analysis of SAGE III extinction Angstrom exponent, averaged from June 2017 to August 2018 in the CATS calibration region, to find a mean/standard deviation of 1.79 ± 0.10. We use this standard deviation as a relative uncertainty for the backscatter color ratio (6%), so we assume a stratospheric aerosol backscatter color ratio of 0.400 ± 0.024. This is now explained in the text on page 8, lines 26-33.

Kremser, S., et al. (2016), Stratospheric aerosol—Observations, processes, and impact on climate, Rev. Geophys.,54,278–335, doi:10.1002/2015RG000511.

Alternatively, this could be avoided by just using the CALIOP 1064 nm scattering ratios directly (see my comment below on this) and replacing the Kar et al. (2018) 532 nm CALIOP calibration uncertainty with the 1064 nm CALIOP calibration uncertainty given in Vaughan et al. (2019).

Unfortunately, it's not that straightforward. The reason that we did not use the CALIOP 1064 nm scattering ratios is the same reason CALIOP does not use the molecular normalization technique at 1064 nm - its high-altitude molecular SNR is far too poor.

I was disappointed that the correlation between the nighttime calibration and the cold plate temperature was not exploited more (although it is not clear if relationship holds outside this 120 day period, see my comment above). I think the authors are missing out on an opportunity to explore improving their daytime calibration using this regression.

See our previous comments on the relationship between the calibration and the cold plate temperature.

I would also suggest adding a short paragraph to the end of this Error Analysis section comparing the calibration uncertainties to other work that has performed a similar normalization, specifically MPLnet and CALIOP. Both of these where mentioned in the Introduction as forming the basis and background for this current study. Some brief context relative to MPLnet/CALIOP would make a nice connection back to your initial motivation and help give perspective to the readers that are more familiar with MPLnet/CALIOP than they are CATS.

This information is well documented in the literature and referenced in this paper.

Validation

I appreciated that the validation of this calibration tough: HSRL/Raman techniques aren't feasible at 1064 nm, so all your left with is comparisons to other lidars who also need to calibration to a molecular signal. Because of this, one cannot treat CPL and EARLINET as absolute truth. Therefore, I suggest re-framing the discussion in section 4 around comparing the two profiles in the context of each instrument's calibration uncertainty (e.g. add uncertainty bars to the profiles Figs. 5-9 that correspond to each instrument's respective calibration uncertainty). That would put these comparisons within the proper context. Without this, it is easy to read too much into apparent absolute agreement as the authors themselves do on page 9 lines 29-30 where the agreement is called "surprising".

In some instances, we have already done this. For example, in Section 4.2 we say "The difference between the two instruments falls within the uncertainties in the CATS ATB (Sect. 3) and the uncertainties in the Polly^{XT} retrievals." We have edited parts of Sections 4.1 and 4.3 to re-frame the discussion as suggested.

The CPL/CATS agreement is NOT "surprising" after considering that the CPL was scaled by an assumed scattering ratio of 1.27! This large factor is quite uncertain and one could chose many reasonable values for it that would strongly impact the comparisons in Figs. 5 and 6. Showing the uncertainties involved would help avoid one reading too much into any agreement/disagreement.

Our "assumed scattering ratio of 1.27" was not hand-picked out of thin air simply to get better agreement between CATS and CPL. Instead, this value comes straight from Table 4 of Vaughan et al. (2010), as discussed on page 10, line 28 of the manuscript. While we saw good agreement (2.28%) in one case, we also found a difference of 20.88% in the other case.

From a sample size perspective, EARLINET is the authors' best bet for a comprehensive comparison. I encourage the authors' not to forgo this opportunity and go beyond only comparing eight nighttime overflights. I encourage the authors to also include daytime comparisons and a large enough sample size to make meaningful statistical comparisons.

The reason for limiting the EARLINET comparisons to these 8 nighttime cases was twofold. (1) The only aircraft CATS underflight data we have are 2 daytime cases with CPL, so we wanted to show some nighttime comparisons to an independent lidar measurement. (2) These EARLINET cases represent times that CATS was in close proximity to the ground sites and the ground lidars were operational. For more comparisons between CATS and EARLINET, we suggest checking out a new paper published in ACPD below.

Proestakis, E., et al. (2019), EARLINET evaluation of the CATS L2 aerosol backscatter coefficient product, *Atmos. Chem. Phys. Discuss.*, https://doi.org/10.5194/acp-2019-45.

For the CALIOP comparison, what is the motivation for comparing attenuated backscatter in cirrus? Since this study is concerned with calibration, why not compare CALIOP and CATS attenuated backscatter profiles as was done in the CPL and EARLINET comparisons? Adding the complication of cloud detection and multiple scattering into this seems unnecessary and out of scope for this study.

Yes, ideally, we would compare heavily averaged attenuated backscatter profiles in some high-altitude region where the aerosol loading is both low and temporally stable. Unfortunately, as clearly indicated by Figure 1 in the ATMD manuscript, the CALIOP 1064 nm backscatter signal is too noisy to support reliable comparisons of backscatter signals having very low aerosol scattering ratios. This low SNR is a consequence of the CALIOP 1064 nm detectors (avalanche photodiodes), which are plagued with a huge amount of dark noise. The magnitude of this dark noise is clearly evident by examining the CALIOP 1064 nm signal distributions around $\beta'_{1064} = 0$ in Figure 10 of the ATMD manuscript.

Aerosol scatting ratios

The text describing the scattering ratios and their presentation in Fig. 2 is confusing and in need of revisions. First, as Reviewer 1 points out, it is not really fair to call this a "molecular" normalization technique since, as Fig 2. shows, aerosol comprises anywhere from 30–50% of the signal you're calibrating! This is a huge challenge/limitation that the authors aren't very up front about (see my comments in the next section concerning this). Considering the need for these scattering ratios and their large contribution to the overall uncertainty, the authors need to be more precise in describing how they are incorporated into the algorithm and build confidence that these values are accurate.

We will address this comment in the next section.

It is unclear how exactly these scattering ratios are applied. On page 4 the authors state that “the CALIOP data is used to estimate the spatially and temporally varying 1064 nm scattering ratio at these altitudes (Fig. 2).” In Fig. 2 zonal mean scattering ratios are plotted in 4 different months. Are these zonal means

yes

what is meant by \spatially and temporally"?

Zonal means are inherently spatial averages. We describe the temporal averaging on page 5, lines 3–4 of the manuscript published in AMTD: “Every 15 days, the CATS team computes 30-day averages of the CALIOP 532 nm scattering ratios between 22 and 28 km.”

Why are only 4 months plotted in Fig. 2? What scattering ratios are used for the months not plotted? More specifics are need here.

The four months were chosen to illustrate the seasonal variability of these scattering ratios and thereby indicate that this seasonal variability is properly accounted for in the CATS calibrations.

Additionally, if zonal means are used, I would recommend putting standard deviations on the curves in Fig. 2 and discussing the amount of variability that is neglected by using mean values (this would need to be included in the error analysis as well).

The standard deviations of the 532 nm mean scattering ratios are quite small; for the January data shown in the paper (left panel of figure 2), the relative uncertainties (i.e., the standard error divided by the mean) for the 2° averages are all less than 0.2% (because the number of samples included in the averages is very large).

For Eqs. (2) and (3): why not just use the CALIOP 1064nm scattering ratios directly instead of assuming a backscatter color ratio?

Per our earlier response, the CALIOP 1064 nm SNR is too low to retrieve reliable estimates of high-altitude aerosol scattering ratios.

As I mentioned above, the uncertainty in the assumed backscatter color ratio is likely larger than just using 1064nm CALIOP data directly. Plus, the error in 1064nm CALIOP scattering ratios has already been characterized (*Vaughan et al.*, 2019) which would make the authors' uncertainty analysis more straightforward. Limb sounding instruments will, by far, give the highest accuracy scattering ratios at these altitudes. Did the authors explore any other alternatives to using CALIOP for getting the aerosol scattering ratios? At the very least, the CALIOP scattering ratios should be compared to the climatology of SAGE II, SAGE III, GOMOS, etc...

Limb sounding instruments do not provide scattering ratios since they don't measure aerosol backscatter, only extinction coefficient. While we could use extinction measurements and assume a lidar ratio, none of the instruments operational at the time of CATS launch provided measurements near the CATS 1064 nm wavelength. SAGE III started operating in June 2017, but data wasn't available until roughly a month before CATS stopped operating. Using SAGE III 1022 nm extinction measurements as a comparable to the CALIOP 532 nm scattering ratio method would make a wonderful follow-on study, but is far out of the scope of the current paper.

Don't oversell the approach
I've touched on this throughout my comments above. There are several statements throughout the paper that are misleading considering the large uncertainty in having a significant amount of aerosols in the calibration region and the reliance on CALIOP to account for this. The authors state in the abstract that "Overall, CATS has demonstrated that direct calibration of the 1064nm channel is possible from a space based lidar using the molecular normalization technique". But this statement is only half true because the CATS calibration relies on having another, already calibrated, lidar in space (CALIOP). You can't characterize this as a "direct calibration" if 30–50% of your calibration (i.e. Fig. 2) relies on inter-calibrating to CALIOP! In essence, the authors follow a similar approach as has been done in previous work: derive a 1064nm calibration from calibrated 532nm backscatter.

CATS is the first space-based lidar to use the Rayleigh normalization technique to directly calibrate a 1064 nm backscatter channel. This is a true first and an important milestone for the backscatter lidar community. It should not be understated just because we have chosen to incorporate independent measurements to better quantify aerosol loading.

LITE, GLAS, and CALIOP did not have the SNR to directly calibrate using the Rayleigh normalization technique at 1064 nm. In the past, this technique has always been called the Rayleigh or molecular normalization technique. However, these previous applications were at shorter wavelengths (355 nm and 532 nm) and in altitude regions with small aerosol contributions. Thus, we understand the comments from one of our referees about calling it a Rayleigh normalization technique. We have changed the phrases "Rayleigh/molecular normalization technique" to "atmospheric normalization technique" and "Rayleigh profile" to "Rayleigh profile corrected for aerosol contributions". A sentence on page 4, lines 16-19 defines this name.

There are several instances of the authors being cagey about this. For example, on page 3 lines 30-31: "CATS exhibits high nighttime 1064 nm SNR, enabling 1064 nm attenuated total backscatter (ATB) direct calibration without any dependence on the CATS 532 nm signal". This statement is very misleading. Yes, you have no dependence on the CATS 532 nm signal, but you do depend on the CALIOP 532 nm signal and, on top of that, a CALIOP 532 nm signal that is conveniently already calibrated for you. If the CATS 532 nm signal was of sufficient quality, you would have certainty used the CATS 532 nm signal instead! Another example: page 13 lines 24-29. Here the authors do say the aerosol loading is higher in the CATS calibration region than for CALIOP. Instead of waiting for the conclusion, the authors should mention this in the introduction and then again when discussing Fig. 2. Additionally, it is important to convey, quantitatively, the difference between the two: CALIOP has aerosol scattering ratios that are less than 1.02 in its calibration regions (Kar et al., 2018, their Figure 2b), CATS has values between 1.4-2.0 (Fig. 2). That is a very significant difference. In spite of all these difficulties, the general pathway the authors have taken to calibration CATS is likely the best approach. But, the authors need to chose their words carefully and convey that their approach is not going to be widely applicable to other 1064 nm lidars since they do not demonstrate an independent calibration of 1064 nm backscatter.

We have attempted to be upfront about the all of the limitations of our technique. For example, in the first paragraph of Section 2.1 we plainly state that "While this altitude region provides sufficient molecular scattering for the Rayleigh normalization technique, the aerosol loading in the lower stratosphere (22-26 km) is also higher than the 36-39 km region used to calibrate 532 nm CALIOP data. To improve the accuracy of the CATS nighttime calibration, the aerosol loading in the calibration region must be quantified, along with the ozone transmission profile, molecular backscatter profile, and polarization gain ratio (PGR)."

We do agree that choosing words carefully is important. We are reminded of this by the comment saying that "the CATS calibration relies on having another, already calibrated, lidar in space (CALIOP)", which is repeated in similar statements throughout this short comment. However, this assertion is simply not true; calibrating CATS does not rely on having another lidar in space. We could easily parameterize this aerosol loading based on a model (i.e., previous measurements and studies), a technique that is utilized routinely in the satellite remote sensing community, to calibrate CATS data with a larger uncertainty. Calibrating CATS 1064 nm data in the mid-stratosphere to < 10% uncertainties does rely on having another source of stratospheric aerosol loading measurements in space. Given the unanticipated demise of the CATS 532 nm channel, we were quite fortunate in being able to leverage the CALIOP data for this purpose.

Speculating on a way to truly do so is good fodder for the conclusion. The authors do some of this already, but I would encourage them to expand that discussion a bit. Have the authors considered the precision/accuracy trade offs between a lowered rep rate and increasing the altitude of the calibration range? If the CATS measurements weren't limited to between 51S-51N how would polar stratospheric clouds impact the calibration in polar regions?

A sentence was added to the second paragraph of the conclusion to elaborate on how decreasing the laser repetition rate of a future CATS-like backscatter lidar could provide a larger data frame, and thus a higher calibration altitude, based on the comments of both referees.

More minor comments/edits

The minor comments were taken into consideration and some of them were adapted in the latest version of the manuscript.