General Remarks

This manuscript presents an analysis of CALIOP aerosol measurements in the Arctic. It includes valuable discussion about the difficulties associated with the CALIOP detection threshold and the difference in sensitivity between daytime and nighttime, as well as strategies for dealing with these difficulties. The paper is well researched and clearly written.

Very recently a paper was published in ACP by Winker et al. with some overlap with this manuscript. The Winker et al. paper has a more thorough treatment of CALIPSO data filtering and gridding which is also relevant to this analysis, so it might be good for the authors to consult that paper before final publication. The present manuscript's unique contributions beyond those of Winker et al. include the specific focus on the Arctic, and comparisons with other instruments. Comparisons of CALIOP with other instruments, particularly in situ measurements, are very rare and always valuable. The idea of making a correction to convert CALIPSO daytime measurements to nighttime equivalent is clever and intriguing, but there remain significant difficulties in handling this challenging dataset.

Comments about the handling of detection thresholds

A major thrust of the paper is the calculation of a "nighttime equivalent" correction, but I'm not sure the paper is entirely convincing that this is the best way to handle the challenge of different measurement sensitivities by day and by night.

The primary difficulty of CALIOP having a non-negligible detection threshold is that when gridding to produce averages, some assumption needs to be made for cases where no aerosol was detected. It's usual to assume these cases have zero backscatter and extinction, but then the gridded averages are biased very low. The authors approach this problem with two different strategies (used together). First, they calculate a "nighttime equivalent" value in an attempt to address the fact that the daytime and nighttime detection thresholds are different. If I'm following the description of the technique correctly, then this correction essentially assumes that for aerosol between the daytime and nighttime detection thresholds, the extinction and backscatter are not zero, but rather equal to the average values for those aerosols that are detected. This seems to me that it must produce an overestimate of the correction, since the layers that missed are in reality more weakly scattering than those that are detected.

Since this correction is only used to scale up the backscatter and extinction as if they were detected in nighttime conditions, this still leaves the problem of a non-negligible nighttime detection threshold. To address this, the authors apply the CALIOP nighttime detection threshold to the correlative measurements. The averaged-up numbers being compared are generally below the CALIOP detection threshold (see page 4880, line 3), implying that there are a lot of zeros included in the statistics and making it harder to feel confident in the results.

Another approach to making comparisons with the gridded CALIPSO dataset is to recognize that the averages defined at lines 26-27 on page 4871 are lower bounds based on the assumption that any aerosols not detected by CALIOP have extinction and backscatter coefficients of 0. Have you considered also calculating an upper bound by assuming that the backscatter of cases reported as “clear air” is equal to the detection threshold? This would set a quantitative value on the uncertainty in the gridded product due to the detection threshold, and the in situ measurements could be compared directly to the envelope. I believe this would be a more direct comparison and the assumptions would be clearer and easier to follow and understand.

The nighttime equivalent correction is an intriguing idea but I have some specific difficulties with the discussion of Figure 3 which forms the crux of this correction.

1. When calculating the average backscatter of the measurements within a grid box, do you mix day and night? Since you have shown that day and night have systematically different backscatter, it would probably be better to use only daytime measurements for this figure and the calculation of the fit.

2. Does scaling up the detection frequency as described in Eqn 2 effectively assume that the undetected layers have the same mean backscatter as the detected layers? This would produce a high bias since the undetected layers are undetected because they fall below the detection threshold. I think Figure 2 shows that even if the fit were perfect, this "correction" would produce daytime values biased 10-15% too high.

3. Are you correcting what is essentially an additive error (layers that are not detected) using a multiplicative correction (scaling up the detection frequency). Does that seem right?

4. I am surprised that there is such a very large difference between daytime and nighttime even at the largest backscatter values. Wouldn’t it be expected that daytime measurements of highly scattering aerosol would be just as good as nighttime observations, as long as the backscattering is well above the detection threshold?

5. Pg 4875, lines 26-27. The agreement between the nighttime and nighttime-equivalent detection frequencies is not really a test of the appropriateness of this correction; I think it’s really only an indication that the linear fit is a reasonably good fit.

Comments about the comparisons with other instruments

The authors claim that the comparisons validate this approach (Page 4876, “We examine the validity of this empirical approach by comparing”), but it seems that the analysis is really focused on validating CALIPSO data (which is certainly of interest) and not designed to adequately validate the correction approach specifically. To do that, the comparison in Figure 5 should separate the day and night measurements so the effect of the correction is clear. The HSRL comparison does not even show a comparison without the correction, so is not really appropriate to the point of validating the correction. Without separating out the effect of the correction, there is little basis for a reader to judge how much improvement the correction makes.

While the statement about “the maximum temporal offset between in-situ observation and satellite overpasses” on page 4877 implies that the comparisons are for matched coincident cases, later discussion makes me believe that this is not the case. I think the comparisons with other instruments are all averaged (monthly means over several years and means over all the campaign flights within a large geographical area) with no attempt to do event-by-event matchups. Since several flights during ARCTAS were along the A-train orbit track I would have guessed that more specific matchups would be possible for the aircraft measurements. I would also think it would be possible to do one-to-one matchups with ground-based measurements. It seems that one-to-one matchups would produce more rigorous comparisons, and ameliorate the need for filtering out episodic large values (which is unfortunate, since as you say on 4879 “CALIOP would only be able to detect the strongest haze events”). Was there any attempt to study coincident matchups?

Comments about Diamond Dust

I’m not quite convinced that the CALIOP cases discussed on page 4873 having large aerosol extinctions but negligible depolarization are likely to be diamond dust. Can you expand the discussion? The combination of large backscatter and small depolarization could occur for horizontally oriented ice (HOI) crystals, which as you say may be a factor in the HSRL record discussed on page 4881, but is probably not a significant factor in the CALIOP record due to the tilt of the laser. Other than HOI, I’m not familiar with cases of ice crystals having large backscatter and small depolarization. I think I understand from your discussion that you refer to airmasses with a small density of ice crystals that would affect the bulk backscatter coefficient but not the bulk depolarization ratio. Bourdages et al. (2009) appear to be referring to radar, which otherwise would not be very sensitive to aerosol, so the relative contribution to backscattering by a few large crystals would be considerably more than for lidar. I see that Hoff (1988) points out that a relatively few ice crystals can contribute a measurable amount of scattering, but even that study uses depolarization as an indicator of the presence of ice crystals, which you imply is not practical for these cases. He does a simple calculation to show the amount of scattering for certain number densities of ice crystals of certain sizes. Can you expand that calculation to estimate the measured particle depolarization for the same cases?

In the discussion of diamond dust with respect to the HSRL measurements on page 4880, depolarization is used as an indicator of diamond dust, inconsistent with the handling of potential diamond dust in the CALIOP data. I think it would be better to handle the data filtering more similiarly, to aid with comparisons. However, HSRL and CALIPSO probably do differ in the relative importance of HOI since HSRL is vertically aligned. HSRL measures lidar ratio which is also affected by HOI. Can’t the lidar ratio be used to filter for HOI?

Other Specific Comments

Pg 4869, line 22-23. This sentence implies that Devesthale et al. (2010) found that 65% of the AOD occurs below 1 km, but I believe they were actually counting discrete layers (that is, 65% of aerosol layers, not 65 % of the AOD). I could be wrong, but please check.

4870, 19-20. The description of the CALIPSO algorithm papers is not quite right. Liu et al. (2009) describe the separation (and assessment) of clouds and aerosol, while Omar et al (2009) describe the subdivision of aerosol into different aerosol types. (Cloud phase classification is in another separate paper.)

4870, 21-23. “iteratively … until convergence”. In most cases, there is no iteration in the assignment of lidar ratio for CALIOP. There are two exceptions. In one type of case the AOD of a lofted layer can be determined via the transmittance, so the inversion can be done without prescribing a lidar ratio. Since the aerosol classification is not needed for this kind of case, it’s not described in Omar et al. (2009) in detail, and probably not what you mean. It’s also quite rare for aerosol. The other exception is cases where the prescribed lidar ratio produces an unstable vertical inversion and is consequently adjusted until the inversion is stable. You are correctly filtering out these cases (on page 4872, you say “where the retrieval algorithm had to adjust the initially selected lidar ratio”) so they are also irrelevant. It would probably be better to strike out the reference to iteration here.

4870, 26. Change to “NASA Langley airborne High Spectral Resolution Lidar”. Since the HSRL used by Rogers et al (2011) to validate CALIPSO is not the same HSRL instrument used in the current manuscript, nor the same research group, it’s better to be specific.

4871, 20ff. Pre-gridded Level 3 products were released in December 2011. Why not use these?

4871, 23. “Fraction of detected aerosol layers” could mean either the number of detected aerosol layers over the number of aerosol layers that are undetected (which is unknowable) or it could mean the number of detected layers over the number of cases with no aerosol detection (i.e. faint aerosol plus true clear air). Can you reword the discussion to make it clearer that you mean the latter?

4872, 3. FYI, Winker et al (2013) recommends a more permissive CAD filter of 20.

4874, 22. “We compare the daily average RH…”. Where is this comparison? What is the result?

4878, 6-7. “difference … is small because of the relatively high values of extinctions”. I’m confused by this. According to line 25 on 4875, the scaling factor ranges from approximately 1.6 to 6.2. So even at high values of extinction, the nighttime equivalent would be 60% larger than the daytime value. So small differences must be primarily due to a large proportion of nighttime measurements in the average, rather than large extinctions, right?

4881, and Figure 8. Since the in situ measures extinction, not backscatter, and since HSRL also measures extinction and CALIOP reports extinction (although in that case backscatter is a more fundamental measurement), why not make comparisons of extinction here instead of (or in addition to) backscatter?

Technical Suggestions

Pg 4868, lines 20-23. Consider rewording this sentence. I’m not grasping the significance of the word “although” here.

4869, 12. “Active remote sensing” should be “passive remote sensing.”

4871, 2. "Was found"

4872, 24. Extra word. Delete the first "method"

4875, 23. Missing period after "segments".

4877, 11 and 17. Gasso should have an accent on the o.

4879, 7. "corresponding to flights around Barrow and Fairbanks". What does this sentence clause signify? Both Barrow and Fairbanks are in the AK box, not the CAR box, and both the DC8 and P3 were based in Fairbanks.

4879, 11-12, "along track measurements" would be clearer if changed to "measurements along the flight track".

4884, 8, "produces" should be "produced"

Figure 3. There is a sort of "rastering" or "moiré pattern" that causes some of the tick marks on both the x and y axes to disappear, with the pattern of which ticks disappear changing as the browser window is resized. It might be helpful to use thicker lines for the axes and tickmarks or change the file type of the submitted figure.

Figure 5. In this layout, the figure annotations are much smaller than the text size. Please consider increasing the text size in the figure or requesting a full-page layout for this figure.

Figure 7. Labels probably should be bigger in this figure also.

Figure 10. The change in scale on the color bars is a little bit confusing. Maybe you could at least make a note of it in the caption.