Responses to Referee # 2

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.

At the time of writing we were not aware of the Winker et al. (2013) paper. In the revised version of our manuscript, we now refer to their work. In our work we tackle some specific issues of retrievals over the Arctic and believe these two manuscripts complement each other. Our CALIPSO data filtering is very similar to the one used by Winker et al. (2013), with some minor differences:

1. *CAD Score. Winker et al. use -20, while we use -50. Differences between these two strategies are discussed later in our reply.*
2. *Ext_QC flags: both this manuscript and Winker et al. (2013) allow for QC flags of 0, 1. Winker et al. also allow for opaque layers (whether or not the lidar ratio is changed during solution, i.e. QC=16,18) but these occurrences are extremely rare over the Arctic, given both the weakly scattering nature of Arctic aerosols and our strategy to not consider profiles where signal attenuation by clouds occur. Thus our strategy is virtually identical to Winker et al.'s.*
3. *Uncertainty flag: we use the same strategy and filter out layers that have an uncertainty value of 99.9 km-1.*
4. *We do not filter out aerosols adjacent to clouds above 4 km.*
5. *We do not filter out noise spikes.*
6. *We filter out negative extinctions (Winker et al. filter out extinctions < -0.2 km-1)*

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 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.

The reviewer describes very well the inherent problem of dealing with the CALIOP undetected layers and generating gridded extinction values. This problem is particularly pronounced in the Arctic, where the aerosol loading is low. We indeed assume that non-detected layers have zero backscatter and extinction for the nighttime retrievals. For the daytime retrievals we assume that aerosols layers between the daytime and nighttime detection thresholds have the same extinction as the layers that are detected. This is our "nighttime-equivalent" extinction. This does indeed leads to an overestimate of the correction, but as described below this overestimate is small (10-15%) compared to large values of the scaling factors (100-300%). While inherently uncertain, this approach allows us to reconstruct the seasonal cycle of Arctic aerosols as retrieved by CALIOP. As pointed out by the reviewer, our nighttime-equivalent still assumes zero extinctions for undetected layers below the nighttime threshold and we address this by applying the same assumption to the in situ observations.

We agree in the remark that a mean backscatter below CALIOP's detection threshold does indicate the presence of numerous zeros in the statistics. However, that is an inherent characteristic of CALIPSO retrievals over the Arctic, particularly during summer and autumn, or at higher altitude, and can not be avoided.

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 compare directly to the envelope. I believe this would be a more direct comparison and the assumptions would be clearer and easier to understand.

This is a good suggestion. The revised Figure 5 (comparing CALIOP retrievals to in situ observations at Barrow and Alert) shows the upper bound and lower bound for the gridded CALIOP mean extinction as a grey envelope. The upper bound assumes that "clear air" has backscatter equal to the detection *threshold, while the lower bound assumes zero backscatter (which we refer to as "standard extinction" in the manuscript). Neither one takes into account the correction for day-night. No threshold is applied to the in-situ observations. The nighttime-equivalent is quite close to the lower bound. The Figure shows that the in situ observations fall within the envelope of the CALIOP upper and lower bounds. It also demonstrates that in summer the upper bound overestimates in situ observations by a factor of 5-10. The seasonal cycle of the upper bound is also too weak compared to observations. In situ observations are much closer to the lower bound, showing that the true extinction of undetected layers is much below the sensitivity threshold. We have included a discussion of this in the revised manuscript.*

We tried applying the same approach of estimating an upper bound in the free troposphere, where aerosol detection frequency is very low (often below 2%, see Figure 1), but found this approach less informative since the upper bound is entirely dominated by the assumption of non-zero backscatter and can lead to a seasonal cycle opposite to what is observed by independent techniques.

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 use only daytime measurements for this figure and the calculation of the fit.

Yes, we use both daytime and nighttime retrievals in calculating the mean backscatter of the detected aerosol layers. We initially used the daytime backscatter in Figure 3, but found that it led to greater scatter and less robust statistics. The overall slope of the relationship remained the same though. In order to obtain a more robust statistical relationship, we thus decided to combine the backscatter of both day and night aerosol layers.

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.

We did show in Figure 2 that daytime values are expected to be 10-15% too high relative to nighttime values. This is an upper bound for the overestimate affecting the late spring and summer seasons when only daytime retrievals exist. In other seasons, we calculate the overestimate to be be less than 7%. Compared to the magnitude of the scaling we apply to the detection frequency (x2-5) the magnitude of this error is small. A paragraph has been added to the text at the end of section 2.3.

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?

Although not ideal, given the limitations of the instrument under Arctic-specific conditions we believe this is a good effort at capturing the pan-arctic seasonal cycle of aerosol extinction. Additive corrections such as making various assumptions on the extinction of the undetected layers are not well suited to the Arctic free troposphere under daytime conditions. As noted above, the results become too dependent on the arbitrary assumptions made as to the extinction of the undetected layers and the seasonal cycle is not captured. This is why we chose our empirical approach based on quantifying this multiplicative correction directly from CALIOP observations of the same environment during both daytime and nighttime (Figure 2 and 3).

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?

Yes, CALIOP retrievals of very dense plumes with very high values of backscatter should yield the same values. Figure 3 represents monthly mean backscatter and detection frequencies for a gridded average, which represent a distribution of aerosol extinctions. For moderately dense plumes a certain fraction of the aerosol layers will be between the daytime and nighttime detection thresholds, leading to a ratio less than unity.

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.

This is correct, we have noted that in the revised version of the manuscript.

Comments about the comparison 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.

The reviewer is correct. The section on comparisons to other instruments examines both validating the CALIOPSO data, but also our day-night correction. The revised manuscript reflects this more accurately. The direct validation of CALIOP is done via the comparison to Barrow and Alert observations (Figure 5), while the impact of the correction in the nighttime equivalent CALIOP extinctions is emphasized in the comparison to ARCTAS observations (Figure 7). In order to address the reviewer's point about the HSRL observations, we have revised Figure 8 to clearly show the impact of the correction.

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 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?

The statement on page 4877 refers to the comparison to the Barrow and Alert ground based nephelometers. Regarding the comparison with ARCTAS, we are aware that a few of the flight paths were designed to validate CALIPSO using highly spatially and temporally coincident matchups. Coincidence is better suited to validate CALIOP products, which has been done extensively with the Langley AHSRL, than to test our correction method. Our emphasis is on the validation of the correction method by using the vertical profiles of extinction, and requires as many points as possible to be tested effectively. In performing our comparisons shown in Figure 7, we consider CALIOP data only for the days when the aircrafts were flown. In this case, spatial averages within the AK and CAR domains are computed with a temporal resolution of 1 day and we do not claim they are one-to-one 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 air masses 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 then 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?

We had initially attempted to use both depolarization ratio and pseudo color ratio alongside extinction to separate ice crystals and aerosols. We plotted various combinations of 2-D distributions of these three variables and found a continuum rather than 2 separate populations. A linear depolarization ratio of 0.1 is often used as a threshold to indicate the presence of ice crystals (Hoff, 1988). However the majority of *aerosol layers with anomalously high extinction had depolarization below 0.1 and setting a threshold on depolarization would have eliminated legitimate aerosols layers. Using a threshold for the pseudo color ratio would also eliminate coarse mode sea salt over the north Atlantic where values are significantly higher than elsewhere in the Arctic. Thus we opted for a simple approach and chose to cap the maximum extinction allowed using a conservative value more typical of highly polluted airmasses, and use that as a proxy for ice particle contamination (diamond dust or other). Expanding the calculation done by Hoff (1988) is beyond the scope of our work. We corrected the reference to Bourdages et al. who refer to the radar backscatter of aerosols mixed with ice crystals.*

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 similarly 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?

The depolarization threshold used for HSRL is intended to filter out clouds and precipitation from the observed scene. Unlike CALIOP, HSRL lacks a scene classification algorithm and arbitrary criteria must be established to separate out Arctic Haze from clouds and precipitation. Thus we chose a depolarization threshold of 0.1 consistent with the choice of Hoff (1988) although the results did not vary significantly (±5%) using thresholds in the range (0.05-0.12)

Unlike the Langley HSRL, the HSRL operated at Eureka is optimized for the study of Arctic clouds rather than aerosols. The HRSL is tilted 4˚ off the zenith to avoid reflection from HOI. This is now mentioned in the revised text. Furthermore, while the backscatter measurement obtained with the Eureka HSRL is robust, the extinction (and hence the lidar ratio) is very noisy because of the very small field-of-view of the receiver (100 micro-radians) leading to difficuties in measuring optical depths less than ~0.1 at ranges less than ~2 km. This prevented us from using the lidar ratio for comparison with CALIOP's to help discriminating between aerosols and diamond dust.

Other Specific Comments

Pg. 4869, line 22-23. This sentence implies that Devasthale 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.

The reviewer is right, Devasthale et al. (2010) refer to the number of aerosol layers. We corrected the text

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.)

We revised the whole paragraph as follows:

"Liu et al. (2009) examine the part of the algorithm that separates clouds from aerosols and assesses its performance. The algorithm that classifies the aerosol layers into six aerosol types (clean continental, clean marine, dust, polluted continental, polluted dust and smoke) is described by Omar et al. (2009). The inversion to obtain the AOD at both wavelengths is then initiated by assigning an extinction-to-backscatter ratio (called lidar ratio), as a function of the aerosol type. The resulting data product is called Level-2."

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.

Done

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.

Done

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

Given the amount of manipulation that was necessary to this study, and specifically the diamond dust filtering method and the aerosol layer base extension described in Koffi et al. (2012), we preferred to work with the Level 2 products which gave us more flexibility than gridded data. We tested the gridding routines used in this work and we were able to exactly reproduce the vertical profiles of Yu et al. (2010) over all regions considered in that study when we made the same choices regarding data filtering and processing .

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?

We ensured the text contains a clear definition:

"For each grid-box we calculate the aerosol detection frequency (f), which is the ratio of the number of detected layers over the sum of detected layers and clear air."

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

The difference between using a CAD filter of 20 and 50 is very small as shown in Figure 1 of Yu et al. (2010). Our choice result in a marginally larger number of discarded layers than using CAD=-20 while at the same time maintaining a greater confidence in the aerosol/cloud classification.

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

A whole sentence was missing from the typeset version (highlighted in green below). We apologize for the oversight and we will make sure it will be included in the final version.

"In order to exclude the possibility that a diurnal cycle in relative humidity (RH) drives the difference in extinction between daytime and nighttime retrievals, we compare the daily average RH for descending (daytime) and ascending (nighttime) for the same spatial region and temporal period for one year of data (2006-07), using the retrievals from the NASA Atmospheric Infrared Sounder (AIRS) on board the A-train AQUA satellite. The two populations differ in the mean by RH=3% (daytime higher). We find that the random variable representing the standardized difference of the two means falls within one standard deviation from zero, indicating that the two populations can be considered statistically indistinguishable."

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?

That is correct. We revised the sentence to:

We find that there is little difference between the CALIOP standard extinctions (Barrow: 9.9±11 Mm-1; Alert: 4.6±5.3 Mm-1) and nighttime-equivalent extinctions (Barrow: 11±11 Mm-1; Alert: 5.0±5.2 Mm-1) because the annual mean is dominated by the winter maximum which occurs under nighttime conditions.

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?

For the reasons discussed earlier, we can not consider the extinction measured by HSRL reliable. Therefore, we decided to perform the comparison with CALIOP in terms of backscatter only, which is a more fundamental measurement and is measured with high sensitivity by the HSRL system.

Technical Suggestions

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

We have rephrased the sentence.

4869, 12. "Active remote sensing" should be "passive remote sensing"

Corrected

4871, 2. “Was found”

We corrected the clause inconsistency replacing the sentence with the new sentence: “No seasonal, latitudinal or vertical dependencies were found.”

4872, 24. Extra word. Delete the first “method”

Corrected

4875, 23. Missing period after “segments”.

This is a typesetting typo. We’ll make sure it will be corrected

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

Corrected

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.

That part of the sentence has been deleted.

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

Corrected

4884, 8, “produces” should be “produced”

Corrected

Figure 3. There is a sort of “rastering” or “moire 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.

We thickened tickmarks and axes and increased the resolution of the 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.

We increased the text size for both legend and figure annotations and also replaced this figure with a different one according to the reviewer’s suggestions.

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

We increased the size of all text.

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.

Using a common color scale is not advantageous so we added the following note to the caption: "Note that the color scale saturation values are lowered at higher altitudes."