Author (Daniel Letros) responses to the referee comments is shown in blue.

We would like to extend our appreciation to you for your efforts in giving us this feedback. It has made the presentation of our work better. Thank you!

General comments on Letros et al. [2025]:

This is a well-executed paper that presents analysis of data taken by the ALI instrument, which records polarized radiance spectra. The value of these measurements is demonstrated through both sensitivity studies and analysis of selected ALI data. A particularly useful capability provided by radiance information is the ability to distinguish between clouds and aerosols along the line of sight.

I have few criticisms of the work presented, and particularly appreciate the degree of testing presented for the algorithm approach taken (which is clearly designed to test its limits, rather than primarily to show its performance in the best possible light). Error analysis is thoughtful and thorough. Some structural suggestions and clarification questions follow, but I have a positive opinion of the paper, and would be glad to see it published.

Detailed comments:

Abstract, line 4: "Tangent altitudes which have signal contaminated by clouds" should be defined more clearly. Throughout the paper, similar wording is used, and the authors' interpretation of this language appears to be consistent: This language is meant to describe tangent altitudes for which a cloud appears along the line of sight. But that's not the only possible interpretation: For example, limb scattering observations can (and frequently are) "contaminated" by upwelling radiation from a broad area below the line of sight, for which the locations and properties of the clouds are not well known. That is a distinct problem from interpreting measurements with clouds along the line of sight, and it isn't the focus of this paper, which the abstract should make clearer.

Yes you are right and our wording should be clearer. In general we are pursuing a quantitative distinction of lines of sight that one would not want to include in an aerosol retrieval - which at a minimum are lines of sight looking through clouds. However, we are aware of the upwelling cloud contamination in practice. To combat this for the flight retrieval efforts in the paper, we limit the retrieval when the radiance profiles see a distinct change in DoP - which (as Figure 7 shows) can be higher than the actual height of the cloud. Exactly how robust of a strategy that is to avoid *all* cloud contamination remains uncertain to us for now but we think it is very fair to say it is not full proof.

We have attempted to clarify the definition of "contaminated by clouds" to be related to the depolarization in the abstract as well as elsewhere in the paper.

Abstract, lines 14-15:+ The abstract ends by declaring "good agreement" between the aerosol extinction coefficients derived here and coincident data (from SAGE III, OMPS and OSIRIS). It would be more useful to make this statement quantitative, or at least indicate what criteria were used to conclude that the agreement is "good." quantified wording.

The statement about particle size information would also benefit from clarification: A short summary of the particle size distribution assumptions made and the relative quality of the various retrieved properties would help a reader who is reading the abstract before deciding whether to proceed further. Added to the abstract concerning the new SAGE particle size comparison. We believe it addresses this concern.

Line 47: "purposes of the SPS is" – should be "are" Corrected.

Figure 2: The spectral response functions shown in parts (a) and (b) appear to be fairly Gaussian – it would be useful to include an estimate of the FWHM (or some other indication of their width). The caption

mentions that the FWHM defines the spectral resolution, but if those values are stated anywhere, I missed them.

The spectral resolution shown in Figure 2(d) is directly the FWHM of the spectral response functions. We note this in the text of the first paragraph of Section 2.1 - however we understand your confusion. In the text we call it spectral bandpass and not resolution as (d) is labelled. We have clarified the wording of the text and also noted the FWHM relation in the caption of Figure 2(d).

Line 93: This paragraph should conclude with a reference to the later sections in which the non-ideal response of the ALI significantly affects the analysis (to motivate the detailed discussion of this feature, and also to guide a reader who is particularly interested in this aspect of the work). Very good thought. Implemented.

Lines 101-102: "external source each pixel measured" should be "external source for each pixel measured." Corrected.

Lines 103-104: Here and elsewhere, the set of images used to characterize the dark behavior is called "large" – how large? And how was the most appropriate large number chosen?

Of course, to characterize the dark (or any other) behaviour, the more images collected the better. Generally we had no specific chosen number of images to collect for any test, and the dark characterization for instance had a variance in the number of images collected at each exposure time. The efficacy of our calibrations (which included if we needed more image collection) was determined by analysis. I.e. the dark calibration was validated by looking at the histograms of randomly selected dark images and seeing the deviation of the mean from zero. Additionally, Figure 4 of the paper exemplifies our evaluation of integrating sphere calibrations.

However, to loosely quantify "large" here we have noted hundreds of images at line 103-104 - which was typical of our calibration work.

Line 111: "Bad pixels" are mentioned in this paragraph (and defined in the next). Later (in Figure 4), some visual evidence is provided that the number of "bad pixels" is a small fraction of the total image, but how small is it (as a fraction of the total number of pixels)?

We consider 7289 pixels out of the 327680 total pixels (2.22%) to be non-ideal and exclude them from any scientific analysis.

Lines 138-139: "Out-of-field" stray light is a serious problem for many limb scattering sensors (usually more serious than "internal" stray light). Is that true for ALI as well? The text says that this is "mitigated with careful baffling" – can you say more about how well this mitigation works, and how its effectiveness was assessed? (If some of these instrument-related questions are addressed more completely elsewhere, then adding a specific reference would be helpful – and I particularly apologize for not reviewing Letros et al., 2024, due to lack of access!)

This is an excellent question, and we do not have this granularity of the ALI design published anywhere currently. The short answer is we do not think out-of-field stray light plays a significant role in ALI performance (although it is of course always a concern of any limb instrument). Our confidence for this comes from analytical study of the design (with commercial optical software) and the lack of any significant evidence of it seen in lab conditions. Furthermore - although obviously not controlled conditions for analysis - we have seen no evidence to suspect it being in the flight data either.

To elaborate a little further, stray light was studied in the lab in two ways: (1) by systematically moving our integrating sphere all around the ALI optical enclosure and imaging with the AOTF off, and (2) rotating the ALI field of view in front of a collimator and imaging with the AOTF on. Test (1) is an excellent test for general light leaks like from gaps in the optical housing. After correction of these images to measure the stray light we observed signal (at worse) on the order of tens of DN. These images would otherwise well saturate the detector (16k+ DN) if the AOTF were on and ALI was looking at the source. However, (1) is not a good test for out-of-field stray as the field angles of the source are not well controlled and the AOTF is off. Test (2) accomplishes this, and as we systematically rotated the collimated beam out of the intended ALI field of view we saw dark images - indicating the baffling was working as designed.

However for full transparency, physical limitations of our lab equipment made rotating ALI in front of the collimator very impractical for steeper angles (i.e. $> 5°$ outside of our field of view) - and unfortunately the limitation is in the vertical (altitude) direction which is of course more important to the science. It is at this stage of evaluation that our confidence relies on the analytical optical design and not on specific testing. Hence, we will always have trouble giving a definitive "no we do not have stray light" but, again, we do not consider it an issue with ALI.

Table 1: I recommend including the estimated albedo for each scan as another column of data. That would make it immediately clear why Scan 3 includes more highly polarized observations than the others (lack of underlying clouds, which suppresses the amount of multiple scattering present).

We refrained from that initially because when Table 1 is provided the albedo is not discussed yet. However, we do think Table 1 can be changed to better provide information about the three scans with the albedo being one aspect. We have altered Table 1 to include more relevant information to the context of the aerosol retrievals (as opposed to the gondola).

Line 155: This paragraph says that the scans presented represent "reasonably nominal conditions" for ALI. Besides the gondola being "relatively stable," how are "reasonably nominal conditions" defined? I assume that the statement about stability refers to the attitude of the sensor – how is this assessed? And do variations in pointing during the integration time of the measurements contribute significantly to the effective altitude resolution of the measurement (or affect measurement quality in other ways)?

Largely, "reasonably nominal conditions" here means the stable attitude of the gondola. Although as we allude to in this paragraph and discuss later in the paper, we also mean to imply atmospheric DoP conditions. As for the attitude, it is measured by IMUs which allows the reconstruction of ALI lines of sight post-flight for the retrievals. The main issue (which you also noted) is how the gondola may change attitude during an image exposure which will cause a loss of spatial resolution. We select the scans we do with this in mind to avoid scans which would have significant spatial blurring.

We refrain from discussion around the gondola attitude and flight reconstruction within the paper for brevity, as we hope to maintain scope around the retrieval algorithm and ALI science demonstration. However, for other readers we have now appended to the first paragraph of Section 3:

[...] At this point in the flight the gondola was steered to maintain a solar azimuth angle (SAA) of 60°. During the flight the gondola position and orientation is recorded which allows reconstruction of all the ALI lines of sight in each image.

... and these in the third paragraph of Section 3:

[...] The balloon gondola was relatively stable for this scan compared to most others taken during the flight (gondola attitude within IMU error during exposures), [...]

[...] The other two scans we select present more difficult observation conditions both in terms of relative gondola stability (although there was still minimal attitude change during exposures, i.e $< 0.1°$ change in pitch over each image acquisition) and observational conditions.

Lines 184-185: The definition of the D matrix should be clarified. Would it be fair to call this a Tikhonov regularization term? Stating that it "restricts elements of the optimization from changing too much" should be reworded - I would describe it as being meant to retard the change in the state vector from one iteration to the next in the early phases of the retrieval. (This is reasonable in a Rodgers-type optimal estimation scheme for multiple reasons: An a-priori profile that differs greatly from the true profile may cause the initial retrieval step to move much too far, or in the wrong direction, for example.)

We agree completely with your feedback here and do like your description of retarding change better than what we had used in the paper. We have adjusted the sentence in the paper to now use your wording.

Line 208: "... as discussed later" – this should include a reference to the section where the material appears. Reference to Section 4.3 added.

Figure 6: The text provided in Section 4.1 suggests that the albedo estimation process used in this work is new (or at least significantly modified, relative to earlier approaches). But many of the details of the algorithm used are described in the figure caption. This makes the figure caption too long, and fills it with information that isn't directly related to the figure, without providing enough detail about the method itself. I recommend moving most of this caption into the regular text of the paper, and perhaps expanding it (or adding references) to clarify how it works.

We have moved much of the caption into the text in an attempt to make this section more elegant.

Line 226: "... atmosphere of known true state albedo of 0.6." What is the "true state albedo" – the actual albedo of the underlying surface? An effective reflectivity (combining the influences of the surface, clouds, aerosols, etc.)? Or something else? This is related to the following note.

We understand your confusion here and we should be more clear. The "true state" albedo of 0.6 is meant to indicate the known surface albedo of the simulated atmosphere (the atmosphere we are estimating the albedo of). This atmosphere also includes GloSSAC aerosol (which has no aerosol in the high tangent altitude region we are matching). However, the estimation is meant to quantify your latter definition of effective reflectivity.

To address this confusion we changed mentions of "true state albedo" to "true state surface albedo".

Line 228: "... 0.654 was found, which we consider to be a reasonable estimation of the 0.6 true state." My initial reaction is that this agreement (worse than 0.05, or nearly 10% relative error) is not especially good... but maybe that's unfair. Does the simulated data used in this example include noise, biases, etc. that are meant to mimic the performance of ALI? Does the stated value (0.654) represent an "effective reflectivity" that cannot be expected to perfectly match the "true" value (0.6), for the reason noted in the line 226 note (or for some other reason)? And maybe most importantly, how much does an error in retrieved albedo matter for the retrieval of the aerosol properties? (As noted elsewhere in the text, tangent height normalization often significantly reduces the sensitivity of limb scattering retrievals to uncertainty about the brightness of the scene.)

Yes, the simulation is meant to mimic ALI in practice (i.e. the simulated albedo estimation includes noise, bias, etc). Also (as we mentioned in the point above) there is aerosol in the simulated atmosphere, and we are using a Rayleigh (no aerosol) atmosphere to match the up-welling radiation in the 33-34 km tangent altitude range. Given this we would expect the effective albedo we estimate to be higher than just the surface albedo of 0.6.

I have added more robust detail about these points in the last paragraph of Section 4.1.

As for the important aspect - the impact on aerosol retrieval. *Without* high altitude normalization an incorrect albedo will of course introduce a bias in your retrieved aerosol. If you have too low of an albedo, then retrieved aerosol will be biased high to make up for the missing scattered signal. Too high albedo, and the retrieved aerosol will be biased low. *With* normalization, we found in simulation that affect of the albedo is essentially nullified. There are two caveats to that statement though: one) in simulations the physics of the forward model perfectly match (i.e. $\vec{b}'$ of Equation 3.4 in (Rodgers, 2000) is null) and two) in these simulations the aerosol was actually zero within the normalization altitude range.

This second point is the big one which is a limitation of our approach from a balloon platform. While we estimate the albedo to give the retrievals the best fighting chance, we found in practice that a normalization approach was best for the measurements Timmins 2022 flight. We don't consider this surprising as broadly speaking $\vec{b}'$ is not null and normalization *helps* to deal with this. Now being within the stratosphere (unlike a satellite instrument) there is no guarantee that the aerosol is zero at the altitudes we are normalizing by. So $\vec{y}$ is constructed relative to the (potential) aerosol signal within the normalization altitudes. In our approach we make the implicit assumption that it is zero here (as we noted in the paper), but if this assumption is incorrect then the retrieved aerosol will still be biased. If it is biased with respect to other instruments (which we admit ALI seems to be with respect to SAGE, OMPS, and OSIRIS) it is difficult for us to say if it is normalization or another effect, like the bimodal particle size we explore.

To more robustly acknowledge these aspects in the paper, we moved the statement beginning at line 479 (which was meant to make a general acknowledgement that we have limitations) closer to Fig. 15 and increased the detail to include many of the points above.

Lines 237-238: This sentence contains a particularly confusing reference to cloud effects: "... relative changes in polarized light can be used as a metric to determine if limb-scattered signal was influenced by cloud or not." Are we talking about the "influence" of clouds in the underlying scene here? That's my interpretation, but maybe the statement also applies to line-of-sight clouds? This should be clarified.

We believe We address this with our response and action to your first comment (*Abstract, line 4*)

Line 268: The reference "Bass and et al, 2010" appears here. That appears to be the only reference that uses "et al." rather than listing the full author list – why? And calling it "and et al." seems redundant.

The "and et al." was our mistake in formatting the citation. It is corrected now to just be "Bass et al.". As for the use of "et al." we are happy to make adjustments according to type setting before complete publication.

Line 280: "standard divinations" should be "deviations."(Corrected) And it would be helpful to explain what "prototyping the algorithm" means(defined). I assume this involved experimenting with various settings – you settled on these particular settings for some reason, but how much did the particular selections that you made matter, in the end?

Our goal in prototyping the retrieval was to get the correct ballpark settings - in this case the settings of the $\mathbf{S_a}$ matrix. The metric to evaluate the settings was simply if it performed well in simulated exercises with varying true aerosol states. Quantifying how these particular settings matter would be a time intensive study. Qualitatively however, naturally the a-priori covariance will dictate how much certainty to put into the a-priori state. The smaller these covariances are the more iterations the retrieval will need to statistically move away from the a-priori state to one which may better describe $\vec{y}$. Of course, incorrect underlying statistics of the retrieval in general may lead the solution to a false minimum or a very wrong and diverging state. We fully admit that our current algorithm as it is tuned (with regards to settings like the $\mathbf{S_a}$ matrix) may not be strictly optimal.

Line 288: "... the θ state is reasonably insensitive" – this should be quantified.

Now quantified with $< 0.5°$.

Line 310: Here (and in the caption of Figure 7), the text refers to a "stark" change in DoP behavior (for cloud identification). The figure provides some visual demonstrations that these can appear very obviously in the profile, but how "stark" must a change be to trigger identification of a feature as a cloud?

The edge detection algorithm has freedom to be tuned (i.e. the amount of smoothing applied or restriction in altitude to even consider a cloud being present) which will adjust what it can detect. Generally though, the change in DoP of the cloud just needs to be distinguishable from the background variance of the DoP profile. So for example in the case of Figure 7 (i), the change caused by the cloud would need to be larger than the background changes (resulting from the noise) in the higher altitudes.

Line 313: "... known form the true state" – should be "from." Corrected.

Lines 375-376: "... variance of the median radius is ... selected to be 0.01." and "The scalar width has this variance set to 0.0001." Shouldn't these values have units? I may be confused about what these definitions mean... but again, how much do they matter in the observed behavior of the retrievals?

Yes, we should note some units. The variance of the radius should have units of μm^2(now corrected), and the number density in Table 3 should be noted with cm^{-6}. The width is unitless. As for how they matter, we refer to our response to your *Line 280* comment.

Table 3: How were these listed number density variance values selected, and how do they affect the retrieval? As mentioned earlier, I'm particularly curious how the solution might be affected when the a-priori number density profile differs from the true atmosphere by a factor of 10 or more.

In regards to the covariance, we will again refer to our response to your *Line 280* comment. However, to your particular comment of a-priori profiles we have added content in the paper around a comparison with the SAGE III particle size data which address that question.

Lines 405-406: "However, of note this a-priori N profile is not constructed from any specific knowledge of the aerosol to be retrieved." What does this mean? I guess the a-priori profile is some kind of climatology, or ...?

We simply mean to say that this a-priori was built as a very simple climatology, and not with a-priori knowledge to the specifics of the observations - i.e. using reports from other instrumentation to build an idea of what to expect. The a-priori profile we use is just a file with steadying increasing aerosol from high to low altitudes, including an extra contribution around 20 km representing the Junge layer.
The text of the paper has been adjusted to more clearly reflect this.

Figure 10: This caption ends with the statement that "This small error primarily results from the horizontal averaging." For a small error, maybe I shouldn't quibble, but I have to ask: What quantity is being averaged horizontally, and how? Is this averaging discussed elsewhere in the paper, and is it obvious how you determined that this averaging is the cause of the observed error?

We believe we understand the confusion. The radiance profile is built by averaging columns of the images together. The averaging is along the horizontal dimension of the image - not in the vertical/altitude dimension of the image. Figure 5 makes note of this. In the case of the retrievals shown in the paper, we binned all 512 pixels on each row of the detector to get the radiance profile measurement vectors. This binning significantly reduces the noise versus using a single pixel of the row. What we are discussing in the caption is that the error bars of the measurement vector are plotted, they are just very small because of the averaging.

We have changed the caption of Figure 10 in hope of being more clear:
[...] Error bars of the measurement are shown, but too small to be easily visible. This relatively small error of the measurement vectors primarily results from the column binning of the ALI images to produce the radiance profiles.

Figure 12: The properties of the bimodal distribution are shown in the figure, but do you have a reference for how these particular properties were selected?

No reference in particular. To alter a size we just increased the radius with a correspond decrease in width. This is to approximate larger aerosols coagulating with smaller ones. Beyond that we simply made the two profiles with distinct particle sizes in altitude with respect to each other.

Lines 459-460: "This is a similar approach to the standard retrieval approach of OSIRIS and OMPS (Rieger et al., 2019; Taha et al., 2021)." For the latter case, the assumed aerosol size distribution is not log-normal. As stated at the start of the review, I appreciate the experiments with bimodal log-normal distributions to test the approach, but were any non-log-normal experiments also done?

No, non-log-normal distributions were not within the scope of the work we did.

Lines 465: "... yields respectable agreement overall." How is this defined? (With the extinction plots presented on a logarithmic scale, it's difficult to read the percentage error well.)

Added percent difference plots and quantified wording.

Line 503: "... atmospheric DoP can be well retrieved" – this should be quantified. quantified wording.

Line 515: "... but we this is a point" – should be "but this is a point." Corrected

Line 519: "... extinction in very good agreement" – this should be quantified. quantified wording.

Line 521: "... showed an overestimation of aerosol extinction" – this should be quantified. quantified wording.

Line 503: Is listing data as "available upon request" adequate? I understand that public release of the full set of ALI measurements may not be possible. But inclusion of the data used and illustrated in this study (as a public "supplement" file) seems like a reasonable expectation, which I thought had become a fairly standard practice (for data that is not "officially released" and archived elsewhere).

We will look into hosting ALI data on Zenodo (or something similar) in the very near future, and we will update the data availability statement when it is somewhere accessible.