Dear Editor and Reviewer,
We have carefully read your feedback and answered in much detail, whenever necessary for the matter. The questions and remarks were very helpful to refine our text and we have done our best to account for all remarks. With our most recent edits we hope to have clearly underlined the most important aspects of the new method, because some points cannot be simplified further than this, contrary to what has previously been suggested by the reviewer. This regards the 'single root cause' of the improvements, which is due to an interplay of (i) consideration of measurement errors, (ii) coupled retrieval of all state space variables and (iii) the implementation of constraints. No single aspect can be regarded as the root cause for the improvements, because all of them have evident influence on the refined solution, as we explain below and have now also made clearer in the manuscript.
We have also gladly adopted most of your suggestions regarding the calculation and display of error, except for mainly Figure 8, where time and lack of methodology did not allow for error bars in the MLE case. However, a reference to the simulation case has been added for a qualitative assessment in this case.

Line numbers in the following comments refer to the "Tracked changes" version of the document, amt-2021-212-ATC1.pdf.

Major Theme 1: there are various stated reasons for why the current method is better than the SCA; it would be good to emphasize if there is just one that is the key reason (which I think there is), and downplay those that are more minor.

This is a good and helpful remark. As you will see in the following answers, there are indeed three reasons for better performance: i) consideration of uncertainty, ii) coupled retrieval and iii) the constraints. Of these, reason ii) and iii) dominate in our opinion, while i) is still a rather important aspect. Also, if one left out constraints in a coupled retrieval or constrained a decoupled retrieval, not much impact would be expected from our side. This was highlighted in the text, see l. 80-90.

First I want to say, in the paragraph at L250-256, thanks to the authors for adding this note about how the MLE and SCA solutions are algebraically the same when the box constraint is removed from MLE and the zero-flooring is removed from SCA. This insight was very helpful for my understanding of both retrievals, and makes the paper better. But now this better understanding has caused me to question several other details, and I think it's possible to have a more consistent picture throughout the manuscript of how the methods differ and what the impacts are.

The various reasons are first introduced at the paragraph at L76-91. (1) "we account for the noise of the signals" (2) the new retrieval is simultaneous with respect to backscatter and extinction (3) constraints are applied (which are applied differently than the zero-flooring of the SCA) and (4) "dominant anti-correlated noise that originates from the cross-talk correction" is "automatically detect[ed] and suppress[ed]".
We clarified the text by rephrasing (4), but reason in a later comment that (1)-(3) are all important aspects, see l. 80-90.

Reading between the lines of the manuscript, it seems the SCA implements its zero-flooring constraint after the mathematical retrieval of backscatter and extinction; …

For clarification we added a sentence close to line 225, mentioning that the zero-flooring in SCA happens indeed during the iterative retrieval (not just afterwards), see Flamat et al. for reference, and that SCA midbin is not being regularized in this way.

...does this mean it breaks the connection with the measurements, such that the final reported backscatter and extinction solution cannot reproduce the measurements well? If that's correct, I think that's probably the key difference, since the new retrieval, in contrast, implements the box constraints as part of the retrieval and requires the retrieved backscatter and extinction to be consistent with the measurements subject to the constraints. So, it's a better retrieval because it is self consistent and preserves consistency with the measurements. Do the authors agree? …

Indeed, we expect as well that the SCA solution with zero-flooring does not reproduce the measured molecular signal perfectly (see https://doi.org/10.5194/amt-2021-181, Figure 7), in contrast to the MLE case. On the other hand, this should not be the case for the SCA midbin (SCA MB) results, which are not regularized ad-hoc and hence correspond directly to the measured signals.

...If so, this is basically the same as the reason marked (3) above, but I believe it should be possible to clarify the introduction, discussion and conclusions to make it more obvious. I believe it would also be helpful to remove or downplay the other distracting reasons that are less informative, or to relate them to this main reason.

In order to highlight this misfit of the SCA solution wie added close to l. 260: "Due to the implemented zero-flooring of optical depth in the SCA, its retrieved optical properties do not correspond perfectly to the measured signals, though J_obs=0 still holds in the case of the unregularized SCA MB retrieval." The important aspect here is that the MLE does still perform better than SCA MB, which is not regularized.

As for the other three reasons: In backwards order, first the "dominant anti-correlated noise". In my first review, I had trouble interpreting the impact of the anti-correlated noise and the authors response says this is actually not important. In that case, I think this reason should be dropped from L84.

We did so by rephrasing into "A considerable gain in the quality of the retrievals is expected [...], because a coupled retrieval in conjunction with a box-constrained set of space variables will allow for important information exchange during the determination of the self-consistent set of optical properties."

At L 520 "With this [the introduction of constraints] the anti-correlated noise in the cross-talk corrected signals can be traced back and effectively suppressed", I offer an alternate version that I think is more in line with the author response and revisions in the Appendix: "Since the box constraints are integral to the simultaneous retrieval of backscatter and extinction, noise is suppressed in both products simultaneously, in contrast to the zero-flooring of the SCA

which works on the channels independently without regard to the fact that the errors are correlated due to channel cross talk (Appendix A)."

We appreciate your effort and adopted your suggestion with minor changes, namely: “Since the box constraints are integral to the simultaneous retrieval of backscatter and extinction, noise is suppressed in both products simultaneously, in contrast to the zero-flooring of the SCA, which considers the extinction variable independently without regard to the fact that the errors are correlated due to channel cross-talk (Appendix A). Furthermore, the results of SCA and SCA MB do not strictly fall into a physically meaningful subset of solutions, e.g., they include negative backscatter coefficients.”

Next, the simultaneous retrieval of backscatter and extinction. I think this is important too, but not in the way I first interpreted the writing. L 516 says "A coupled retrieval may improve the precision". However, the authors have now convinced me that if it weren't for the constraints (box constraints for MLE and zero-flooring for SCA), both retrievals are the same algebraic noise-fitting solution, since they have the same number of measurements as unknowns. So, in the absense of the constraints, it doesn't matter at all if the backscatter and extinction are retrieved sequentially or coupled. Rather, as the authors emphasize, it's the better implementation of constraints, an implementation that minimizes disagreement with the measurements, that improves the precision. It's much easier to implement these better constraints in a simultaneous retrieval, of course. So, I think it's something like "A retrieval that implements constraints simultaneously with the retrieval of backscatter and extinction improves precision".

We fully agree to the point, that the interplay between coupling and constraints is the root cause for precision improvements, as you figured out correctly. One without the other will not have the desired effect. We therefore clarified this in the text as well, see l. 510 to 530.

Finally, does accounting for the noise in the signals improve the retrieval? As the authors have pointed out in the discussion and revisions, there are the same number of unknowns as measurements, and so the solution (in the absense of the box constraint) is just the algebraic solution. In that case, it doesn't matter what the measurement error covariance matrix looks like; the same solution will be found. What about with the box constraints? No, I think with no prior or regularization term, I believe the minimum cost is still independent of the measurement error covariance matrix, although that minimum won't be zero. In the implementation of the retrieval, if the iterations are cut off at some thereshold that's dependent on the measurement error covariance matrix, then of course that matrix would impact the solution in that way, but if the algorithm actually converges to the minimum cost function (which I believe is the intent here) then I think the measurement error covariance matrix will not impact the solution. Am I correct?

This is not correct, because it is simplified too far. The entries of the covariance matrix have an influence on the solution in the constrained case (but not in the unconstrained case). Consider the following toy model:

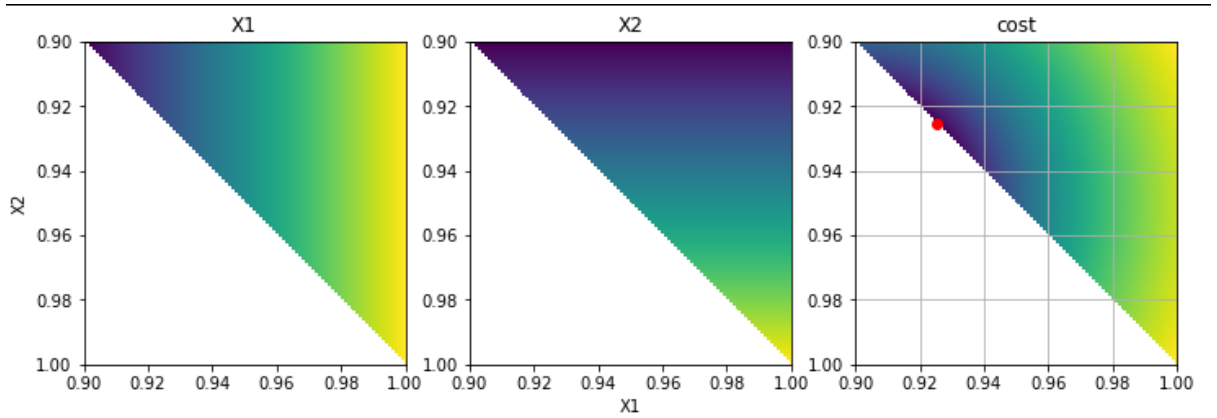
Let’s model something like the ratio of received molecular signal S_meas to the expected (simulated) molecular signal S_sim and let’s call the ratio S_meas/S_sim = X. Now, let’s also assume we measure this ratio (directly, i.e. crosstalk is known) at two different distances r_1

< r_2. Then, we also know that due to additional aerosol attenuation X(r_1) >= X(r_2) must hold, which is our constraint. Shorthand, we just note X_i = X(r_i). Now, assume a measurement noise covariance matrix on X, which is diagonal (uncorrelated measurements) with variances V_1, V_2. We call X_i the state space variable and Y_i its measurement Y_i = X_i + epsilon with noise term epsilon.
*The cost function then is J = 1/V_1 * (X_1-Y_1)**2 + 1/V_2 * (X_2-Y_2)**2.*

Now there are essentially two different cases / scenarios
i) Y_1 > Y_2
Assume e.g. Y_1 = 1 and Y_2 = 0.9. In this case, we can minimize the cost function with X_1 = Y_1 and X_2 = Y_2, which is in line with the constraint X_1 > X_2. Here, the variances do not contribute to the solution and the cost at the solution is J=0.

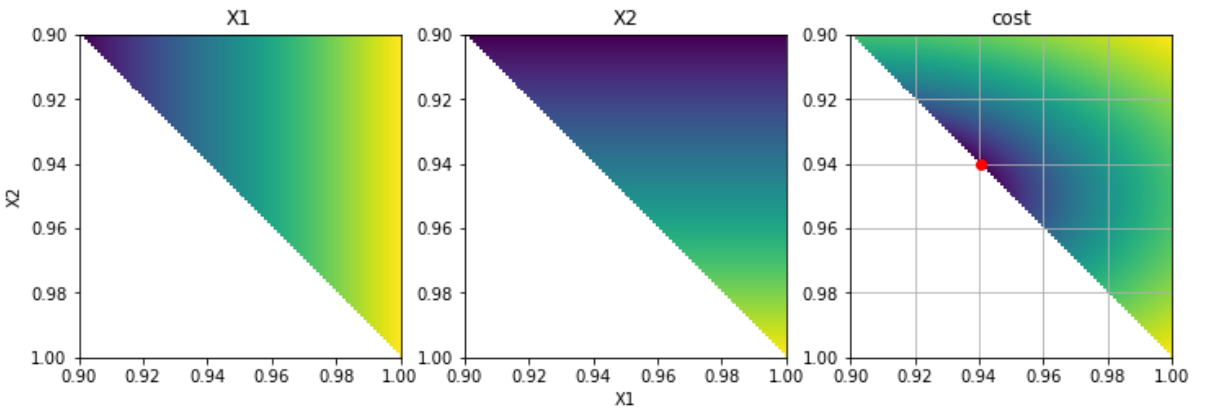
ii). Y_1 < Y_2
Assume e.g. Y_1 = 0.9 and Y_2 = 0.95. In this case, the solution X_1 = Y_1 and X_2 = Y_2 does not fulfill our constraint of monotonically decreasing signal amplitude and can therefore not be chosen.
What classical SCA does (in principle) is to say: Let's set X_1 = Y_1 in the first place, and since X_2 cannot be greater than X_1, we make the best guess of "no attenuation" between the points and assume X_2 = X_1.
*This option corresponds to a cost of J = 1/V_2 * 0.05^2 = 0.0025/V_2*

Now, the MLE considers that both values Y_1 and Y_2 are uncertain. If V_1 = V_2 = V, then the state space and the corresponding cost function would look like in this figure (note that "half" of the state space is cut away due to the constraint):



Here, the red dot marks the minimal cost. Which suggests X_1 = X_2 = 0.925 as global minimum. By setting X_1 = X_2 one can reproduce this solution (minimize J for X_1 after substituting) and see that it corresponds to the weighted mean of the measurements Y_1 and Y_2, which becomes the arithmetic mean when V_1 is equal V_2:
X_1 = X_2 = (Y_1/V_1 + Y_2/V_2) / (1/V_1 + 1/V_2) = 0.925
Not only is the remaining cost at this solution lower than in the SCA guess, namely
*J = 1/V_1*0.025^2 + 1/V_2*0.025^2 = 2/V*0.025^2 = 0.00125/V,*
*but we also see that the solution explicitly depends on the entries of the error covariance matrix (not their absolute but their ratio). If e.g. V_1 = 4*V_2, then Y_2 would have higher*

weight in the solution, which became X_1 = X_2 = 0.94, see also illustrated in the figure below:



This toy model generalises immediately to the problem at hand: The positivity constraint on optical depth is equivalent to a monotonically decreasing molecular signal ratio X with increasing distance from the satellite. The measured signal ratio X is noisy, and so, case i) and case ii) take almost equal share. Especially when the range bin thickness is reduced, the noise (and its estimate) will increase and the MLE solution at the interface of bins with different width will give higher weight to the thicker bin. We hope this clarifies that the relative magnitude of the covariance matrix entries will have significant influence on the solution in the constrained case, whereas the absolute scale of the uncertainty solely determines the planned convergence criterion, which optimally does not impact the solution.
We made a remark in line 292: "Note that the position of a cost function minimum is invariant to scaling of the covariance matrix $\mathbf{S}_y \to \lambda\mathbf{S}_y$ with scalar λ, while the relative magnitude of its entries is important when being subject to constraints, i.e., in a general case where $J>0$ holds at the minimum."

If so, then almost everywhere that the measurement error covariance matrix is mentioned is confusing and somewhat spurious, and should probably be reconsidered. E.g. Line 80 (mentioned above, where better performance is attributed to accounting for signal error); and L 273 - I understand of course the desire to avoid the complication of the off-diagonal terms in the covariance matrix, but if the covariance matrix doesn't impact the solutions, perhaps this isn't very relevant; and L 279 - "As pointed out by Povey et al (2014) unbiased estimates are a prerequisite" - in Povey et al, it's a prerequisite for an optimal estimation retrieval that has a prior term that must balance with the measurement term, but this retreival does not have that feature.

As demonstrated above, the choice of the error covariance matrix will influence the solution in presence of constraints.

Major Theme 2: There's a need for better characterization of the error distribution of the box-constrained MLE retrieval results. The authors have shown that the new retrieval produces a better-looking solution than the SCA and SCA MB retrievals, but they also have shown that it is still quite impacted by measurement noise and the resulting error is quite significant (100% or more). For that reason, the results should not be used without an understanding of their uncertainty. The simulation cases, therefore, are a hugely important part of this manuscript. It is also important to include some kind of estimate of the spread of

solutions for the included real data cases as well. Here are the easiest ways I can think of to do this.

Figure 7. Each solution appears to be an average of multiple bins. Can the figure include an indicator of the actual spread of solutions obtained from each averaging interval?

This is a very neat idea, we therefore display now the range of maximum and minimum lidar ratio values within the bins used for each data point and updated the caption. This gives the reader a clear idea of the retrieval method's behaviour without hypothesising on the signal noise.

Figure 8. I appreciate that this case is harder, since it is only two individual profiles. One approach is to produce another simulation with profiles taken from the solution of this real data case (i.e. same backscatter and extinction as one of the profiles or the mean of them). If it is difficult to use the instrument simulator to do this, then I think it would be accpetable to use the assumed measurement error covariance matrix (i.e. from Eq 13) to estimate it. (This is the clarification of my earlier suggestion that the authors requested.)

We agreed to delete the appendix on analytical error propagation, as it is not used throughout the manuscript anymore. Therefore, we would want to mitigate confusion by relying on the non-representative, analytic way of calculating uncertainties here, without having it documented anymore. Due to the time constraints, we were not able to find an alternative for the error propagation, so we have to keep Figure 8 without error bars on the MLE results. But we suggest to add the following text within section 4.4 for a qualitative error assessment:
"By comparison to the similar simulation case I, we expect a backscatter coefficient error in MLE retrieval on the order of 30 \% within the aerosol layer below 3.5 km and an extinction coefficient error on the order of 100 \% for individual range bins. A comparison of SCA and SCA MB results with the simulation case I and the ground truth also suggests, that the currently reported error in the L2A product is no reliable estimate."

Ultimately, the authors would like an analytical error calculation since only an analytical solution is believed to be fast enough for practical use in real data processing. However, since both the input errors and the forward model (when box constraints are included) are non-linear, any analytic solution with assumptions will be hard to accept until sufficient research and analysis demonstrate consistency with existing numerical solutions for real data. So, a fair number of numerical solutions will be required to be calculated anyway. This underscores my hope that it's not unreasonable to wish for numerical results for the specific cases in the manuscript.

We agree that any analysis of an analytical approach will need to rely on a manifold of simulation cases. But for this comparison to be made, we would need to develop an analytic model / an alternative approach in the first place. You also acknowledge that such a comparison will take considerable effort. Hence, we will have to work on the error characterisation in the future work but cannot include it within the rather limited time frame of this revision.

I would also suggest completely removing the analytic error propagation. E.g. the paragraph at lines 324-334 and Appendix C. The authors know it does not correctly represent the error propagation of their retrieval because it doesn't include the box constraint. For that reason, they have deleted all the results relating to this error propgation. So, it should not still be included in the methodology. I think everything after "whereas" at line 325 should be deleted (as well as Appendix C). If desired, the authors can make a short statement acknowledging that an analytical error propagation for the current retrieval would have to include a way of representing the impact of the box constraint, which is a topic of future work.

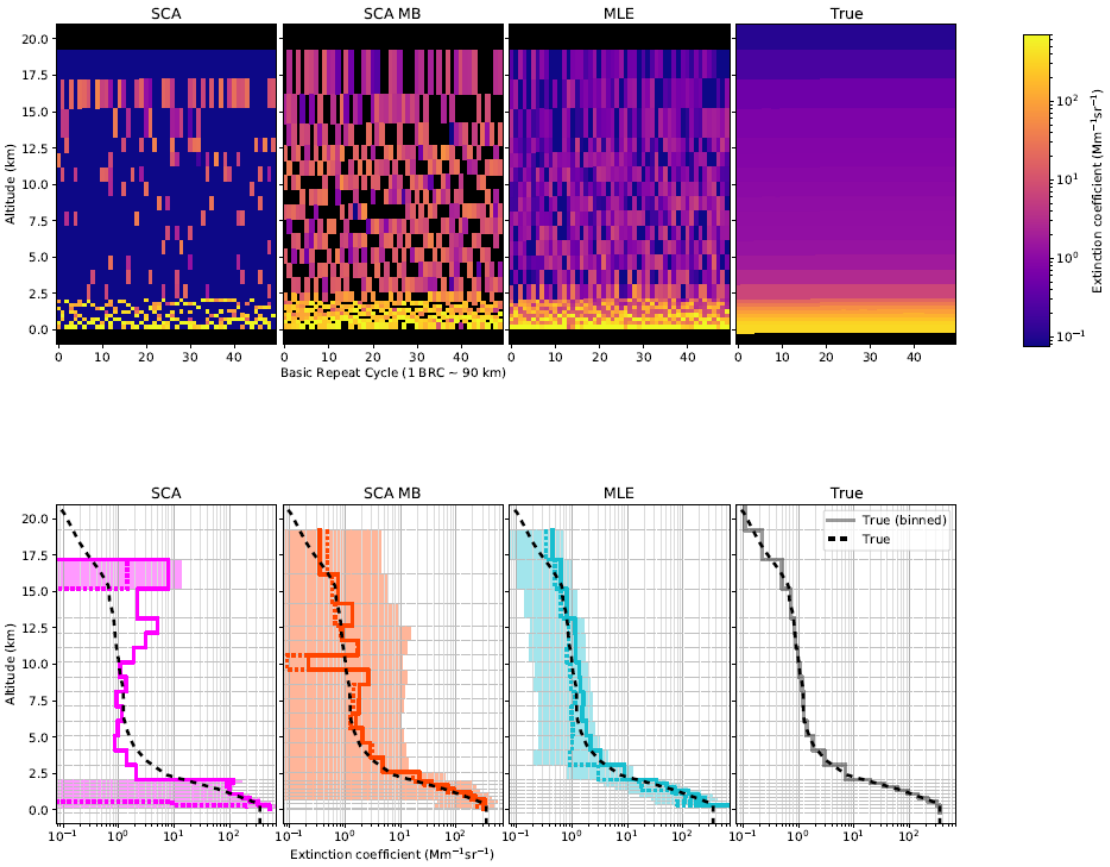
Given the earlier discussions, this is a reasonable measure. We shortened the paragraph at lines 324-334 significantly and added a note regarding error propagation from linearisation: "While a similar analysis can be made for the MLE, we find that especially the obtained extinction uncertainties would strongly overestimate the actual variability of MLE results due to the omitted constraints in such procedure. Hence, future work regarding the implementation of the box-constraints in the error estimation is pending."

We could potentially keep Appendix C for further reading, but it is indeed obsolete with respect to the results. So we decided to remove it as suggested, together with all references thereof.

Minor Theme 1, related to Major Theme 2. The representation of the errors for the simulation cases can be improved.

Figure 2 and Figure 4 make clear that the standard deviation is not a good representation of the spread of the errors in the box-constrained MLE method (so perhaps also for the SCA and SCA MB as well), in that the shaded area for the lidar ratio goes well below 2 sr, although the constraint makes it impossible that any solutions lie in that range. I encourage the authors to remake the figure with a different visualization of the spread that is more representative of the actual range of the results. An alternative is to use a percentile spread, such as the 25-75% limits. (In figure 3, since standard deviation is shown along with a bias statistic, I think it's more acceptable).

The primary goal of the visualization in Figures 2 and Figures 4 is to compare the methods. For your request I replotted the data in these figures using a 25 and 75 percentile. Doing so, another problem occurs in the SCA case (see figure below): Since the majority of extinction values below the first bin are zero, there is often no visible spread as both the 25 and 75 percentile are also zero in mid latitudes. Unless we take higher percentiles, this problem cannot be solved, but too high percentiles are not representative of the bulk data anymore. Although the distributions for SCA and MLE are non-symmetric due to the constraints, we suggest keeping the standard deviation as an established measure of spread (and error) for the sake of comparability of the retrieval methods and with Figures 3 and 5. We appreciate that penetrating towards the negative x-axis is unnecessarily confusing for the MLE case and therefore clipped the error bars at zero for the MLE lidar ratios. We hope to have explained this point well and that this is a satisfactory compromise, as we believe that we cannot find a statistical measure of spread that perfectly suits all different empirical distributions for SCA, SCA MB and MLE variables.



Previously I asked about the potential for bias caused by the box constraint algorithm and the characterization of the error output to account for potential bias. The authors have added a line late in the paper that suggests that an observed bias may be due to the box constraint, and I appreciate that, but they also admitted they do not have a good understanding of the algorithm. I think more analysis is required to understand how the algorithm affects the distribution of solutions. I suggest one way to gain a better understanding is to show a histogram of the solutions from the synthetic cases, particularly a histogram of the lidar ratio with mean and median marked. For instance, it would be good to see the behavior near the edges (e.g. 2 sr), whether the histogram looks truncated or rather "piled up".

The reason why we do not provide the LR statistics is due to the fact that the LR histogram is not representative of the results. That is because, when the algorithm retrieves optical depth close to zero, the lidar ratio can be arbitrary (and indeed bunches at the first guess and the borders, etc.). In order not to confuse the reader, we report the lidar ratio defined from the mean backscatter and mean extinction, effectively throwing out the influence of "(almost) empty" bins with diverging error. Therefore, it is true that the lidar ratio is directly retrieved. But in contrast, backscatter and extinction are always well defined properties, which is why they are preferred. So, showing a lidar ratio histogram would mislead the reader, as all samples would be assumed to have equal weight.

Figure 3. I like the new Figure 3 very much; it is very helpful to see the behavior specifically in the region with significant aerosol, and helpful to see a bias calculation paired with the standard deviation, and to have equations at line 364-365 specifying the statistics. I would

also like to see panels showing the statistics for lidar ratio (which is, after all, a directly retrieved quantity from the MLE retrieval.)

First of all, thanks for the positive feedback on the improved visuals. The reason we did not include a seperate panel on lidar ratio is to be economical in terms of space. Mean and error of lidar ratio appear clearer in Figures 2 and 4 than for extinction and backscatter due to their linear scale and the piecewise constant ground truth.

Figure 5. Figure 5 should have the addition of the zoomed in boxes, like Figure 3 has.

Thanks for the remark, but considering the wide spread of the results on the x-axis, the visualisation does not considerably improve by zooming in solely on the y-coordinate. For Figure 3, this was different, as there was considerable "bunching" in both directions close to the ground. We therefore suggest saving up this space.

Minor theme 2. Items related to the discussion of ratioing of signals and discontinuities where the range bin size change.

L 367-370 (a). First, I suggest replacing "seems to be triggered by the refined range bin" with a more definite description of the observation, holding the hypotheses for the next sentence. That is, "the bias is colocated with the change in range bin size". I suggest this because I think the change in range bin size only makes the problem more obvious, but does not actually cause the problem (more below).

We included this small change.

L 367-370 (b). Next, about the first theory about the bins that are not uniformly filled: does this make sense? Wouldn't the requirement for uniformly filled bins be more badly violated by large bins and better met by small bins? If the suggestion is that it is worse here due to the dramatically increasing slope of the aerosol extinction profile, then the bias would logically be colocated with the slope but only coincidentally colocated with the change in bin size. Flamant et al. is quoted elsewhere as predicting a bias in extinction due to this reason, but is it also expected to affect backscatter, as here? If this part is kept, Flamant et al. should be referenced here, with a specific description of their work showing how non-uniformly filled bins cause a bias.

We made this point referring to the gradient in the ground truth profile, which is steepest (coincidentally) in the region with the lowest bin size. On second thought, I would argue that with this gradient direction, one would expect underestimated backscatter, since the molecular attenuation happens mostly in the lower part of the bin (increasing X), while the aerosol backscatter Y remains roughly similar. Hence, the backscatter coefficient beta ~ Y/X would need to decrease in this area by applying the hypothesis of uniform bins.
Therefor, we deleted this sentence.

L 367-370 (c). Finally, I find the other theory more convincing (ratioing of noisy signals). But why doesn't it apply to the MLE retrieval as well?

A glimpse to the reason was provided by our statement "..., because here mean(beta) will become biased high increasingly with increasing uncertainty of Xi." Now, we completed the argument as follows: With MLE, we constrain the possible values for X_i to a physical subset, which makes them effectively less uncertain compared to SCA and SCA MB. Hence, a lower bias is obtained. This was added to the text.

L412-413, "due to its noise suppression capabilities" is presumably the answer to my question but I find it somewhat vague. Can it be made more specific?

With the addition to the prior subsection (see answer above), we suggest to keep this statement as is and hope to have clarified this point.

L579. "if this varying reliability of the signals is not taken into account". I believe from L367-370 that the bias is linked to taking the inverse of a noisy signal, not because of a discontinuity in bin-size, although it is more noticeable because the discontinuity in bin-size results in a discontinuity in the bias.

Thanks for stumbling over this statement. It is our fault we did not make clear that this paragraph is dedicated to the extinction variable. We completed this argument as follows: "Hence, if this varying reliability of the signals is not taken into account, biases or oscillations in the extinction variable can potentially be triggered by zero-flooring whenever range bin heights change. This is due to the fact that extinction essentially depends on the moving ratio of noisy signal values along an atmospheric column. The mean absolute of this ratio increases with increasing noise, which may then lead to a bias after zero-flooring negative values in SCA, see also (reference to https://doi.org/10.5194/amt-2021-181) for a graphical explanation and a showcase without flooring."

Miscellaneous minor suggestions in line-number order.

L 37. Illingworth et al. 2015 is a secondary reference with respect to the idea that the indirect effect of aerosol on clouds is the largest uncertainty in radiative forcing. I suggest referecing some primary sources, or a major climate change review such as the IPCC report.

We now reference the IPCC report directly.

L303. "not invariant under variable transforms". While admittedly I only quickly skimmed Zhu et al (1997), it caught my eye that they say the algorithm is indeed invariant under transforms with the exception of the first step away from the first guess. This suggests to me that efforts to find a better first guess might be better rewarded than attempts to find a better transform to deal with a poor first guess. The aerosol-free atmosphere is a difficult first guess to work with. The standard HSRL technique provides backscatter in one algebraic step. Perhaps consider calculating backscatter from the ratio of the channels, and with this estimate optical depth using your first guess lidar ratio of 60. (I am not suggesting this is required as a response to this review but offer the suggestion in case it's helpful.)

Just above the lines you are referring to, Zhu et al. state in section 4.1: "However, complete scale-invariance was not possible to achieve; indeed the limited-memory algorithm itself is

*not invariant to linear transformations in the variables." This is also what we observe. The statement you refer to is "However, the algorithm is invariant with respect to scalar multiples of the variables and the objective function, and we have been able to maintain that invariance in the code with only a few exceptions." We interpret these statements such that J -> lambda*J and v -> lambda*v with scalar lambda, cost J and variable vector v are symmetries of the algorithm. However, the algorithm is not invariant to the general case v -> M*v with a matrix M, as the first statement about linear transforms implies. Not being invariant to linear variable transformations implies that there is no general invariance to nonlinear variable changes (e.g. taking the logarithm) either.*

L312. I'm curious about the description of running 40000 iterations so that "the estimate should fit as close as possible to the signal data". Does the cost function really continue to decrease for 40000 iterations? I believe it's common for the cost function to begin to jump around after a while and not continuously decrease. I agree with the point that cutting off the iteration prematurely leaves some unnecessary impact from the first guess (especially if the measurement error covariance matrix is not strictly correct), but I do think it should be cut off when the cost function ceases to decrease.

In our current implementation it continues to decrease, but with very low speed. That is why we checked additionally with the total cost criterion, in order to assess at which point no additional information can be extracted from the measurements (whenever the modelled signal values are on average maximum a standard deviation away from the real signal).

L 378. "the feedback" is vague. Does it mean that when the SNR decreases there is a greater proportion of negative solutions that get filtered out by the zero-flooring, and therefore bias the mean solution?

This is a reference to the topic of line 579 (see comment above). We explain this feedback in the Appendix A as written above and added a reference to line 387 for the interested reader.
L380. At the start of the added section, it would be good to say "with the exception of the bin closest to the surface". This is mentioned in the following paragraph, but the first paragraph is confusing with this omission.

We fully agree and inserted the remark.

L 469. Copolarized lidar ratios of 80sr to 120 sr for depolarizing desert dust are attributed to Wandinger et al. 2015. Does that paper really present copolarized lidar ratios? Or is this a calculation of the authors' based on non-polarized lidar ratios from that Wandinger et al. 2015?

Wandinger et al. explicitly present the discrepancy in Aeolus observations for different types of aerosols, the provided estimates are extracted information from the plots.

L 415-421. What does the averaging kernal look like below the cloud in the region that has up to 100% bias and up to 500% relative error, but where the average lidar ratio "remains quite accurate"? It would boost confidence in the conclusion, if the averaging kernal also shows that the optical depth is not reliable but the lidar ratio is.

We are not sure if a look at the averaging kernel can help here, because the high-bias is mainly a non-linear effect (ratioing of noisy signals). We believe this might be a misunderstanding of the terms accuracy and precision? The precision of individual lidar ratio estimates is of course very low, indeed. The presented lidar ratio is obtained from the average backscatter and average extinction over 1000 realisations which have high individual errors. Not the individual lidar ratio is accurate, but this average. To motivate this we added: "This suggests that the noise induced biases in extinction and backscatter do almost balance, which can be motivated by the fact that both variables depend on X_i^{-1} (the aerosol optical depth equals the normalized log-derivative of pure molecular signal and can be rewritten in terms of the ratio $X_{i+1}X_i^{-1}$)."

L 517. Geometric overlap? Is that relevant for a satellite lidar like AEOLUS? Is this sentence meant as a more general discussion that also encompasses ground-based lidar?

See below.

L 517. I can't see the cross-talk calibration as part of any tradeoff between a coupled or sequential retrieval, since errors in the cross-talk calibration will significantly impact either style of retrieval.

This sentence was included to highlight why processing schemes for ground-based lidars such as developed by Marais et al. do prefer not to implement a simultaneous retrieval (as stated in lines 87-89). For the specific case of Aeolus it is superfluous, which is why we removed the statement in the manuscript's conclusion.

L 522 (approximately). It should be repeated in the conclusions that in the simulations moderate amounts of aerosol still cannot be distinguished from zero (despite the fact that the new retrieval does significantly better than the existing one). This can be part of the motivation for the future work with signal accumulation. (And by the way, the manuscript has quite a good explanation of the motivation for the scene-based retrieval strategy.)

Right, the moderate aerosol amounts cannot be distinguished from zero on a single bin basis, unless one averages over bins, because then the uncertainty decays with approximately 1/sqrt(N).

We included: "It is important to note that despite the improvements, moderate backscatter coefficients of about 0.1 Mm^{-1}sr^{-1} can still not be distinguished from zero on a single bin basis. Higher precision can only be achieved by signal accumulation or averaging of the backscatter coefficient estimates."

L 594. "the contribution from the Mie channel in the particle-free atmosphere is pure noise". But C1=C4=1, so the molecular backscatter is distributed evenly across the two channels, so I don't think this is true. Perhpas "the signal in the Mie channel in the particle-free atmosphere is more than half noise"

Entirely correct, what we meant to refer to was the signal Y. We adjusted the statement according to your suggestion.

Wording suggestions:

L35. I suggest avoiding the awkward parantheses. Specifically, I suggest removing the parentheses around "optical" and simply deleting "(change)"

Adopted.

L40. Consider deleting "so-called". "So-called" usually has a connotation that the speaker does not agree with the label or as a way of using an informal-sounding name in a more formal context, neither of which apply here. I know the intent is "what is called" but it's redundant here anyway, so could just be deleted.

We deleted so-called from the text, whenever it seemed redundant.

L276-277. The sentence that starts "This accounts" is confusing, and I'm not sure I really understand it. Can the authors please reword this?

We replaced it with “This overlap is about...”

L297. Consider changing to "can cause the retrieval results to underestimate the true particle extinction by a factor of 16, and therefore underestimate the lidar ratio".

Adopted.

L314. I didn't follow what "even in unfavorable conditions" refers to.

This sentence can as well be removed, so we did so.

L386-387. I suggest deleting "likely due to the diminishing influence of the lowermost optical depth on the cost function". The non-sensitivity of the cost function at this point does not determine that it will over or underestimate, but it does show that a big error was expected, and it's not a speculation. The second part of the sentence can stand on its own without the "likely" part.

This is right, we rephrased it into saying, that deviations from the ground truth were expected.

L394. Not everywhere but nearly everywhere.

Adopted.

Figure 6 labels and caption. "Lidar ratio" should be "copolarized lidar ratio" (everywhere, but particularly important where measured data is described).

All figures presenting measurment data have been updated with the term “co-polarized”.

Figure 7 caption. The sentence beginning "The upper error bound" is no longer relevant.

This was deleted.

L505. The sentence "It should be stressed...lower error margins" is no longer relevant.

Deleted as well.

L579. Change "if" to "since"

Adopted.