

We thank the reviewers for the careful reading and the suggestions that helped us improve our manuscript. The answers to the reviewers will be given after each comment in bold font.

Reply to Referee #1

The paper is a nice addition to the literature of lidar measurements of lidar ratio and particle linear depolarization ratio of different aerosol types, with good support from comparisons with other methodologies. Interesting findings include support for a clear separation between the lidar ratios of Arabian and Saharan dust, and at least one case where the in situ suite suggested a pollution component in a dust mixture where the lidar could only detect dust. A section with dust mass and volume concentrations derived from a chain of assumptions is less interesting partly because a lack of error characterization makes it difficult to know how quantitatively useful these estimates are, yet even the qualitative comparison revealed a possible weakness with models' dust transport making another useful conclusion.

Thanks for this valuable comment on the dust mass and volume concentration. We tried to better account for the errors and uncertainties that come along with the calculation of the dust mass and volume concentration to hopefully give more reliable data.

While the manuscript is generally good, there are sections that seem frustratingly like a first draft. These include statements that are unsupported or vaguely approximate where precision is called for, important data presented in tables or figures without much discussion or analysis in the text, and figures with inconsistencies in presentation that make them difficult to relate to each other or with the text. Also, in many cases it is difficult to find a description of which instruments were used to derive specific data items in the figures and tables. (In most cases, I think I can infer which instruments, but can't find text to confirm my impression.) There are probably also a least a few actual errors in numbers in the text and figures, so these should all be checked thoroughly. I think it's important to improve these points and not ignore them, but I think all my suggestions should really be quite straightforward and easy for the authors to improve.

Thank you for pointing this out. We carefully revised the manuscript.

Some important specific points:

How were the lidar values that are quoted in the abstract and conclusion derived? From single measurements, or from averages over the profiles of the selected case studies, or aggregates over every point in the mission identified as a specific type? Does it combine POLIS and POLLY-XT measurements or refer only to POLIS? This should be stated clearly in whichever section is most relevant to the calculation and again in the conclusion.

The mean values give the mean for the different types including all measurements that have been performed for the specific type and listed in Table 1. We modified the text accordingly.

Error bars and quoted uncertainties throughout refer only to systematic errors. It's excellent that the team has characterized the systematic uncertainties so carefully, since that is sometimes neglected. However, this methodology also mentions significant

averaging, so there is random error that's not represented. There is also significant natural variability and it's important to have an estimate of the variability, to better support conclusions about whether different studies observe comparable values. It seems like multiple measurements are used to derive the quoted values of lidar ratio and PLDR (i.e. in the abstract and conclusion), so a standard deviation would be easy to calculate. (Measurement random error would probably have to come from a propagation of errors that the authors may not be prepared to do, but if it is available, this should be reported as well).

Thank you for pointing out the necessity of error calculations. For comparability to other studies, we added the standard deviation to the mean values shown in the abstract and in the conclusion. However, the two groups providing the lidar data did not do the same completeness of error propagation. Thus, to keep the data comparable we choose to only give the systematic errors and standard deviation.

At line 125, it's stated that the manuscript will "intercompare the measurements and resulting classification" (between POLLY_XT and POLIS). This intercomparison was not done. While both sets of measurements appear in Table 1, there's no discussion of them. Table 2 does not show classifications for the two lidar separately and although the separate classifications appear in the scatter plot of Figure 9, there's no way for a reader to identify simultaneous measurements and again no analysis or discussion. Indeed it is also difficult to tell from the text whether both or just one of the instruments' results were used in various places, such as the summary values quoted in the abstract, and for the lidar classification in Table 2.

The reviewer is right, we did not intercompare the typing using POLIS and Polly^{XT} measurements. We added a paragraph after the general typing schemes:

'As both lidar POLIS and PollyXT were located site by site for during the A-LIFE campaign, we can use the measurements to check if and how the analysis of different lidar systems with different algorithms done by different research groups affect the outcome. As mentioned in Section 3.4, we found now significant differences in the retrieved extensive optical properties (i.e. the extinction coefficient). With the PLDR and lidar ratio presented in Table 2 and Figure 9 we can also check if there are significant differences in the retrieved intensive optical properties. Although the mean values for the lidar ratio (both wavelengths) partly differ by 10 sr or more, considering the uncertainty range of the retrieved values we found no significant differences. For the PLDR at 532 nm we found an agreement of the mean values within 0.02 between the two instruments. For the PLDR at 355 nm the differences of the mean values are as large as 0.06 for the dust dominated day around 20 April 2017. However, considering the uncertainty range, the differences are not significant. Differences can occur from different averaging (time and height) as well as from differences in the lidar performance (e.g. signal strength).

Applying the classification scheme on the intensive optical properties we find good agreement of the results between the different systems and the different wavelengths. Only for the pollution case on 11 April 2017, the classification at 355 nm and 523 nm show light differences in the assigned aerosol type. The difference in the PLDR (0.12 ± 0.02 at 355 nm and 0.04 ± 0.02 at 532 nm) causes this difference. While the aerosol type was classified as polluted dust at 355 nm, it was classified as pollution/smoke at 532 nm. POLIS measurements are missing for that day to check the validity of the PollyXT classification.'

I would like to know more about the case where in situ typing indicated a pollution component in an air mass that the lidar could not possibly type as anything other than pure dust. This was called out in the conclusions and I agree it's an interesting finding, so it deserves more attention. Specifically what chemical or other measurements in the in situ suite indicated pollution? Is there an estimate of the mass fraction or any other measurement that would support the statement that it's a minor component (as stated in the text) or alternately "moderate" (as stated in the table) (And by the way, that also needs to be made more consistent). Would this amount of pollution be expected to impact either the radiative properties or the aerosol-cloud interactions? It might be useful to highlight this case, and/or other "difficult" cases as case studies, besides the pure type cases highlighted earlier.

The investigation on how pollution within mineral dust layers affect the optical and radiative properties of these layers goes beyond the scope of this manuscript. But as it is an important topic it is covered in an extra publication by Teri et al. (2024). In addition, we added the optical properties of these days (5, 6, 27, 29 April for Arabian dust, and 21, 22 April for Saharan dust) measured in-cabin of the Falcon aircraft in the supplement.

Specific points by location in the manuscript:

Abstract: Please state in the abstract that the quoted error bars represent systematic error and not variability.

Done. In addition, we added the standard deviation.

(Introduction)

Line 42 or wherever seems best to the authors, please discuss whether the analysis in this study assumes the mixing is external mixing rather than internal, and any implications.

We added this in the Methodology chapter (2.4 Aerosol typing and aerosol type separation based on lidar measurements:

'Based on findings from former studies on Saharan dust (e.g. Petzold et al., 2011), we assume a two-type external mixture of mineral dust and pollution. This assumption is in good agreement with the coordinated in-situ measurements (see Section 4).'

Line 47. The argument for the need for typing is vague and somewhat unconvincing. It could probably be improved by being more specific. E.g. specifically how would radiation calculations use typing information and what quantitative properties could be estimated better by knowing the type of particles?

We added a sentence why aerosol typing can help to determine the effects of the different aerosol layers. To not go to much into detail here, we added the references to prior publications on this topic:

'Remote sensing data from airborne or spaceborne measurements provide information on continental and global scale. But they cannot directly derive the particles' microphysical properties or chemical composition, and thus their radiative effect and capability to act as

cloud or ice nuclei. However, those properties strongly depend on the type of particle (e.g. Groß et al., 2013; Wandinger et al., 2023).'

(Methodology)

Line 112. "and for the daytime measurements". Please expand on this to explain how the lidar ratio is used for the daytime measurements. The information is in the Discussion section on page 19, but the Methodology is where I think most readers would look for it first.

Thanks for pointing this out. We moved the explanation we gave in the Discussion section to the Methodology section.

Section 2.4 and 2.5. There are eight or so different assumed values in these two sections to make the leap from the lidar-derived particulate depolarization ratio and backscatter to dust fraction and volume and mass concentrations. Clearly this must lead to a lot of uncertainty in the output. What is the estimated uncertainty on the assumed parameters and on the calculated outputs?

The uncertainty of the estimated dust fraction is about 10-20 %, of the estimated volume and mass concentration it is about 10-15 %. The uncertainties are given in Figure 10, and we added the uncertainties in the text.

Line 156. The procedure for classification with AERONET measurements should be discussed in the methodology section (currently part of 3.5.1)

We moved the description of the procedure for AERONET base classification to the methodology section.

Line 165. How are the two different sources of dust distinguished using in situ data?

The two dust types are mainly distinguished based on the information of the dust source region.

Line 169 and Section 2.8. It's confusing that the "in situ" classification is stated to use FLEXPART, and then there's also a separate FLEXPART classification. If FLEXPART is really used as an integral part of what's called in situ classification, then this needs more explanation so that readers can understand what is really based on measurements and what's based on a model. If it's only to separate Saharan and Arabian dust then it would be better to have the in situ classification in table 2 only indicate "dust" and leave the FLEXPART classification as a completely separate entry.

The airborne A-LIFE data set was classified into 4 main classes (Saharan dust, Arabian dust, mixtures with and without coarse mode) based on a threshold value of in-situ measured coarse mode number concentration and simulations with the Lagrangian particle dispersion model FLEXPART. FLEXPART results were used to decide whether a measurement was classified as a dust layer, and to distinguish between Saharan and Arabian dust. Each of the four main aerosol types was further separated into three sub-classes (clean, moderately-polluted and polluted) based on threshold values of the ratio between refractory black carbon mass and coarse mode number concentration. Details about the A-LIFE aerosol classification scheme can be found in Weinzierl et al. (2024, in prep.), and Teri et al. (2024).

L183 "an aerosol type". The FLEXPART aerosol labels look like they are probably components that can appear simultaneously, rather than a single aerosol type per one-minute section. Please clarify. That is, in Table 2, where groups are listed, such as "SD, OM, SO₄, SS", is the reader to understand that these are multiple components found together? Consider changing the wording to say something like "were assigned aerosol components" instead of "an aerosol type".

We followed the reviewer's advice and changed that accordingly.

(Results)

Line 193. Please specify which satellite measurements.

We used a number of information amongst them, measurements of MSG. We added this in the text.

Figure 1. It is difficult to pick out data for specific dates mentioned in the text, exacerbated by having 6 minor ticks for each 5 days, a difficult interpolation to do by eye. Can the axis please be reformed to be similar to Figure 2 where each minor tick represents one day?

You're completely right. It is hard to spot specific days. Thank you for your comment. We updated Fig. 1 to have exactly 5 ticks for 5 days.

Figure 1 and Figure 2. It would also be very helpful to annotate the x-axis on both figures with lines to indicate the major events and case studies that are discussed in the text.

We modified the Figures accordingly.

Line 210 "Similar to the Arabian dust events ... almost wavelength independent". I don't see that the Arabian dust cases can be described this way, especially the first. Please reword.

We reworded the sentence as follows:

'Both events are characterized by low Angstrom exponent indicating no or low wavelength dependence. But while the Angstrom exponent (440-870 nm and 380-500 nm) during the Saharan dust event shows typical values of ~0.2 (Toledano et al., 2009; Groß et al., 2011) the Angstrom exponent (440-870 nm and 380-500 nm) during the Arabian dust event shows slightly larger values of ~0.6 for 440-870 nm and ~0.8 for 380-500 nm. Both events show a large contribution of the coarse mode particles to the overall AOD at 500 nm.'

How does the AOD summed from the lidar extinction compare to the AERONET AOD? This could help verify that the AERONET is not impacted by clouds, highlight any potential issues from day night differences, and potentially help support the subsequent discussion about the mass concentration differences between the lidar calculations and AERONET.

We investigated the agreement of lidar and sun photometer derived AOD for some of the coordinated measurements and found good agreement. However, this cannot verify that the AERONET algorithm filters out all contribution of clouds, as the AERONET and lidar measurements are different in resolution and sampled volume. We do not include the AERONET measurements in the discussion about mass concentration.

Line 232 What's the justification for calling these cases "pure"? The in situ classification in the table does not seem to support calling them all pure.

The reviewer is right, the in-situ measurements indicated enhanced coarse mode particles in during these events but still indicated that the major aerosol type was pollution aerosol. We rephrased the sentence to *'with a dominance of'*

Figure 3. While the backtrajectories show the direction the airmasses came from, they do not seem to go far enough back to show the sources. Both the pollution and Saharan dust tracks are just as dispersed in altitude at the earliest time shown on the track as at the measurement location. Wouldn't we expect to see tracks converging in altitude to the injection height at the source? Later, at line 280, it says "(Figure 3) indicated the western Saharan region" but that's not actually true, because the trajectories don't go far enough back to reach the western part of the Sahara.

This is right, the back trajectories shown in Figure 3 shall highlight that in fact the airmasses are coming from different direction. For a detailed source analysis, we used the FLEXPART analysis.

Line 244. Again, which satellites?

Amongst others MSG and MODIS

Figure 2 shows two different Angstrom exponents. The text frequently refers to the Angstrom exponent, but doesn't say which one.

We added the wavelength pairs we used for the Angstrom exponents.

Line 248. How were the PLDR values 0.24 ± 0.02 and 0.27 ± 0.01 arrived at? They really do not seem to be averages of the values displayed in Figure 4 between 0.5 and 2.0 km altitude, as implied. Some subset near the top of that range may average to 0.27 at 532, but there doesn't seem to be any single value above 0.23 at 355, so a subsetted range can't really explain the discrepancy there. Are these correct?

The reviewer is right, for 355 nm this was a typo. The right value is 0.21 ± 0.02 . We corrected this typo.

Line 273. Similarly I can't understand how the values at 355 and 532 differ by 0.01 when the profiles in Figure 5 seem to completely overlap. Is it because the 532 nm profile goes up a bin or two higher in altitude than 355 where the values are largest? It also seems strange that it wouldn't be acknowledged in the text how similar these two lines are. Is there any error here?

We checked our analysis and the averaging range for both wavelength and found no change in these values.

Figure 6 has a clear error in how the PLDR error bars are plotted. Both green and blue error bars are plotted around the green line with none around the blue line. It's also remarkable how much less vertical variability the green line has. Is there any difference in the smoothing/resolution between these?

Thanks for pointing this out. We corrected the figure. No, there are no differences in smoothing or resolution.

The quantities quoted in the case studies seem to be somewhat sloppy estimates. For instance, the 380-500 Angstrom exponents on April 5 on Figure 2 cover the range from about 0.8 to 1.3 and are described as "characterized by low Angstrom exponent of about 0.5" (line 246). Even if it's the long wavelength Angstrom exponent that's meant, 0.5 seems to be just a bit lower than any of the values shown on that day. The 355 nm lidar ratio is described as "wavelength independent" (line 250) but it is not specified that this applies only to POLIS and not to POLLY_XT.

We changed the description of the case studies and also added that for the detailed case studies only the POLIS measurements were used.

Line 267 describes the AERONET AOD as "about 0.15 at 500" and Angstrom of "about 1.5" but the AOD never gets to 0.15 on that day and seems to vary around about 0.12, and 1.5 is at the high end of the variability of the Angstrom exponent. Instead of reading "about" this or that, I would rather see a quantitative average and standard deviation over each day (or whatever period is desired) for a more accurate discussion. Same for the discussion of the lidar values which are also described as ">" or "around" (Line 271) instead of quantitatively.

We modified this paragraph to give a more quantitative description.

'For the AERONET based aerosol typing we use the scatter plot of Ångström exponent (440-870 nm) vs. AOD at 500 nm as proposed by Toledano et al. (2009, 2011). Values of the Ångström exponent (AE) larger than ~1.2 serve as indication for smoke/pollution, independent of the AOD. Ångström exponents of <0.5 serve as indication for dust (Toledano et al., 2009, 2011, 2019) or marine aerosols. Following Toledano et al., 2011, a threshold of AOD=0.15 is used to separate dust and marine aerosols. Measurement points with AOD<0.15 and AE<0.5 are classified as marine, while measurement points with AOD>0.15 and AE<0.5 are classified as dust. Ångström exponents between 0.5 and 1.2 serve as indication for mixtures. We further subdivide this value range in dust mixtures for AE larger 0.5 a value of the AERONET derived Fine Mode Fraction larger 0.5 (AE values of ~0.8). Values with Fine Mode Fraction < 0.5 (AE values >~0.8) and AE values smaller 1.2 are classified as polluted mixture.'

Line 286 "Those values have been reported recently". That is an odd way to phrase this unless the same exact values were observed in another study. I assume the authors mean that the current values are similar or consistent with previous studies. Please specify the values from the other studies so readers can see this for themselves. This applies in other parts of the manuscript also. Please always explicitly quote the values that are being compared.

We followed the reviewer's advice and changed this phrase to 'Similar values for Saharan dust after a transport of several days were also reported from recent studies'.

Line 296. I don't fully understand how extinction can be retrieved in the daytime but lidar ratio can't. My understanding is that the extinction retrieval itself is quite difficult if the signal is noisy, but I believe the backscatter retrieval is not as sensitive (is that correct?) so it would seem that if conditions are sufficient to retrieve extinction, lidar ratio would not be much harder. I understand this is probably explained fully in prior papers by the research group, but a summary of the explanation for this in the methodology section would be appreciated.

For the day-time measurements the extinction coefficient is retrieved with the Fernald-Klett algorithm, as described in the Methodology section. As described, this algorithm uses the lidar ratio as an input.

Line 298. It's not really the measurements' signal-to-noise ratio that prevents the retrieval of PLDR at high altitudes, is it? That is, the measurements probably have sufficient SNR to retrieve a reasonably precise volume depolarization ratio, but the low values of the scattering ratio prevent converting it to PLDR with reliable accuracy because of the singularity in the conversion equation.

Actually, it is the small signal to noise ratio that limits the retrieval of the PLDR from the measured VLDR, but it is also correct that the smaller the backscatter ratio is, the higher the SNR must be to retrieve the PLDR with a given relative uncertainty. We include this in the sentence as:

"As the signal to noise ratio was too small to retrieve the PLDR with sufficient accuracy for overpasses at flight altitudes >7 km with low backscatter ratios, we restrict our evaluation to the extinction coefficient in those cases."

The switch from time (in Figures 1 and 2) to overpass number (in Figures 7) is not very user-friendly. Would it help to plot Figure 7 against time like Figs 1 and 2 (even though the symbols will not be linearly spaced)? Even if not, please consider annotating the x axis with lines indicating the dates of the major aerosol events identified in the text of the paper to help orient the reader to the new x-axis. A callout to the appendix in the figure caption might also help slightly.

Although we think, that the figure becomes less well structured and the comparison of in-situ and lidar measurements less clear, we agree that the change to overpass numbers is not very user-friendly. Thus, we followed the advice of the reviewer to change the x-axis to stay with the formatting of the former figures.

The lowest panel of Figure 7 is very busy and hard to read with so many different symbols, especially with many missing data points. Maybe consider moving the dust and non-dust extinction up to the PLDR plot on a second y axis. That would also have the benefit of making it easier to compare dust fraction to PLDR. Indeed it might be more informative to show it as dust fraction rather than as separate dust and non-dust components. In fact, the PLDR and dust fraction data could very well be a separate Figure 8, since they have nothing to do with the in situ comparison, as far as I can see, and separating them would make them easier to understand.

We agree that the dust fraction and the PLDR has nothing to do with the in-situ comparison. Thus, we follow the reviewer's advice to separate the comparison from the dust contribution. As we already investigate the dust contribution with respect to the dust fraction on the

backscatter coefficient and the volume, we decided to simply remove the dust AOD and non-dust AOD in Figure 7.

Indeed, could there be an intercomparison of dust fraction with in situ? In situ has a particle size distribution and coarse mode fraction, so is there any comparison available with some version of the dust fraction derived from PLDR?

To have such an intercomparison is beyond the focus of this publication, but the in-situ group is working on such activities.

Line 303. "In the lowermost layers, the values range..." Why is there no analysis or discussion of the agreement in these more important layers? There is sometimes disagreement beyond the error bars. This is probably not of much concern, but since there was a stated intent to compare the two, a comparison should be discussed both when there is agreement and when there isn't.

We added a discussion on the disagreements of the in-situ and the lidar derived values:

'Although the lidar and in-situ derived extinction coefficients show the same behaviour, differences between both methods are obvious on dust dominated days, with the in-situ values exceeding the lidar derived extinction coefficients. The largest differences we found was about 0.05 km^{-1} within the Saharan dust layer. The differences result partly from the different methods used to derive the extinction coefficient and partly from the different volume sampled by in-situ and lidar measurements.'

Line 304. "The lidar ratio ... helps to distinguish". This is either a non-sequitor or a typo since the sentences before this talk about extinction, and after it talk about PLDR.

The reviewer is right, this was a typo and is corrected now.

Line 306. Include some discussion of what happens when the PLDR is 0.32 while 0.3 was assumed as the value for pure dust. It looks like maybe no dust contribution is reported in that case (such as overpass #30)

We mentioned this topic in Section 2.4: *'It has to be considered, that a deviation of the actual measured PLDR and the one used for the type separation can lead to and over or underestimation of the contribution of the two assumed aerosol types. And thus, it is important to investigate the uncertainties of the type separation, including also uncertainties of the input values.'*, and discussed the resulting overestimation along with Figure 10: *'As we find a deviation of the measured value of the PLDR of 0.32 on 27 April 2017 and the input value used for type separation of 0.3 the retrieved dust fraction to the bsc and the volume is overestimated, leading to mean values slightly larger than 1.'*

Figure 7 caption. Which instrument are the dust and non-dust contributions from?

We added that those values were derived from the lidar analysis.

Figure 7. I believe the dust and non-dust contributions are derived via the PLDR and backscatter from the POLIS instrument. Why don't the dust and non-dust extinctions add up to the total 532 nm POLIS extinction? For instance, the first two low-altitude overpasses (#1 and #3) have very similar non-dust extinction (gray) and very different

dust extinction (orange), but very similar total extinction (green) (and in fact the one with larger dust component has smaller total).

Right, the dust and non-dust extinction coefficients were derived from POLIS measurements. Having only assumptions for the input parameters to calculate the dust and non-dust contribution (PLDR and LR for dust and non-dust component) can lead to uncertainties that, summing up the two values, lead to differences compared to the directly measured extinction coefficient. Following the suggestion of this reviewer, we removed the dust and non-dust extinction coefficient in this comparison.

Figure 7. How does overpass #32 have a larger dust extinction than total 532 nm extinction and how should this be interpreted?

Please see previous comment.

Lines 329-335. In the section describing how aerosol types were derived from AERONET, it's somewhat difficult to understand which parts are quoting observations of the Toledano et al. references and which are talking about the methodology for typing in this particular study. This confusion is exacerbated by the use of "about" preceding almost every number in the paragraph. This makes the numbers sound very vague whereas this paragraph needs to describe what discrete thresholds were actually used in the study to set the AERONET types in Table 2. I hope the paragraph can be revised to be more clear and quantitative.

We significantly revised the description of the AERONET classification to be clearer.

'For the AERONET based aerosol typing we use the scatter plot of Ångström exponent (440-870 nm) vs. AOD at 500 nm as proposed by Toledano et al. (2009, 2011). Values of the Ångström exponent (AE) larger than ~1.2 serve as indication for smoke/pollution, independent of the AOD. Ångström exponents of <0.5 serve as indication for dust (Toledano et al., 2009, 2011, 2019) or marine aerosols. Following Toledano et al., 2011, a threshold of AOD=0.15 is used to separate dust and marine aerosols. Measurement points with AOD<0.15 and AE<0.5 are classified as marine, while measurement points with AOD>0.15 and AE<0.5 are classified as dust. Ångström exponents between 0.5 and 1.2 serve as indication for mixtures. We further subdivide this value range in dust mixtures for AE larger 0.5 a value of the AERONET derived Fine Mode Fraction larger 0.5 (AE values of ~0.8). Values with Fine Mode Fraction < 0.5 (AE values >~0.8) and AE values smaller 1.2 are classified as polluted mixture.'

Figure 8. It would also be good if the thresholds for distinguishing types were shown as lines on the left panel.

We color-coded the different measurement points according to the type classification scheme.

Figure 8. It would probably help the reader if the points in the scatter plot were color-coded in some way that would help pick out specific events. Perhaps color code all of the points by date, or maybe color code only the points that correspond to the key aerosol episodes that are highlighted in this study (the Arabian and Saharan dust events and the pure pollution events).

We followed (partly) the advice of Reviewer 2 and removed the daily mean values. We kept only those points that were used in Table 2.

Section 3.5.2. Strangely, it's really not clear how the lidar aerosol type classifications are done, despite it being the main theme of the paper. I think but am not completely sure, that they were done using only the lidar ratio and PLDR, (and not relying on backtrajectories or other information, because that would make it impossible to use those for verification). Is that correct? If so, what are the thresholds? (Or if the method is more complex than simple thresholds, describe it here). Please also show the thresholds in Figure 9. Which lidar instrument is used for the classifications in Figure 9 and Table 2? It seems like perhaps both are included in Figure 9, but the table has only one column for "lidar".

The lidar based aerosol typing uses PLDR and lidar ratio as described in Section 2.4 Aerosol typing and aerosol type separation based on lidar measurements. A detailed description of the method including a figure with the decision tree and threshold can be found in the publications by Groß et al., 2013 and Groß et al., 2015. As the method is available and the information easily accessible we do not think that it is necessary to repeat it here.

Data from both instruments are shown in Figure 9, indicated by two different symbols. This is also stated in the Figure caption. For Table 2 we did not separate the results from the two lidar systems as (except for one value) both agree in their typing. Following the suggestion of this reviewer we added a more detailed description comparing the results of the two lidar systems, highlighting their good agreement. We added a note, that both lidar systems agree in the lidar based classification for Table 2.

Table 1. Extinction should be included in the table, since it's also shown in Figures and discussed in the text.

As the properties given in Table 1 are the values used for aerosol typing, we do not want to include the extinction coefficient, as it is not a property used for aerosol typing.

Table 1 caption. "For this evaluation" What exactly does that refer to? (I don't think it can mean the whole manuscript, since POLLY-XT does appear in figures.)

We changed the caption to make clear that the values given in the table were derived from nighttime measurements and removed POLIS, as also the Polly^{xt} night time measurements were used.

Figure 9. The blue (polluted marine) point at 60 sr looks rather odd. It seems like it is between points labeled pollution on both the lidar ratio and depolarization ratio axes. If this is real, I would like to read an explanation or discussion of this point (If the data are being classified by the two lidar quantities using thresholds, without using other non-lidar data, I don't see how this point and the black pollution point at about 0.12 PLDR could both be as they are shown. Maybe one is an error, or maybe I misunderstand the method. Either way requires a revision, even if it's just to make the explanation of the method more explicit.)

We corrected Figure 9.

(Discussion)

Line 400. "verified by the evaluation of air mass source regions". If the back-trajectories are used in both the "in situ" classification and the "lidar" classification, doesn't that complicate the ability to usefully compare them to each other?

We did not use back-trajectories in the lidar classification.

Line 410-411. (repeat from above) I would like to know more about this more difficult case where in situ typing indicated a pollution component in an airmass that the lidar could not possibly type as anything other than pure dust. Specifically what chemical or other measurements in the in situ suite indicated pollution? Is there an estimate of the mass fraction or any other measurement that would support the statement that it's a minor component (as stated in the text) or alternately "moderate" (as stated in the table) (And by the way, that also needs to be made more consistent). Would this amount of pollution be expected to impact either the radiative properties or the aerosol-cloud interactions? This could be an important conclusion, that depolarization indicating pure dust could still include measureable amounts of pollution, especially if it would be expected to have relevant impacts. It might be useful to highlight this case, and/or other "difficult" cases as case studies, besides the pure type cases highlighted earlier.

There is a separate publication by Teri et al. (doi.org/10.5194/egusphere-2024-701) handling this topic. We refer to this publication in our study. Especially in the supplement of this publication, the requested information can be found.

We corrected our expression about 'minor' and 'moderate' to be consistent.

Table 2. AERONET classifications include a surprising number of similar sounding categories, "Dust and pollution", "mixed dust", "polluted mixture". What is the difference between these, both qualitatively and quantitatively?

The reviewer is right, there are a number of similar categories which actually do not reflect exactly the classification by Toledano et al. but were driven by the a priori information from source investigation and in-situ measurements. We revised the AERONET classification part to be closer to the method proposed by Toledano et al. and with that independent from a priori information influencing the classification. With that we reduced the number of categories to marine (not available in this study), dust, dust mixture, polluted mixture and pollution/smoke. We revised the description of the AERONET classification to be clearer.

Table 2. AERONET 11 April. I'm surprised by "mixed dust". In Fig 2, this date has some of the lowest AE. Wouldn't that make it unambiguously pollution based on how the AERONET classification was described? (As I've said already, I don't find the description of the method very clear, so I could very well be wrong, but would like to understand better).

The low AE in the AERONET measurements point towards larger particle sizes (likely dust) but the mean AE is just above the threshold of 0.5 to be classified as Dust Mixture instead of dust.

Table 2. Lidar column. Which lidar's classification is this?

Please see prior comments

Table 2. Lidar column. "Mixed pollution" (11 April and 25 April) is not a category in Fig 9. What category is this meant to correspond to? Please make the category labels consistent across all figures and tables.

Thanks for pointing this out. We corrected it.

Table 2. "mixed pollution" 25 April. Why would this be "mixed pollution" rather than just pollution, where it seems to fall on Figure 9? (if the values are 0.03 PLDR /60 sr at 355 nm, as I think they are)?

This is right, and we corrected it.

Table 2. FLEXPART 25 April. What is the meaning of "SS(dust)"?

The sea salt (SS) contribution was independent from '(dust)'. There was a chance for some minor dust contribution. However, the way we indicated it in the Table was unclear. Thus, we corrected it.

Line 450. "while FLEXPART estimates only about 30 ug/m³". What about dust mass fraction or total mass? That is, does FLEXPART agree better with the total mass, but disagree that the mass is mostly dust, or have a similar fraction but disagree with the total (or neither)?

We checked the total mass of FLEXPART but it does not show a better agreement. Thus, we still assume that FLEXPART underestimates the dust mass.

Figure 10. What is the interpretation if the volume fraction is bigger than bsc fraction, or the reverse, or when there is a large difference? More specifically, how is it possible to have a case like overpass #10 where the volume fraction is apparently zero but the backscatter fraction is near 60%?

Looking at the uncertainty range in the retrieved bsc fraction and volume fraction, we think it would be going to far to interpret differences in the volume fraction and bsc fraction. To highlight that, we included the uncertainties in Figure 10, showing that the differences are within the uncertainties. We also added some text to make that clearer.

'Both dust fractions agree well within the uncertainty ranges, except for 11 April 2017 when the dust fraction to the bsc significantly exceeds the volume dust fraction. This difference might be caused by a wrong assumption of the contributing types, and thus of the chosen conversion factor.'

(Summary and Conclusion)

Line 464-478 The summary and conclusion section has a confusing flow (especially the first long paragraph). It would be good to reorganize it to group similar thoughts together. It might be useful to have a larger number of distinct paragraphs.

We restructured the Summary and Conclusion with specific focus on the first paragraph.

Line 465 "compared to the lidar ratio found for Saharan dust". This would preferably be followed immediately by giving the the lidar ratios for Saharan dust.

We added the values for Saharan dust.

Line 467-469. Please put the mean values of the PLDR of the measured Arabian dust immediately after mentioning it (at the beginning of the sentence). As it is, it reads as if the quoted values are the values "for Saharan mineral dust close to the source region".

We followed the reviewer's advice.

Line 469. There also seems to be a typo, since the numbers for 355 and 532 are the same here, but different in the abstract.

The sentence was misleading. We changed that by also following the reviewer's former advices.

Line 470 and 471. "their PLDR of" and "their lidar ratio". This sentence is also constructed ambiguously, with the grammar suggesting that these refer to pollution, or perhaps to the Saharan AND Arabian layers. Only by cross-referencing with the abstract can I see these are the values for the Saharan observation.

See previous answer. We followed the reviewer's advice and changed it.

Line 471-473 and Line 467. These sentences about agreement with previous studies should be placed together or combined. These two very similar ideas placed far apart is part of what makes the flow hard to follow.

We moved the later sentence to directly follow the first one.

Line 474 "0.05" at 532 nm. The abstract says 0.04.

We corrected that.

Line 481-482. "The derived volume fraction of the dust... showed a lower contribution to the total volume compared to its contribution to the optical properties". I don't understand this conclusion. Perhaps spelling out what optical properties would help, but I think it is probably referring to the lidar-derived extinction and PLDR. Since the derived volume fraction is assumed from a simple formula applied to the optical properties, is this just saying that the relationship is a curve rather than proportional? (not a new discovery).

The optical properties are just mentioned in the prior sentence.

Line 487. "Although it could predict the dust transport in general". Was this discussed in the Discussion section?

No, we did not include a special discussion on dust transport. We believe that this is beyond the scope of this paper.

Line 486-488. "Models generally assume". This is the first appearance of this idea; it should be in the Discussion section.

We followed the reviewer's advice and added this also in the discussion.

Line 491-492. "in-situ derived total mass concentration exceeded". In the Discussion a partial explanation of this was given, that it reflects different sampling resolutions. It would be good to repeat that in the conclusion.

Thanks for this helpful advice. We repeated the explanation here.

Minor points:

Line 40 replace "process" with "processes"

Done

Line 41 replace the comma between optical and microphysical with "and"

Done

Line 205 "indicated by the greenish to reddish colors": please indicate which panel is being referred to.

Done

There should be a callout to Fig 2, probably somewhere between lines 200 and 210

We included that.

Line 217 replace "neglectable" with "negligible"

Done

Line 298. Probably intends too small rather than too large.

Corrected

Line 304. Please split this paragraph into two or more, since it covers a number of different topics. I suggest splitting at line 304 after "campaign".

Done

Line 309. Consider including the reference to the Tesche et al. paper(s) that give the methodology for deriving the dust contribution here again.

All the references to the methodology are included in the methodology section. This includes of course the publication by Tesche et al.

Figure 9 color legend. Please make the dots in the legend bigger. It is very difficult to distinguish the colors. The text in the legend is a bit on the small side too.

We changed that.

Table 2 and Table 1. Include overpass number as one of the columns, to make it easier to compare with Figure 7.

That is a very good suggestion. We added the overpass number in Table 2. However, we did not add it in Table 1 as these values are independent of the overpasses and refer to nighttime measurement.

Table 2 caption, describe why some rows have no lidar or FLEXPART classifications.

For those layers no lidar classification and FLEXPART classification was possible. We added this information in the figure caption. We still keep these height ranges as it completes the comparison with the AERONET classification.

Table 2 and Figure 9 (possibly elsewhere too). "Dust mixture" in Fig 9 is "Dust mixture (marine)" in Table 2. The position on Fig 9 does seem to suggest this is meant as exclusively a marine + dust mixture. If so, consider changing the name to reflect the more specific meaning.

Thanks for pointing this out. We changed it.

Figure 10. blue "BSC" dot missing at overpass #30. Why?

The value was out of the range shown in the Figure. We corrected that.

Line 477. "the different classification schemes". I suggest spelling out what these are, to help the summary section stand by itself.

We mentioned the methods in the prior sentence, but to be clearer, we changed the phrase to 'We compared the findings of the lidar based classification to aerosol typing based on in-situ measurements, sun-photometer, and FLEXPART transport simulations. The different classification schemes showed a very good agreement, although the sun-photometer based classification can only give a column integrated value.'