Response to reviewer 3

We thank the reviewer for their insightful reading of the paper and many comments. This reviewer says "The paper provides a much needed quantification of aerosol absorption properties derived by satellite products, by inter-comparison and comparison with AERONET".

My main criticism is a poor description of the algorithms in the method section, except for the POLDER-SRON part, and the lack of interpretation of the results.

The descriptions of the algorithms is deliberately brief because we want to focus on the evaluation and intercomparison. We will add more detail to the FL-MOC and GRASP sections, and present a first interpretation of our results in the Discussion, referring also to the papers this reviewer suggested in his next comment. In particular, we want to add the following to the discussion:

"The two POLDER products perform better against AERONET than the other two products, with typically (but not always) higher correlations, smaller biases and regression slopes closer to 1 (one) for all three parameters AOD, AAOD and SSA. However, dearth of measurements makes it very difficult to 1) meaningfully compare evaluation metrics amongst the products and 2) draw global conclusions. Theoretical evidence (Hasekamp et al. 2007, Hasekamp et al. 2010, Hasekamp et al. 2019a) suggests that retrieval schemes for absorptive properties will benefit from using polarisation measurements at multiple view angles which would support the idea that the POLDER products perform better. In addition, the OMAERUV product is based on measurements from a sensor with substantially larger pixels than POLDER and will struggle to resolve the fine-scale structure of aerosol plumes."

We'd like to point out that a full interpretation of our results is outside the scope of a single paper that already is quite large and concise. It would require dedicated numerical experiments, as for instance done by Holzer-Popp et al. *AMT* 2013. Even for AOD products it is still challenging to attribute retrieval errors in actual data quantitatively.

The paper's claim to be the first to show this may be correct, but there have been a number of papers in the past to show the minimum amount of information content that is needed before AAOD and SSA can be expected to be retrieved with some degree of accuracy.

These papers are theoretical studies, assuming e.g. only random errors at various stages of the retrieval process (co-authors O. Dubovik and O. Hasekamp conducted such studies). These studies did not consider the possibility of long-term averaging of data. Our study uses actual satellite data and does consider the beneficial impact that averaging has.

Still these theoretical studies show the need for polarization measurements at multiple view angles, probably explaining why the POLDER products appear to do better. As mentioned earlier, this is now discussed in the summary.

For the MOC (What is this? No description given) and OMI algorithms no information is given at all, only a reference to other papers.

MOC (or FL-MOC as we call it in the paper) stands for Fu-Liou MODIS OMI CALIOP. This abbreviation is now explained in the updated text. Its algorithm is briefly explained in Sect. 2.1.1. It is not a retrieval per se but a consistent reinterpretation of the combined data within their stated uncertainties. We have added some more detail.

For the GRASP algorithm it should be made clear in what way it differs from the POLDER-SRON

A good idea. Although we already discuss the (in)dependence of the SRON and GRASP algorithms in Sect. 2.1.6, they differ in many ways: different cloud screening, different solution methods, different estimation of surface contribution. We have added additional explanation in the Sect. 2.1.6

Currently, the authors only present the errors or biases, but no explanation in terms of the algorithms' treatment of the different derivations of the AAOD and SSA.

See our previous comments.

A discussion of the independence of AERONET observations should be included here.

It is not clear to us which independence the reviewer refers to? Clearly AERONET has its own limitations and makes its own assumptions but these are not related to any of the satellite retrievals we discuss in this paper. The uncertainty of AERONET inversions we discuss, and we refer to several papers (incl. Dubovik et al. 2000 and Sinyuk et al. 2020) that analyse this uncertainty in great detail.

The paper treats the accuracy of AOD, AAOD and SSA derived from satellites. These are all connected parameters, but should not be interchanged, which seems sometimes the case.

We deliberately switch back and forth between AAOD and SSA to provide a better picture of how these products behave (if we discuss SSA, results for AAOD can be found in the supplement and vice versa). Ofcourse there is a strong connection (SSA=1-AAOD/AOD) but they need to receive separate evaluation. Even if we know the uncertainties in AAOD and AOD, this does not teach us anything about SSA uncertainties. The reason is that AAOD and AOD errors may or may not be correlated. In the first case, SSA uncertainties can actually be fairly small.

l. 426. 'Over ocean, SSA products tend to correlate better than over land. The two POLDER products correlate better than any other satellite pair (r =ấ Lij 0.8 over ocean for AOD > 0.75).' The next paragraph starts like this: l 432. 'ost surprisingly, POLDER- GRASP-M and POLDER-SRON show a fairly systematic difference in SSA (-0.04), independent of AOD (there are regional variations).'

How are we to interpret these seemingly contradictory statements? Are we not talk- ing about SSA? Are " two POLDER products" not the same as "POLDER-GRASP-M and POLDER-SRON" SSA? Or are the statements not contradicting? Probably the latter, but the reader has

to check his/her own sanity a couple of times first, before this become apparent. In a technical paper like this consistent phrasing and grammatical structuring is even more essential than normally, and the lack of that in the current paper makes it hard to read.

They are not contradictory statements. Two products can show a large systematic difference (a bias) and yet correlate highly. The simplest example is the case where both products actually agree with the truth except for a bias in one product.

Other minor issues are listed below:

2 l35 heating can also destabilise the boundary layer (Johnson et al, 2004), semi-direct effect are now called fast adjustments and can be both negative and positive in forcing.

Thanks. We are familiar with this paper but somehow forgot to include it. We suggest to keep the term semi-direct effect because it is most often used in the context of absorbing aerosol. Fast adjustments can also be the results of non-absorbing aerosol (see e.g. Zanis et al. ACP 2020).

l158 The terms bidirectional reflectance distribution function (BRDF) and bidirectional polarisation distribution function (BPDF) are not explained.

We have added explanations of the acronyms.

Figure 4. Half of the difference plots are the same (but vv) and can be removed.

It is true that half of them are the same, but we feel that this layout makes it easier to intercompare datasets.

l305 "The scatter plots show good correlation with AERONET." This is a meaningless term. The idea is to quantify the goodness, or accuracy. Please, rephrase to The scatter plots show the correlation of the satellite AOD with AERONET AOD.

We do not understand why this is meaningless. Obviously, a good correlation does not preclude the possibility of significant biases. That is why we also study biases. However, a good correlation suggests that the satellite retrieval is sensitive to the same characteristics of observed scenes as AERONET.

On the other hand, a small (global) bias does *not* prove in any way that a product is suitable. See Schutgens et al. *ACP* 2020 on how global biases in AOD are meaningless indicators of product performance.

l326 product -> products

Corrected.

l346 Section -> section

Corrected.

l349 0.006

l. 365 If cloud contamination if such a big problem, why is it not (additionally) removed?

But it is removed, as best as possible. However, cloud screening is not a straightforward process and products often differ more in their estimate of cloud cover than in their estimate of AOD (Schutgens et al. *ACP* 2020).

Figure 12. For POLDER-GRASP-M an additional minimum AOD threshold is used before calculating AAOD and aggregating SSA (l 166.). The threshold is not mentioned in the paper. However, it is not 0, as suggested in the caption of Fig 12. This should be clear in the Figure and/or the text.

The thresholds are now mentioned (AOD at 440 nm > 0.3 over land and AOD at 440 nm 0.02 over ocean). As we originally explained: we assume the SSA aggregate describes the same scene as the AOD aggregate (calculated without AOD threshold) and from these two a new AAOD is calculated. This new AAOD performs better against AERONET than the original one. As a consequence, the new dataset contains AAOD values at AOD lower than 0.3 (over land) or 0.02 (over ocean).

l 432. 'Most surprisingly, POLDER-GRASP-M and POLDER-SRON show a fairly systematic difference in SSA (-0.04), independent of AOD (there are regional variations). A major exception would be cases over the deep ocean at low AOD (< 0.1) where this bias dissappears.'

Is this not a result from the fact that no absorbing aerosols are left over the 'deep' oceans? I expect deep oceans refer to those remote parts far from the land (thus aerosol sources), where only clouds and marine aerosols are left ? One would not expect any signal left for those areas. In that case it would make sense refer to 'remote oceans' or something.

Thanks for bringing this up. After some further investigation, we do not think this is a correct statement. Rather, at low AOD over ocean there appears to be a hemispheric contrast in this systematic difference (already visible in Fig~S1) whose cancelling leads to a small global systematic difference. Currently we have no idea what may cause this systematic difference. The text has been adapted.

dissappears -> disappears

Corrected.

l 440. 'than is present in' -> compared to

Corrected.

l 442. 'It will not be easy to increase Inversion L2.0 observations' -> "It will not be easy to increase THE NUMBER OF Inversion L2.0 observations"?

Corrected.