Response to reviewer 2

We thank the reviewer for their many useful comments. The reviewer says this is a paper on "an important topic".

However, a discussion or even mention of the acceptable or scientifically required uncertainty level in single scattering albedo (SSA) is lacking in this paper.

Some discussion regarding the accuracy in SSA required for aerosol radiative forcing or other applications need to be included in this paper.

We now mention the GCOS (2011) SSA requirement of accuracy within 0.03 and stability within 0.01 per decade. The rationale for these requirements seems based on typical regional and yearly variations in SSA. However, SSA requirements are different for different applications and the GCOS requirements are meant to provide a general broad estimate (Popp et al. *Rem Sens.* 2016). In part 2 of our study we will show that current SSA capabilities allow useful evaluation of models. We discuss this in the Introduction.

Lines 46-47: Your current sentence: "From this inversion, columnar properties AOD and AAOD can be derived." No, this is inaccurate since total atmospheric column spec- tral AOD are measured from sun photometer direct sun observations. The AERONET inversion matches the measured AOD almost exactly since very high accuracy in measured AOD is assumed in the inversion algorithm.

We have corrected the text.

Lines 65-66: It should be mentioned here that SSA uncertainty is significantly larger at longer wavelengths, as this is pertinent to estimation at the 550 nm wavelength. In the paper you currently only give the uncertainties in 440 nm from Sinyuk et al. (2020), while the uncertainty at 675 nm is also relevant and are provided in this same reference.

This is a good point and we have added this information. However, we find no systematic increase in SSA uncertainty with wavelength in the Sinyuk analysis: SSA uncertainties for AOD (at 440 nm) =0.2 from 0.037 to 0.048 at 440 nm and from 0.035 to 0.045 at 675 nm.

Lines 67-68: It is strange to continue using this outdated estimate from Dubovik et al. (2000) as the SSA uncertainty actually decreases below 0.03 as AOD increases above 0.4 at 440 nm (see Sinyuk et al. 2020)). Also the Dubovik et al. (2000) estimate is only for 440 nm, while Sinyuk et al. (2020) provides values for all 4 retrieval wavelengths in the V3 database.

The Dubovik study was more limited and because of that generated more "quotable" uncertainty intervals. We have added a range of uncertainties provided by Sinyuk et al.

Line 75: Be clearer here that the 2nd part is a separate paper from this one (I think).

Correct, we have modified the text.

Line 187-188: Again, please give the AERONET SSA uncertainty estimates at 675 nm from Sinyuk et al. also since both wavelengths are being used to interpolate to values at 550 nm, for subsequent comparison to satellite.

See our previous comments.

Line 190-193: It is not strictly accurate to say that the AERONET values of SSA were underestimates since the in situ data also have significant uncertainty due to numerous assumptions and also have less sensitivity at low aerosol loadings. This uncertainty of in-situ data and the lack of a 'gold standard' for SSA should be conveyed here.

We did not mean to imply that the flight campaign data are more reliable but we see how that could be inferred. Text has been modified.

Line 260-261: Please be clear here, are the Inversion L2 data from AERONET only for AOD(440)>0.4? Note that AERONET produces L2 inversions of aerosol size distri- butions for all AOD levels. It is only the refractive indices and therefore SSA that are limited to AOD(440)>0.4.

They are the Inversion V3 L2 data for AOD(440)>0.4, this has been added to the text.

This Figure 3 may in part just be comparing results for dif- ferent AOD levels since only moderate to high AOD are included in L2 refractive index retrievals, while most data measured globally is of AOD<0.4. If so then please include in the interpretation of this figure some discussion as a comparison of lower AOD to higher AOD cases.

That would be our interpretation as well. We will clarify this in the paper. The point is of course that 1) depending on the chosen dataset, evaluation results will differ; 2) there is no way of knowing what this implies for AAOD (apart from modelling studies as done by Dubovik et al. and Sinyuk et al.)

Line 277: Can you please explain why it would be expected that POLDER-GRASP-M has relatively low SSA over land while it is expected that OMAERUV has relatively high SSA over land? If there are references to previous investigations then they should be cited here.

This is a consequence of the AAOD results discussed earlier in the same section. If satellite AOD is rather similar, a lower value of AAOD will translate into a higher value of SSA. Text has been clarified.

Line 289-290: Please elaborate with another sentence or two here in describing what aspect of the non-collocated data accounted from such a large change in correlation. Was it different regions or different time periods sampled?

For sure, a different time period because now the collocated POLDER data only cover 2006 (SRON's data period). GRASP evaluation for 2010 shows worse agreement with AERONET than for 2006 and 2008, in particular lower correlation and high RMSD. However, only for SSA is the difference statistically significant (per year, GRASP has about ~ 150 collocated data with AERONET so statistical noise cannot be ignored). Interestingly, the AERONET data collocated with GRASP shows slightly higher AOD (10-20%) in 2010 than the other years. There is a shift in AERONET sites with GRASP collocations for 2010 (e.g. more Amazon and coastal sites) but dearth of data makes it impossible to say anything about the consequences.

It can be very difficult to establish exactly what underlying factors in different samplings lead to differences in evaluation statistics. For another example, see Schutgens et al. *ACP* 2020. There we had much more data to work with but still could not identify said factors. Obviously different sites at different times are involved in the different samplings but that in itself explains little.

We have added a few lines explaining these issues.

Line 318-319: So, is this filtering of the POLDER-GRASP-M significantly different from all the other data sets? If so, then what is the value in including that particular dataset in Figure 8 since it may therefore be very misleading. Additional discussion is warranted. In section 2.2.4 you discuss some sampling issues regarding SSA and AAOD with this particular satellite product but it is unclear whether this applies to the AOD dataset.

All satellite datasets apply filtering, in the case of POLDER-GRASP-M it is just a bit more transparent. Filtering among dataset can be very hard to compare because the raw data will be very different (wavelengths, view angles, pixel sizes). We have added more explanation in the description of the GRASP algorithm.

As explained in the data section, the aggregated AOD, AAOD and SSA datasets from a single retrieval scheme have exactly the same sampling (time, longitude, latitude). For all retrieval schemes except GRASP, the aggregated data are based on original L2 retrievals of AOD, AAOD and SSA that have (again) the same sampling. The case of GRASP is explained in Section 2.1.4. We recognize that the original wording was confusing and we have tried to improve this.

Line 332: Please note in the text that the scatter deceases significantly for AOD>0.3 which is similar to the L2 threshold in AERONET. Also, I assume that the wavelength of all parameters in Figure 9 is 550 nm. This should be included in either the plot labels or in the figure caption to make it clearer the readers. Also please explain better the green color bars in the left-most 3 columns of Figure 9. Include this clarification in the figure caption.

As explained in the paper, all data are at 550 nm. The green colourbars show percentage of total number of observations. This has been added to the caption.

Line 337-338: At low AOD the major sources of uncertainty in AERONET retrievals of imaginary refractive index and SSA are biases. These are sky radiance calibration (this is independent of the direct sun calibration), extra-terrestrial solar flux, and BRDF from a MODIS product. The analysis in Sinyuk et al. (2020) of the U27 (see paper) only considers these bias errors and does not attempt to evaluate random errors. Therefore you can utilize the Sinyuk et al. paper to get an estimate of the biases in AERONET retrieval data. However these biases for a given site and deployment can be either high or low since it is not known what the direction of the bias in calibration and BRDF are for a given site and date. Additionally the bias direction of the solar flux error is also unknown however this remains constant for all sites and dates.

This is interesting to know. Maybe the reviewer can help us understand better. We wonder e.g. how biases are defined? MODIS BRDFs show a seasonal cycle which opens up the possibility that the errors in AERONET SSA due to incorrect BRDFs may not be biases at all on sufficiently long time-scales. Why is the solar flux error constant for all sites and dates (by the way, we assume that the reviewer means it is constant at a single site but varies by site)? Also, from our reading of the Sinyuk et al. paper, the errors may be termed biases but they are calculated as if they are random errors (i.e. uncertainties are assumed in input parameters). So the Sinyuk “biases” can only be interpreted as the standard deviation of possible biases.

Line 341: This should be taken with more than a grain of salt since the data sampling at high AAOD is extremely sparse for POLDER-SRON. This statement ("it seems to underestimate AAOD by 25% at high AAOD") seems much too strong given the weak data sampling constraints here.

Here, we are only describing what the data tell us. Actually, we often point out in this paper that the low data count precludes extrapolation to the global case. Which we then try to address by a systematic satellite intercomparison.

Line 346-347: Note that AERONET retrievals of SSA also have a maximum value that is only slightly higher than 0.99 due to a minimum constraint on the value of the imaginary refractive index.

We have found values as high as 0.996 in the AERONET L2.0 data. A histogram of values suggests no cut-off. In the AERONET L1.5 data we find a maximum of 0.997. Here the histogram does suggest a deliberate cut-off.

Line 364-365: Yes, agreed that cloud contamination is a likely issue with the relatively large satellite pixel sizes for POLDER and OMI. Also please clarify if the global statistics (the first - labeled in purple in Fig 10) are for all AOD levels or for only AOD(550)>0.25 as in the third comparison in green. It would be very useful to look at the GLOBAL histogram for different AOD levels and include information from that into the text of the paper or even possibly added as part b to the figure.

As the last line of the caption says, AOD > 0.25. We now clarify this in the main text. We have looked different lower thresholds and see that the GRASP vs SRON bias is independent of it unless the threshold becomes very low (in which case the bias disappears).

Line 368: Please avoid clipping off the x-axis labeling in Fig 11 of all but one plot. It would be easier to read if all were labeled.

The final figure will be shown as a column, in which case a single x-axis allows us to dedicate more space to the actual graph.

Line 374-375: The diagonal x-axis labeling is confusing and awkward. Please try to improve the readability and visual discrimination between these comparisons in Figure 12.

We have tried different versions of this graph and this one seems best. The different pairs of satellite are clearly separated in the figure. Ultimately this will be a one column figure.

Line 433-434: It should be noted in the manuscript that the good agreement cannot be due to actual skill in SSA retrievals from the POLDER algorithms over these low AOD conditions over ocean when the absorption signal is far too low for any reasonable accuracy in remote sensing retrievals. This good agreement is likely just due to other factors such as assumptions and/or constraints that were made in the algorithms. Also note that cloud contamination is probably a greater issue over oceans than over land.

We will modify the text. We do not see a constant difference in SSA between GRASP and SRON but any difference distribution will have a bias of -0.04. See also Figure 9.

Line 441-442: This statement does not make much sense. It is not the sensors but the physics of the retrieval problem that limits the accuracy of the SSA retrievals. Also cal- ibration uncertainty is a major factor in SSA retrieval uncertainty and this not a sensor problem per se. Another major

factor in the retrievals is uncertainty in the underlying surface BRDF, especially at low AOD levels, and this has nothing to do with AERONET sensors, but is a required auxiliary input data set.

This sentence referred to the *number* of observations, which is strongly affected by the scanning strategy which in turn is partly dictated by the sensors. The sentence omitted the word 'number' and indeed other aspects affect the number of observations as well so we have rephrased it.

Line 442-443: Please be clear here that you are only referring to the impact of the lower AOD threshold imposed in L2 and not some other aspects that exist between L1.5 and L2. Note that L2 retrievals exist for all AOD levels but only for the size distribution retrievals. Therefore the investigator can match all L1.5 retrievals with these L2 low AOD retrievals in date and time to already get a SSA product in L1.5 that has all the quality controls and cloud screening of L2 except for the AOD threshold levels. The text of the paper needs to be revised here to correct this misunderstanding.

This is a good point and we've included it in the paper.