Response to reviewer 1

We thank reviewer 1 for their detailed reading of the paper. Their numerous suggestions for textual improvements have almost all been applied. Reviewer 1 summarizes the paper as “clear and well written and interesting. “

Line 7 - Abstract - "2) their application to the evaluation of AEROCOM models." Should make a little clearer that no models were used in this paper. The first paragraph of the abstract does that but then this sentence muddies the waters.

Ok, we have rephrased this to read: “This study consist of two papers, the current one that deals with the assessment of satellite observations and a second paper that deals with the evaluation of models.”

Line 2 (and line 56) "several satellite products of AAOD have appeared" maybe change ’have appeared’ to ’have been developed’?

Ok.

Line 4 (also line 91) "super-observations" are they super because aggregated to 1x1x30min?

Yes, they are not better than the original observations but just aggregates. It’s a common term in data assimilation.

Line 33 "The species that absorb most visible sunlight" change to "The species that absorb the most visible sunlight"

Ok.

Line 38-29 "In particular over bright surfaces (ice, deserts, clouds) can the forcing due to absorbing aerosol be significant" change to "In particular over bright surfaces (ice, deserts, clouds) the forcing due to absorbing aerosol can be significant"

Ok.

Line 41 "black carbon may affect the Hadley cell" affect how or what - Hadley cell circulation?

Yes, “circulation” now added.

Line 44 "absorptive aerosol" I prefer ’absorbing aerosol’, but I’m not sure if there’s official agreement on this!

We don’t think there is an official agreement on this. However, we were inconsistent in our choice of words. We changed everything to “absorbing”.

Line 47-50 - Could cite Laj et al. AMT, https://amt.copernicus.org/articles/13/4353/2020/, 2020 for these global surface absorption measurements. They present a review of the available data.

Ok.

Line 50 "Moreover, these are surface measurements." This is true, but perhaps should state why this is a problem? Surface insitu measurements do have advantages over AERONET and satellite

retrievals in that they operate continuously (day/night regard- less of clouds) and are less limited by loading requirements. They are definitely sparse though!

Ok. The sparseness was mentioned in the preceding sentence. We have now added "surface measurements that do not measure the full atmospheric column"

Line 64 "error prone" is it that they are more error prone or just more uncertain? (I'm

not a stats person so not sure those are the same or different!)

Uncertain is a better phrase. Replaced.

Line 62-70 perhaps comment on whether anything is known (or not) about bias in AERONET retrievals of AAOD/SSA rather than just on uncertainty in the retrievals?

That is a very good question, and an important one. However, my discussions with the AERONET team gave me the impression this is not clearly understood. As a result, the given uncertainties may be site-specific biases or random errors. In reality, they will probably be a bit of both. Our study suggests that satellite retrievals contain both biases and random errors (the latter amenable through averaging). I have added some explanatory text.

Line 72 "AERONET hardly covers" change to "AERONET only sparsely covers" or something like that.

Ok.

Line 84 "observational model datasets" change to "observational datasets"?

Thanks for spotting that.

Line 90 is 'L2' defined or a well-enough known abbreviation? later, on line 103, it's spelled out as level 2.

Within the satellite community it is well understood. We've added a brief explanation and a reference. L2 data are estimates of geophysical variables on the spatio-temporal sampling pattern of the radiances.

Line 96 therefor –> therefore

Ok.

Line 101 should 'MOC' be defined (and FL-MOC)? Also section header is 'FL-MOC' but the text in this section just uses 'MOC' but later in figures and text it's referred to as FL-MOC.

Ok.

Line 132 'provided' instead of 'provides'?

OMI stills operates.

Line 140 "also the fraction of spheres is included in the" change to: "the fraction of spheres is also included in the"

Ok.

Line 144 define BRDF?

Bi-directional Reflectance Distribution Function (https://en.wikipedia.org/wiki/Bidirectional_reflectance_distribution_function), a more realistic modelling of surface reflectance than the Lambertian assumption. Acronym is now explained in the text.

Line 156 "Aerosol is assumed an" change to "Aerosol is assumed to be an"

Ok.

Line 157 "aerosol components and are retrieved" change to "aerosol components which are retrieved"

Ok.

Line 158 define BPDF?

Bidirectional Polarisation Distribution Function, see als BRDF. Acronym now explained in text.

Line 158 "The aerosol is assumed a mixture" change to "The aerosol is assumed to be a mixture"

Ok.

Line 192 "Andrews et al. (2017) only had observations over two sites" Andrews 2017 did include comparisons of insitu flight profiles from other sites in addition to the two main sites they studied.

That is true. We've changed the text.

Line 231 change '&' –> 'and'

Ok.

Line 242 add comma after 'i.e.' also after 'e.g.' on various lines (question for editor?)

Apparently this differs in American (comma) and British usage (no comma). We'll stick with no comma for now.

Line 275 "Rocky mountains" change to "Rocky Mountains"

Ok.

Line 280 "The impact of AOD will later be discussed." change to "The impact of AOD will be discussed later."

Ok.

Line 284 extra space before the word 'which' in parentheses

Removed.

Line 316 "observations and was" change to "observations and so it was"

Ok.

Line 324 "underestimate AOD and AAOD" change to "underestimate AERONET AOD and AAOD"

Ok.

Line 324 "amount in case" change to "amount in the case"

Ok.

Line 326 product –> products

Ok.

Line 347 "have hard cut-off" change to "have a hard SSA cut-off"

Ok.

Line 350 put '2019a' in parentheses

Ok.

Line 360 "corrollary" check spelling - only 1 r? i.e., corollary

Ok.

Line 370 "satellite SSA still" change to "satellite SSA values still"

Ok.

Line 386 put 'in general' in commas: ', in general, '

Ok.

Line 402 "and use them to evaluate AEROCOM models" This line in the conclusions suggests that AEROCOM models are used in this paper. Perhaps rephrase and say "in preparation for evaluation of AEROCOM models" instead?

Ok.

Line 409 "could suggests" change to "could suggest"

Ok.

Line 410 "to obtain best" change to "to obtain the best"

Ok.

Line 416 In conclusions refer to 'AQUA dark target', but in text refer to AQUA-DT. Per- haps be consistent?

Ok. We now use Aqua-DT.

The appendix is weirdly interspersed with the figures.

I suppose this will improve with the final version but I'll keep it in mind.

Line 722 "observations of AOD" but then in next sentence require AOD and AAOD. Should it be "observations of AOD and AAOD"?

Yes, corrected.

Line 727 (and 733) "(here 30min x 1 x 1) exceeds" in the main text use format "1 x 1 x 30 min" make consistent, i.e., length x length x time or time x length x length

Ok, changed to 1 x 1 x 30 in appendix.

Line 735 "MAN data" It makes me laugh to ask, but what is MAN data? do you mean SAT data?

Maritime Aerosol Network, basically AERONET on ships, see https://aeronet.gsfc.nasa.gov/new_web/maritime_aerosol_network.html . Acronym now explained.

Throughout the text, the words 'criterion' and 'criterium' are used. I'm not 100% sure if they have exactly the same meaning or not, but they seem to be used same way. Maybe just choose one?

Criterion seems to be the correct one, unless one speaks of cycling in which case criterium is a valid possibility https://en.wikipedia.org/wiki/Criterium .

Figure 5 - indicate what the slashes on the Taylor diagram points represent.

These indicate biases (normalized to the standard deviation in the reference dataset), see also Sect 3.1. Explanation and reference to Sect. 3.1 now included.

Figure 7 - what are the 'remainder' sites if they are not land or ocean? Remainder sites not mentioned in text. Also explain what 'OLSB s' is in figure legend (the rest of the abbreviations in the legend list were obvious)

These sites cannot be identified as ocean or land according to our criterion, e.g. cases where 50% land and 50% ocean. "OLSB s" stands for Ordinary Least Squares Bisector slope. It appears a small section on error metrics was dropped from the text and has now been included again.

Figure 8 - "products used Schutgens" change to "products used in Schutgens"

Ok.

Figure 9 - "except right-most column" change to "except the right-most column"

Ok.

Figure 11 - work on arrangement of plots and make sure x-axis label shows on all of them

In the final version, these figures should appear in a single column and only the bottom figure will have a visible x-axis label and values.

Figure 12 - larger font at top? changing text so not angled might provide more space

This is difficult to do without making the bottom labels spread all over the page. The figure is meant for a single column or it would take up a lot of white space.

Figure A2 - caption says "both observations and model data in this paper" but there was no model data in this paper.

Corrected.

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.

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.

Response to review nr 4

We thank the reviewer for their useful comments and questions. We hope we have been able to use them to improve the paper. Reviewer 4 describes the paper as "very meaningful for the comprehensive comparison between these AOD, AAOD and SSA among POLDER-GRASP-M, FL-MOC, OMAERUV and POLDER- SRON"

Line 100, FL-MOC is an acronym for what?

Fu Liou MODIS OMI CALIOP. The acronym is now explained in the text.

Section 2.1.5, the DirectSun dataset and its validation is detailed in terms of the range of AOD values and type of aerosol. However, the dataset used in this study is not DirectSun product itself but a dataset (Kinne list) that developed using DirectSun product, it is unclear to be the uncertainties of Kinne list and its uncertainties. Please add more direct discussion of the data quality of Kinne list.

The Kinne list is only a list of AERONET station names that have been deemed of better maintenance quality and better spatial representativity than other stations. The data themselves have not been filtered or modified. Official AERONET uncertainty estimates should be applicable.

Line 199, what's suitability here means?

Representativity. The paragraph details several studies into representativity of AERONET sites.

Line 200, be consistent with DirectSun or Direct Sun through the manuscript.

Ok, changed to DirectSun everywhere

Section 2.1.6 this part is not clear. Are POLDER-GRASP and POLDER-SORN sharing the same treatment of surface reflectance?

Only in the sense that they both use the same *functional* form for the BRDF (Litvinov et al. 2011). The parameters of this BRDF are estimated independently by GRASP and SRON. While both algorithms use POLDER measurements, different screening procedures, inversion methodologies and prior assumptions are made.

How the sizeable uncertainty in Line 207 is determined? Do you mean a rather low correlation means independent? Probably some references are needed for detailing the reliance and difference among the satellite datasets.

Yes, low correlation in AAOD suggests that these retrievals are essentially independent. Ours is the first paper that intercompares AAOD amongst datasets so no other references can be given.

Descriptions within Lines 220-225 is not enough to to understand the results in Figure 2.

We analyze how well the satellite data correlates with AERONET depending on the collocation criterion (the 'closeness in time' of the AERONET and satellite measurement). Obviously, a tighter criterion is better but also means less data to analyze. We show that results differ markedly between AOD and AAOD: for AOD, relaxing the temporal criterion has a mild impact, for AAOD there is a big impact. We interpret this as AAOD drifting in plumes over the AERONET sites (necessitating a tight temporal criterion) while AOD (due to its multiple sources) is spatially more distributed.

Figure 3, I didn't understand the two numbers (236 and 9083) here, could you please further explain? Is that the downward error bar of FL-MOC and OMAERUV have exceed the limit of -0.2? How the error bar is computed?

These numbers are not very important but they give an indication of the number of collocated satellite – AERONET data pairs used in the evaluation. The given numbers are averages over all 4 satellite datasets (because FL-MOC yields far fewer collocations than, say, POLDER-GRASP due to the narrowness of the CALIOP swath). Indeed error bars were cut off at the axis for esthetic reasons. Again, this is not very important: it's the top part of the error bar that matters as it shows that there is a statistically significant difference between these two evaluations (DirectSun and Inversion). The error bars were computed using bootstrapping and only represent statistical noise (not measurement uncertainty), see Sect. 3.2. We have added a reference to this section in the caption.

Line 729, do you mean there is a threshold for minimum number of observations, if yes, the threshold is ?

We do not use a threshold but the code allows us to specify a threshold. Sensitivity study of the impact of this threshold on AOD evaluation can be found in Schutgens et al. *ACP* 2020. A threshold does improve agreement between satellite and AERONET (presumably because AERONET becomes more representative of a larger region, although extra cloud masking could also be an explanation) but the impact is small.

Figure 7, Why the number of sites are different in three comparison?

The number is the number of collocated data pairs (satellite-AERONET). They differ for the satellites because of the different orbits and data treatment of each satellite/retrieval algorithm. The same number of AERONET sites is used.

Line 228, the resolution of FL-MOC (2 deg.) is not consistent with the abstract (all products are aggregated unto 1 deg.)

Actually, FL-MOC was also aggregated to 1 degree. However, for the purpose of evaluation with AERONET (only) the spatial collocation criterion was relaxed to 2 degrees. See Sect. 3, line 227-229.

Section 3.1, the specific figure number should be mentioned here, like Figure 5 or Figure 6, because there are two Taylor diagrams in the figure list.

Section 3.1 is a general description of the Taylor plot (very common in model evaluation) and can be used to interpret both Fig. 5 and 6. The only difference is that Fig 6 also includes uncertainty ranges (not a standard aspect of the Taylor plot). This is explained in the text and caption. We have added a few more words to the caption of Fig. 6 to clarify this.

Line 244, why this is unbiased?

There are two definitions of RMSE (one is sometimes denoted RMSD, D for difference):

$$\varepsilon = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x_i - y_i)^2} \text{ and } \varepsilon = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x_i - y_i - \bar{\bar{x}} + \bar{\bar{y}})^2}$$

Here ε is the RMSE of a dataset x_i vs a reference y_i . The upper bars in the second expression denote averages. The second expression is called the unbiased RMSE as the averaged difference between x_i and y_i was substracted. Unbiased RMSE is mathematically very similar to the standard deviation.

Line 254, how the bars are calculated should be added to the captions.

Ok. It was mentioned in the text several times but it is a good idea to add this to the captions.

Line 270-271, the sentence should be rewritten to "These hotspot identified by three products all cover these polluted regions like. . . "

Sentence changed to "Three products agree on AAOD hotspots in China and India, that are known polluted regions"

Line 271, what would be the possible reason for the exception?

The relevant text is: “The products all agree on a major AAOD hotspot from (likely) African Savannah biomass burning. Three products agree on known polluted regions like India and China also being AAOD hotspots (OMAERUV, which is relatively featureless, is the exception). “

The OMI sensor uses much larger pixels than e.g. POLDER. Signals from small-scale absorbing aerosol will be aggregated over those pixels prior to retrieval, leading to lower AAOD. In our evaluation with AERONET data we also see the impact of such small-scale structures: a strong decorrelation with increasing temporal collocation criterion (see Fig. 2). This is explained by plumes drifting over the site, that require tight temporal constraints for satellite evaluation. An explanation has now been added to the paper.

Line 280, please direct the readers where the discussion will be by giving the section number.

Good idea, added.

Line 286, than found-> than that found

Replaced by ‘those’

Line 300, what ‘several’ means specifically?

Either 1, 2 or 3 year(s), depending on product(s). See also Table 1. A reference to this Table has now been included in this sentence.

Line 301, the comparison should be made with caution in terms of what?

Due to temporal sampling issues (e.g. Colarco et al. 2014, Schutgens et al. 2016). The products not only are available for different years, but within each year also have different samplings (due to e.g. orbit or cloud masking). Extra explanation was added.