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.