Dear editor and reviewers,

We would like to thank you for your detailed comments and suggestions for improvements to our paper entitled 'Coherent uncertainty analysis of aerosol measurements from multiple satellite sensors'. We summarize below the changes in the revised version of the manuscript and explain how these changes address the specific concerns of the reviewers. In response to certain reviewers' comments that are related, we may reference a response provided elsewhere, if we consider that it has addressed the current issue comprehensively, rather than repeat the response multiple times, so as to avoid duplication.

In addition to answering the specific concerns of the reviewers, we have also updated certain data analyzed in the manuscript and made corresponding adjustments to the reported statistics. These updates can be briefly summarized as follows:

1) SeaWiFS AOD version 003 data has been replaced with the recently released version 004 data.

2) AERONET data has been updated based on the latest available AERONET archive, and 13 stations were added to our analysis as a result of this update.

3) POLDER data filtering has been updated to make it comparable with the POLDER filters in Bréon et al., 2011 (please see Q.1.1 below for more details).

4) While the original study compared OMI AOD at 388 to AERONET AOD (interpolated or extrapolated at 388 when not directly available), the revised study has updated this comparison to using AOD at 388 nm from OMI only when this wavelength band is also available from AERONET, but compares them at 500 nm otherwise (please see Q.2.14 below for more details).

Finally, in response to Q.2.16, we updated the outlier filtering procedure so that our outlier analysis at each individual AERONET station is now based on a median retrieval error at this specific station, in contrast to the original work that used a median retrieval error across all locations. As a result, in the revised version of the manuscript, the effect of removing outliers is less pronounced, particularly in Table 3 and Figure 10. Nevertheless, we believe this update was necessary to reflect the regional relative performance of the sensors more accurately.

Thank you,
Maksym Petrenko and Charles Ichoku

Anonymous Referee #1

Q.1.1. *Just a suggestion that the authors may consider. There was a paper published recently with a very similar objective and that could get cited: Bréon, FM, A. Vermeulen, J. Descloitres, 2011: An evaluation of satellite aerosol products against sunphotometer measurements. Rem. Sens. Env, 115, 3102–3111.*

-- Thank you for the useful reference, we have discussed its relevance to the presented work in the revised manuscript as follows:

'Finally, a recent study compared AERONET retrievals with a set of 5 spaceborne aerosol products archived at the ICARE Data and Service Centre, including POLDER, MODIS-Aqua (Dark Target retrievals), MERIS, SEVIRI, and CALIOP (Bréon et al., 2011). Although that study was based on a similar collocation framework as that used in the current study, our study focuses on a different set of sensors that provides a more extensive set of over-land spaceborne aerosol products. Furthermore, the presented study is based on the analysis of the spatio-temporally averaged and outlier-screened data, whereas that of Bréon et al. (2011) is predominantely based on the analysis of individually collocated spaceborne and ground-based data points that are the closest in space and time that would correspond to the central values in our collocated data subsets (we report a similar analysis in the Digital Supplement to this paper).'

-- Further, we have updated our POLDER data filters to make the results of our study more comparable to the results of the study in the reference:

'Since there are no formal recommendations on the acceptable range of these flag values, we have adopted thresholds suggested for the 'quality of inversion' flag in (Bréon et al., 2011), specifically 0.5 for land retrievals and 0.2 for ocean retrievals.'

Anonymous Referee #2

Summary of major comments

Q.2.1. *50km spatial average goes in the direction of level 3 (unclear about requirement for central values)*
--Please see our detailed answer to Q.2.13.

Q.2.2. *The generally successful outlier filtering cannot be applied to level 3 products*
--Please see our detailed answer to Q.2.6.

Q.2.3. *Data-volume as other (than accuracy) element on data-use is left out of recommendation*
--Please see our detailed answer to Q.2.12.

Q.2.4. *AVHRR data (offering links to the past over oceans) are not included in the inter-comparison*
--Please see our detailed answer to Q.2.8.

Q.2.5. *POLDER fine-mode AOD data should not be evaluated over dust dominated regions*
--Please see our detailed answer to Q.2.9.

General comments

Q.2.6. *Particular appealing are increased data-sets capabilities with upper end outlier data removed (while surrendering about 5 to 10% of the data). Hereby the chosen method linked to the local median seems a sensible approach. Unfortunately, such data-set improvement is not possible for externally aggregated level 3 data-sets, which begs the question, if in an additional step these improved level 2 products could be aggregated into improved level 3 products, as level 3 products will continue to be the primary evaluation choice in global modeling (due to their compactness and similarity in scale).*

-- Indeed, since the described test requires a reference data (i.e., AERONET), it cannot be directly applied to remove outliers in level 3 data, which are aggregated from level 2 retrievals and do not provide flexibility for comparison with AERONET. Nonetheless, we believe that by studying or otherwise analyzing the outliers and the corresponding retrieval conditions, as identified by the described statistical test based on the collocated data, we could possibly develop appropriate mitigation measures in the retrieval algorithms or design specific data screening strategies for each of the products. (We have added this explanation to the end of Section 5 of the revised manuscript).

Q.2.7. *Another issue, which did not get much attention, is the difference in data-sampling, mainly due to sensor swath capabilities. While data accuracy is desirable, accuracy without coverage is less interesting. Thus often lower accuracy is acceptable, if in turn better spatial and temporal context can be provided. Even though this analysis works with (multi-?) seasonal data, the co-locations and the volume offered by CALIPSO will be much lower than that of MISR and that of MISR in turn will be much lower than MODIS. Fewer samples mean fewer coincidences with AERONET references, so statistics based on a different number of sites and a different number of samples at those sites are not (strictly) comparable (when deciding on the regional/seasonal best retrieval).*

--Data volume (sampling) aspect is indeed very important and is addressed in various parts of the paper. For example, columns 2-5 of Table 3 provide insight into the total volume of data available for each of the sensors, including the impact of seasonal changes in retrieval conditions. Fig. 11 and 12 use marker sizes to articulate the

available data volume of the best retrievals, and Section 7 explains that some anomalies in the IGBP-based statistics stem from the differences in the available number of collocated data points. To supplement this information, we have made a number of changes to the revised version of the manuscript, including:

* Section 3. Specified swath width parameter of each sensor and discussed the limited swath of CALIOP.

* Section 6. Added to a discussion of data volume columns in Table 3:

'The second column of Table 3 (Nfilt) outlines the total volume of the collocated quality-filtered data available for each of the sensors depending on the boreal season. Although sensor swath width (Section 3) and data quality (Section 4) properties are among the main factors that determine the available volume of data (e.g., MODIS has approximately 4 times the swath width and 4 times the data volume of MISR), it can be seen that the seasonal changes in retrieval conditions also have a very considerable impact on the data, where Summer retrievals can have 2-4 times as much collocated data points as Winter retrievals. The relative data volumes of the studied data products provide an important context to be carefully considered when interpreting the statistics discussed in the remainder of this paper.'

*Section 6. Discussed data volume aspect represented by marker sizes in Fig. 11 and Fig. 12.

'Additionally, as indicated by the smaller relative sizes of certain markers in Fig. 11 and Fig. 12, although some locations might be covered by highly accurate spaceborne retrievals from certain sensors, if such sensors offer limited coverage and data availability, their accuracy advantage may ultimately produce only limited impact, highlighting the auxiliary but still important role of the less precise but more extensive products.'

Q.2.8. *There is some disappointment that AVHRR data-sets are not included in this inter-comparison. Despite their algorithm simplicity AVHRR offers competitive data over oceans and a link back in time.*

-- Since our study uses ground-based AERONET data, which are mostly available over land, as the reference dataset, our prime focus was on spaceborne products that retrieve AOD over both land and ocean. Although there are several provisional land AOD products that are based on AVHRR data, the main standard AVHRR AOD product is still over ocean, and therefore is better evaluated on the basis of maritime sunphotometer measurements, which are not included in this study. Nevertheless, we are considering adding AVHRR product to our MAPSS framework so it can be included in future studies.

Q.2.9. *Another issue is the use of POLDER fine-mode AOD data which should not be compared to the total AOD (except over urban and wildfire seasons, where fine-mode AOD contributions dominate. Thus, the large biases over regions affected by dust and the overall low (-est) correlation coefficients put POLDER in a rather poor light.*

-- While we generally agree and consistently reiterate this aspect of the POLDER land data throughout the paper, it is important to remember that POLDER Land AOD is distributed as 'AOD corresponding to the polarized particles (mainly anthropogenic aerosols)', e.g., see http://www.icare.univ-lille1.fr/parasol/?rubrique=aero_list . Moreover, POLDER does not make any specific recommendations on the regions suitable for using its AOD retrievals as a proxy for the total AOD, nor for using POLDER AOD as a pure fine-fraction AOD. Since these factors might potentially leave the users of the POLDER AOD confused about the applicability of the data, we feel that our study provides certain important insights on the characteristics of this dataset as compared to other available AOD products, as well as outlines such geographical regions where POLDER AOD can be treated as a total AOD (e.g., see Fig 14). The revised manuscript clarifies this special role of the POLDER data in our study in Section 3.

Q.2.10. *In that context also there are little to no discussion on CALIPSO data and its poor correlation (there may too few CALIPSO data to perform a confident evaluation). This brings me to the question of bias (as CALISPO data tend to be biased low). Any evaluator wants to know first if the (satellite) reference is biased – in what direction and by how much, as function of region, season and AOD.*

-- There is a known issue in the CALIOP retrieval algorithm that leads to underestimation of AOD, as reported in 'An accuracy assessment of the CALIOP/CALIPSO version 2/version 3 daytime aerosol extinction product based on a detailed multi-sensor, multi-platform case study', Kacenelenbogen et al., 2011. We have added this reference to the revised version of the manuscript.

Q.2.11. *In this contribution there is only limited information given. I do not like the generality of linear fits and the scatter plots often remain discouraging, even with outliers removed. In addition, there is almost no info on biases for low AOD (0-0.2) or for median AOD (0.2-0.5), as scatter plots are offered only for the large 0-5 AOD range.*

-- We have chosen the same scale for all of our plots so that the plots from different sensors can be readily compared to each other. Another consideration is that without a ground reference, it is virtually impossible to detect an underestimation of high-AOD events in the original Level 2 data, and so it is important to explore how each product behaves across the complete range of AOD, including the extreme values of up to 5.

In the revised manuscript, we have updated scatter plots to include 0-0.5 AOD insets. Also, since we could not fit all AOD range specific data into the body of the main

manuscript, we have updated the Digital Supplement to tabulate the statistics based on the suggested ranges of AOD.

Q.2.12. *(continued from Q.2.11) It [study] confirms the general sense of complexity and limitations by satellite remote sensing of aerosol properties, but recommendations remain vague also somewhat ignoring the data volume aspect, which is also an element for a decision on data use.*

--Please see our answer to Q.2.7.

Minor comments

Q.2.12. *4642/20 using the 50km mean value may help in the comparison among different sensor products but goes in the direction of level 3 comparisons. We really learn about the satellite products more from comparisons of the central value to the ground reference data.* ***Q.2.13.*** *It also remains unclear, if the central value was a requirement in matched to AERONET. If comparison involve satellite data without its central value, then the 50km evaluation is less meaningful. Having a central value also would give insights into the central value's regional representation (and hereby helping to address a site's regional representation)*

-- We performed a similar analysis based on both mean and central values; still, the manuscript reports results based on the mean values, while the results of the central value-based analysis are provided in the digital supplement to the paper. We choose this arrangement to ensure the clarity of our presentation as well as to keep the paper's length within reasonable limits. Furthermore, we added a discussion of pros- and cons- of both approaches in the revised version of the manuscript and clarified the requirements of collocation:

'In this paper, results are reported based on the analysis of the mean values. Although not reported in this paper because of the space considerations, a similar analysis was performed based on the central values and is reported in the digital supplement to this paper. It is appropriate to use the mean values in this paper, so as to maintain the uniform sampling criterion across the different sensors and their respective retrieval pixel sizes to facilitate a fair intercomparison. On the other hand, an analysis based on the central pixel values such as that reported in the digital supplement can provide further details on the effect of difference in sampling aerosol products from individual sensors, as well as more accurately characterize the performance of the sensors in the presence of a strong point source of pollutant particles. Additionally, it should be noted that since the mean value of a sample can be computed even if its central value is missing, the reported analysis of the central values is based on a somewhat reduced volume of the collocated data points when compared to the reported analysis based on the mean values.'

Q.2.14. *4643/8 the AOD (via Angstrom) interpolation (and especially the extrapolation into the UV) is only sensible for AERONET AOD data, but not for satellite retrieved AOD data (with a-priori absorption assumptions). This complicates any combination of different satellite data products (e.g OMI in the context of MISR or MODIS)*

-- We have updated our approach to evaluating OMI so as to avoid extrapolation into the UV and further discussed the issue of the wavelength dependence of AOD in the revised version of the manuscript as follows:

'It is pertinent to note that this interpolation process might introduce an additional source of uncertainty when intercomparing the aerosol products. Also, because of the wavelength dependence of AOD, the difference in the compared wavelengths of the spaceborne products should be considered when intercomparing the relative performance of the products. Furthermore, although many AERONET stations provide observations in the range of 340nm-1020nm, certain stations report AOD in the range of 440nm-1200nm. For such stations that have no measurements in the UV region, we have evaluated OMI AOD at 500nm instead of AOD at 388nm, in order to avoid additional extrapolation biases.'

We have also updated Table 1 to reflect this change in the evaluation of OMI.

Q.2.15. *4649/19 the outlier detection if simply based on the retrieval is interesting and important (and calls for the developing of associated level3 prodcuts). However, here (if not, state so) the outliers are based on regression line deviation, thus seem to involve a reference data. If this reference data is AERONET, then it will be really difficult to create an outlier removed global data-set.*

--Please see our answer to Q.2.6.

Q.2.16. *4650/3 the five times above median values only finds the upper-end outliers*

-- While the test at a particular AERONET station indeed identifies as outliers such data points that grossly overestimate or underestimate AOD relative to AERONET measurements, please also keep in mind that the cut-off threshold is defined based on the median relative error at this specific station. Therefore, depending on the station, this threshold can be set relatively low. We added a discussion of this issue in the revised version of the manuscript:

'When applied to the collocated AOD data, this test removes those spaceborne retrievals at a particular AERONET location that grossly overestimate or underestimate ground-based observations as compared to the median retrieval error at this specific location. This is especially useful since many spaceborne retrieval algorithms tend to either under-estimate certain high-AOD events because of a pre-set maximum AOD

threshold, or over-estimate AOD in the presence of clouds as well as under very low-AOD conditions. However, it should be noted that the test may not remove such possible outliers that have a relatively small error.'

Q.2.17. *Table 2 the applied quality criteria are not quite clear for cases where more than one QA criterion is listed*

-- We clarified this issue in the revised manuscript as follows:

'For CALIOP, a column is accepted only if all layers found in this column meet all listed QA conditions.'

Q.2.18. *Table 3 almost all slopes are below 1.0 … this is surprising to me as on an event basis I would have expected the opposite. This is probably related to some satellite retrievals inability to catch high AOD events, also since almost all intercepts are positive! Also considering regional differences the presentation of comparisons via linear fit lines is somewhat misleading. Also some slopes of POLDER are very low apparently related to the use of fine-mode AOD POLDER data. I would leave then POLDER data out of the table or would only compare to the AERONET fine-mode AOD. Also the low slopes of CALIPSO need some explanations.*

-- We have addressed this issue in the revised manuscript as follows:

'For most of the sensors in Table 3 and Fig. 5, the slope of the fitted regression line is below 1.0 and the intercept is slightly above 0. This can be explained by the limitations of the spaceborne retrieval algorithms that tend to 1) overestimate low-AOD events when the AOD signal is very weak and almost indiscernible from the surface signal, resulting in a portion of the surface signal being mistaken for an AOD signal; 2) underestimate high-AOD events because of the very weak surface signal, where a portion of the AOD signal might be mistaken for a surface signal. Furthermore, most algorithms have a pre-set limit on the highest possible retrieved value of AOD (e.g., 3.0 in MISR), which may further affect the reported statistics. Finally, certain censors have peculiar features that impact the overall characteristics of their data. Among such features are: sensitivity to sub-pixel cloud contamination in OMI retrievals that leads to an overestimation of AOD (Torres et al., 1998), sensitivity to fine particles in POLDER land retrievals that lead to underestimation of AOD in coarse mode dominated regions (Herman et al., 1997), and also frequent under-estimation of AOD by daytime CALIOP retrievals (Kacenelenbogen et al., 2011). Since without a ground reference it is near impossible to recognize an underestimation of high-AOD events or over-estimation of low-AOD events in the original Level 2 spaceborne data, it is especially important to explore the behavior of each product across the complete range of AOD values.'

Q.2.19. *Table 4 please separate between spatial and temporal correlations, and hereby differentiate between seasonal and inter-annual correlation, if possible (also try rank correlations as a few outliers can dominate the result). Also the correlations do not address the bias.*

-- For volume considerations, we have added seasonal versions of Table 4 and 5 into the Digital Supplement, while bias is addressed in the new Table 6 and Figure 15 (see Q.2.20).

Q.2.20. *Table 5 RMS is a mixture of bias error and variability error. Such an error distinction could be insightful.*

- We have added Tables 6 and 7 that show the bias and variability components of RMSE in Table 5. We have also added Figures 15 and 16 that show bias and variability for Figure 14.

Q.2.21. *Figure 4 I am not clear about the colors: to me it looks as if it indicates an event frequency. The 'red' linear fit lines are too meaningful (as Table 3) and even surprising for OMI. I also suggest different plots for OMI and MODIS (also as AOD values at shorter OMI wavelengths are generally larger) and would use the extra space to show in the lower three panels a scatter plots enlargement only for the 0.0 to 0.5 AOD regions.*

-- We have updated density plots in Figure 4 and 6 with 0.0-0.5 AOD region insets and an alternative density function that bins data into 0.1 bins (0.05 bins for insets) and colors each bin based on the percentage of all data points that fall into this bin.

Q.2.22. *Figure 5 This plot is rather small and loaded with info. Give am explanation why some errors for particular days are only one-directional and otherwise symmetric?*

-- We clarify this in the revised manuscript as follows:
'In this [bottom] figure, the magenta line indicates daily means of AERONET AOD at 440nm, bar heights reflect the number of all-QA (top half of the figure) and best-QA (bottom half) data pixels in each spaceborne sample, and error bars represent the mean relative accuracy of each sample computed based on its pixels. As an example, consider AMODIS DB (turquoise) retrieval on Dec. 13. Even though the mean AOD based on 22 pixels in this retrieval was within 10% of the corresponding AERONET AOD, all these 22 pixels were marked as having a bad QA.'

Q.2.23. *Figure 6/7/8 … too small to detect detail*

-- Part of the reason for this is that ACPD/AMTD publications have a horizontal layout. The legibility of the figures will improve appreciably when ACP/AMT ultimately publishes the article with a vertical layout, and we plan to work with the typesetter to ensure that the provided figures are enlarged or otherwise modified to a suitable size.

Furthermore, the electronic PDF version of the submitted paper provides high-resolution figures that allow for zooming into a desirable scale.

Q.2.24. *Figure 9 POLDER outliers (no surprise) are related to areas where corase mode aerosol dominates … it also might be nice to indicate (possibly by different symbol shapes) if outliers are high or low with respect to AERONET (to demonstrate potential aerosol type biases)*

--Please see our discussion of POLDER data in Q.2.9. We have also updated Figure 9 to indicate the prevailing directions of the biases.

Q.2.25. *Figure 10 the impact from the removal of outliers is convincing. However, if is it based on comparisons to AERONET then this extra filter is of less use, as then an outlier removal in non-AERONET country will be difficult to achieve.*

-- Please see our answer to Q.2.15.

Q.2.26. *Figure 11 I assume these are temporal correlations at each site (still there may be biases to considered).*

-- This is correct. We have updated captions of Figure 11 and Figure 12 as the following:

'Spaceborne datasets with the best correlation (R^2) of the retrieved AOD to the AOD measured by each individual inland (top) and coastal or island-based (bottom) AERONET site.'

Furthermore, while listing and discussing biases of the individual sites in the context of each individual spaceborne sensor would require a prohibitive amount of space, some of these biases were explored in the context of individual sensors (e.g., in 'Quantitative evaluation and intercomparison of morning and afternoon Moderate Resolution Imaging Spectroradiometer (MODIS) aerosol measurements from Terra and Aqua', Ichoku et. al, JGR'05) or individual sites. Additionally, we are also working on an interactive tool that would allow data users to explore the statistics of individual sites and sensors in more detail.