Reviewer 1:

Suggestions for revision or reasons for rejection (will be published if the paper is accepted for final publication)

I reviewed the previous version of this paper. My main issue with the previous submission was that there were too many comparisons and similar unclear acronyms being used, which made it hard to find a coherent message. In this version the authors have reworked and streamlined the paper; I appreciate these efforts and think it makes some aspects easier to follow. However, I think some more work is needed. Some figures are still hard to understand (lots of nearly-overlapping similar dots, large numbers of panels which also tend to look similar). Some of the text is still unclear, out of order, or not needed (e.g. on line 166 there are 8 references for the simple statement that the AHI retrieval uses the Cox-Munk ocean surface reflectance; none of these references are to the Cox and Munk papers describing the actual equations). So it is still difficult to me to pick out a main message other than merging seems better than not merging. I think the relative merits of fusion over just doing the simple bias correction step should also be explored. From the figures, in general, it is difficult to pick out the key points and main message.

I therefore recommend some more revisions for clarity. I would be happy to review the revised version. I do not want to discourage the authors: this will be a good paper for the journal, and as I said in my previous review it is important that aerosol products from these geostationary sensors are given more attention. It is just not quite there yet, in terms of text and graphics.

→ Thank you for your valuable comments. We tried to reflect the comments by the reviewer, which improved the readability and contents of the manuscript.

Line 170: the use of 0.02 mg/m3 Chl still does not seem reasonable to me. I see the text has been changed from the previous submission, but “less contaminated” does not make sense.

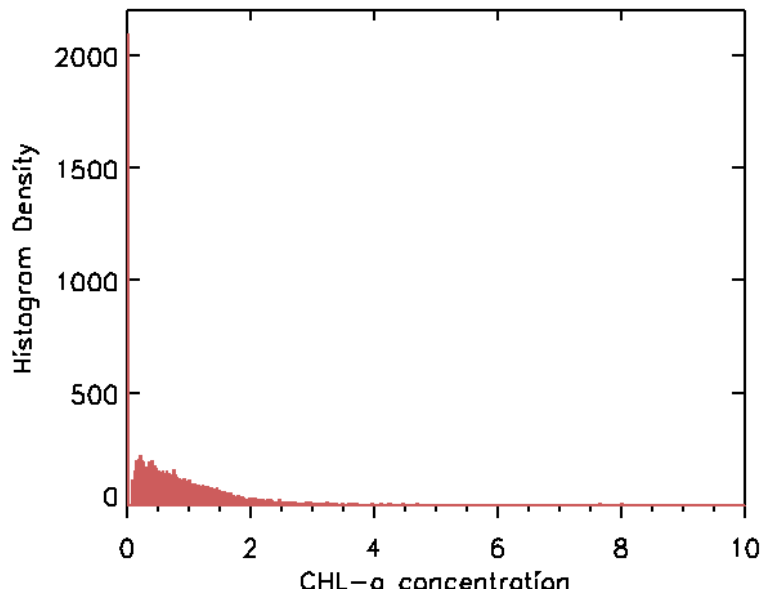


Figure 1. Frequency distribution of JAXA CHL-a concentration in [mg m-3] for 1 May 0400UTC.

→ As JAXA CHL-a are retrieved for limited pixels with its own cloud masking, thus aerosol retrieval requires assumed CHL-a values for the remaining pixels without CHL-a products. Cloud masking is different for JAXA CHL-a and the AHI aerosol algorithm. Thus, 0.02 mg/m^3 was assumed for the pixels without CHL-a data retrieved, to maintain the retrieval area of the AES product, which increased retrieval points by up to 2000 (as shown in the upper Figure). In the figure 1 of CHL-a frequency distribution, CHL-a concentration of 0.02 mg/m^3 corresponds to a very small value in the frequency distribution. Thus it was described as 'less contaminated' in the viewpoint of aerosol retrieval. Table 1 shows the result of AOD calculated with different CHL-a values. With the original input of 0.913 mg/m^3 CHL-a, AOD was retrieved to be 0.223. When 0.02 mg/m^3 CHL-a was assumed as in this study, AOD retrieved was 0.225, which is very similar. AOD retrieved with 50.0 mg/m^3 CHL-a was calculated as 0.222. Thus, the maximum difference in AODs due to the difference in CHL-a concentration is as small as 0.003.

Table 1 Sensitivity test for CHL-concentration of AES products under the following condition (SZA = 19.43 deg, VZA= 41.45 deg, RAA= 133.1 deg, and wind speed = 5.41 m/s).

CHL-a concentration (mg/m^3)	**Retrieved AOD at 550nm of AES algorithm**
(original) 0.913 mg/m^3	0.223
(test) 0.02 mg/m^3	0.225
(test) 50.0 mg/m^3	0.222

According to Werdell and Bailey 2005, the in-situ CHL-a concentration using the High Performance Liquid Chromatography (HPLC) technique is 0.021-48.99 mg/m^3. Also, the NASA bio-Optical Marine Algorithm Data set's (NOMAD) minimum value for comparison with AHI CHL-a in Murakami 2016 paper appears around 0.02mg/m^3.

ref)

Werdell, P. J., & Bailey, S. W. (2005). An improved in-situ bio-optical data set for ocean color algorithm development and satellite data product validation. *Remote sensing of environment*, *98*(1), 122-140.

Murakami, H. (2016, May). Ocean color estimation by Himawari-8/AHI. In *Remote Sensing of the Oceans and Inland Waters: Techniques, Applications, and Challenges* (Vol. 9878, p. 987810). International Society for Optics and Photonics.

Lines 172-190: most of this text is a discussion of the relative merits of the MRM and ESR techniques. As such it probably belongs in the introduction, where these techniques are discussed, rather than in the algorithm description section.

→ Thank you for your comment; we moved these sentences to the introduction in lines 107-128 of the revised manuscript.

Lines 207-208: “the accuracy of GOCI, according to NDVI, has a negative bias for V1 and mostly a positive bias for V2” I don’t understand what “according to NDVI” means here. Can this be reworded?

→ Thank you for your comments. This sentence was revised in lines 253-254 and as follows:

~~ the accuracy of GOCI AODs was dependent on the NDVI values. V1 has ~~

Line 263: The Sayer (2013) paper here in fact says the opposite of what the authors cite it for: it says that dAOD is not Gaussian, at least for Deep Blue data. See Figure 5(b) of that work. That is why the Deep Blue team have to define the expected error envelope and normalize to make it Gaussian. I think the authors are doing something similar here (in Figure 2) but the wording of the text implies the opposite.

→ Sorry for confusion. We removed the reference (line 335) and added Q-Q plot with the description in the manuscript in lines 344-349.

Figure 2: it would be good to add a line showing the theoretical QQ plot for a Gaussian distribution, to make it easier to show how close the data sets are. Also, it is not clear from the text/caption exactly what is plotted here. The text says "dAOD" but the caption says "z score". So is it dAOD divided by RMSE? This should be fixed.

→ Sorry for confusion. We added standard normal distribution (mean=0, std.=1) in the Q-Q plot. . The y-axis is normalized to std, and the text was revised. At both ends of the Q-Q plot, the sample quantile is more skewed than the theoretical value. However, it shows a symmetrical shape with respect to the point of 0.5 on the x-axis, so it still follows the Gaussian distribution.

Lines 275-282: I think this text is saying that, for each NDVI bin and hour of day bin (Figure 1), the mean bias is calculated and subtracted from the retrievals. Is that right? If so, can this paragraph be streamlined? If that is not right, can what is done be written more clearly?

→ Thank you for your advice. We revised these sentences in lines 354-356 of revised manuscript.

.

Section 4: this first part (page 9) uses the word "overestimate" a lot. However what the authors seem to mean is "this satellite combination is higher than that satellite combination". There is no discussion of the AERONET ground truth here so it is impossible to say whether one (or both) things being compared is overestimating or underestimating. It would be better to say that the two are "offset" relative to each other, as that only implies a difference, while "overestimate" implies an error.

→ Thank you for your comments.

We modified the word 'overestimate' to 'offset'. And we added validation result of respective satellite product with AERONET AOD summarized in table 2 (in Section 3.5) before Section 4(results section).

Figures 4, 5: these have 11 panels. Do we really need all these comparisons? Is there some better way to convey the intended message from these comparisons? It is hard to know exactly what details I am supposed to focus on here. I feel the figure is mixing both the comparisons of individual products with the results from individual merges. Maybe it would be best just to show campaign averages of the 4 baseline satellite products here so we can see the differences in them. Or show panels a-e (i.e. FM1 and the satellite products) in one figure and compare the other merges somehow. I am sorry I know it is difficult to present a large amount of information like this, and I am also not sure of the best way.

→ Thank you for your advice. Figures 4 and 5 were replaced with the difference of respective satellite product and average FM1 AOD to contrast the characteristics of each product. FM1 AOD is considered to be the representative fusion product as it includes.

Figure 6: this is interesting and to me shows that the MLE method helps at this site in low AODs. However it is hard to see some details because there are lots of near-overlapping dots from the measurements every 1 hour or so. And lots of big gaps from night time. So perhaps this plot could have the points plotted at daily scale instead? Or else add an additional figure which focuses on a short period (maybe a day or a few) so the x-axis is zoomed more effectively, and we can see how the products resolve the diurnal variability?

→ Thank you for your advice. The figure was revised as Fig 9 to the daily mean AOD during the KORUS- AQ period and corrected to show the diurnal variation of AOD from 11 May to 14 May.

Lines 345-357: I agree that the fused (especially MLE) products are better here. But it is not clear to me how much of this difference is due to the fusing, and how much is due to the bias correction step. Really we have two separate stages here: going from (1) satellite products to bias-corrected satellite products, and (2) from bias-corrected satellite products to the fused bias-corrected products. Unless I have misunderstood what is done here. There is obviously value in doing a bias correction (step 1), but the next step is less clear. Can the authors somehow separate this in the analysis?

→ The bottom figure shows the MLE results validation with the bias correction at the top and without applying bias correction at the bottom. Overall, RMSE, MAE, etc., are very similar. However, FM1 and FM2 of MBE slightly improve, but FM3 plays a role in worsening. However, the bias correction result for both %EE and %GCOS shows better accuracy in low AOD is improved.

Furthermore, the description of the bias correction effect was added to line 654-675 using FM3 and F3.

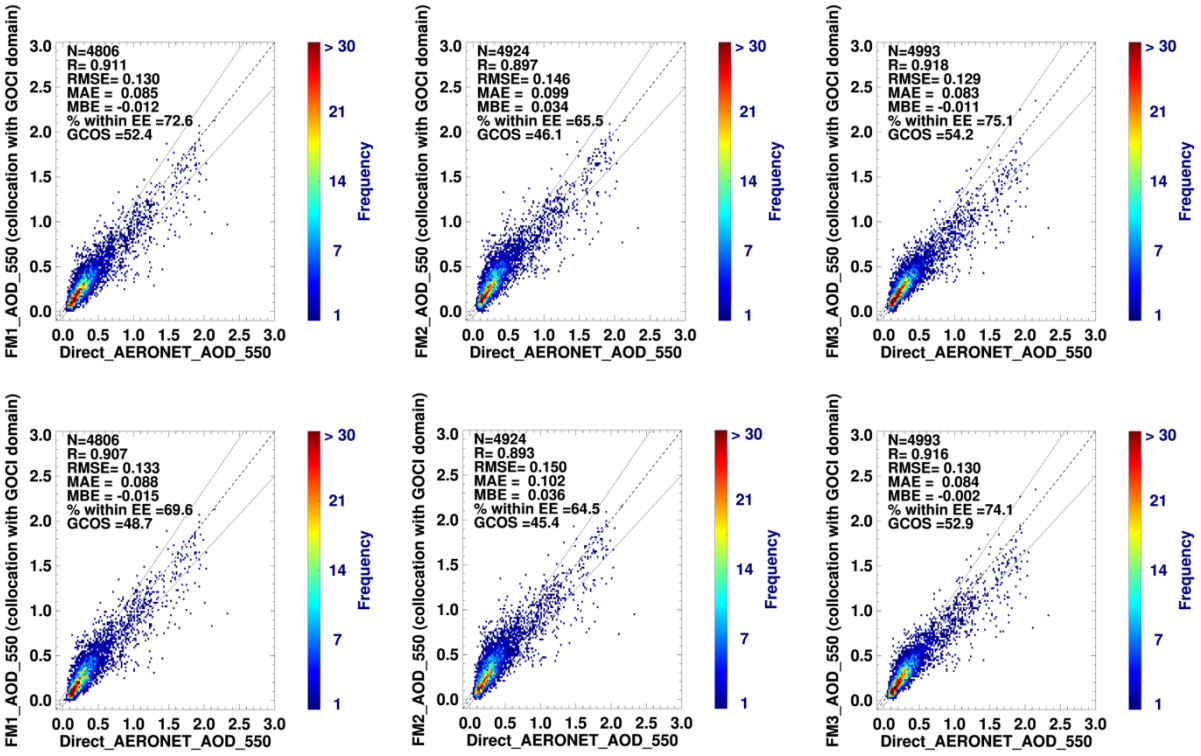


Figure 2. Validation of FM1-3 products with AERONET AOD during KORUS-AQ campaign (Top : Bias correction, bottom: no bias correction)

Lines 373-374: why compare to the MODIS DT EE when this is the expected uncertainty for a different algorithm and a different sensor? If the authors just want to provide a common reference uncertainty to benchmark against, ok, but then the paper has to be clear that this is not an "expected" error for any of the data sets (or merge) used here so is something of an arbitrary reference. Perhaps the GCOS goal (greater of 0.03 or 10% of AOD) would be a better comparison point since that is an international target not tied to one algorithm and sensor.

→ EE was used to compare common reference uncertainty to benchmark against MODIS as commented by the reviewer, and GCOS fraction was also added in lines 385-389. Also, we have specified 'EE' with respect to AERONET as EE for MODIS DT.

Section 5: it is a little strange to have the AERONET data and matchups described here when they were first used in Section 3.4 for the bias correction step. Some of this material should probably be moved earlier in the paper.

→ Thank you for your comments. We moved this part to the end of Section 3.3, because AERONET was also used for calculating RMSE.

Figures 7, 9: again, with 11 panels, it is hard to know what I should be looking for here. What is the main message of these figures? Are they necessary? Maybe it would be better if it were rearranged, with all "raw" satellite products on one row, all "averaged fused" on the middle, all "MLE" on the bottom? Or maybe it could be replaced with some plots of overall dAOD (combined to a smaller number of panels) and we hopefully see that the distribution of dAOD is narrower for the MLE results than the others? I am not sure but think somehow this should be streamlined. I am not sure we need to see 11 (mostly similar) scatter plots.

→ Thank you for your advice. We added a Table 2 and Table 3 to summarize the results and replace these figures.

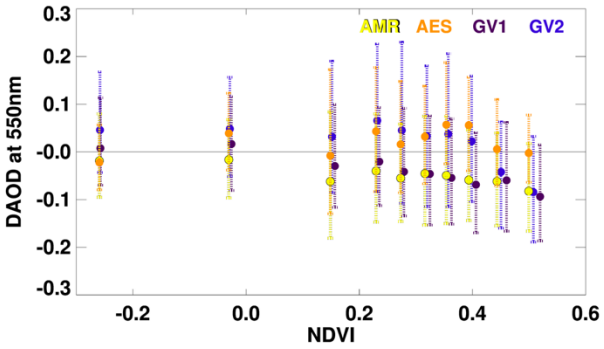
Figures 8, 10: same comment about 11 panels and hard to know what is the main message I should be extracting here. I think the authors need to think about what they are trying to show here. If it is that MLE has a higher fraction in EE than others, then make a plot showing that more directly. There are a lot of colored dots and a lot of white space and nowhere to easily direct the reader to which panel(s) they should be comparing to tell whether something is better or worse.

→ The figure was revised as Fig. 6 the GCOS fraction of respective satellite product for each AERONET site, F1, and FM1 AODs.

Figures 11, 12: why is figure 11 shown as dots and horizontal envelope lines while 12 is dots with vertically drawn bars? I get that the lines and bars are conveying the same information but it would be better to present these two consistently (especially as the figure 12 caption says "as Figure 11"). I personally find the style of Figure 12 easier to see and get the main message (the bias corrected merged products are flatter with smaller errors). However it would be good to pick different colors for the top panels (raw data) compared to the other rows. This is because the eye will naturally compare the same color in each row but that is not relevant here because the colors of the satellite products in the top are not directly relating to the merges in the other rows. It makes sense to match colors between rows two and three because F1-F3 are equivalent to FM1-FM3, just not with the top row.

→ Thank you for your comment. We tried to add an error bar as requested by the reviewer, but what maintained the existing solid line due to poor visibility. Attached the picture indicated by the error bar below. It also improved

visibility by displaying colors consistently.



Section 6: if I understand correctly the authors recommend MLE over simple mean as the merging technique. This should probably be in the Abstract.

→ Thank you for your comments we added sentences in abstract (lines 32-34).

Section 6: In the Conclusions I would also appreciate some simpler, higher-level statements rather than repeating numbers about EE, etc mentioned a few pages earlier in the paper (no need to write them about twice). For example, about what the differences between the simple mean fusion experiments F1-F4 tell us, and between the individual MLE experiments FM1-FM3 tell us. For example the fact that FM1 is better than FM2 but F1 is worse than F2 is interesting. Comparing F1 with F2 (or FM1 with FM2) should tell us something about the quality of the NRT vs non-NRT data so it is interesting that the opposite results are obtained between simple mean and MLE approaches.

→ Thank you for your advice. We added these sentences in Section 6 (lines 753-765).

Reviewer 2:

General Comments

This revision addresses many of the previous comments but the manuscript would still benefit from further distinguishing between results within the GOCI domain and the full AHI domain, simplifying several of the figures to improve legibility, making use of additional tables to summarize statistics rather than relying on reading very small print in figures, and drawing additional conclusions. Specific suggestions are offered as follows.

→ Thank you for your valuable comments. We tried to reflect all comments by the reviewer, which improved the readability and contents of the manuscript.

Specific Comments

The addition of Table 2 greatly helps in distinguishing the sensitivity to the domain size, and a similar approach should be used throughout the manuscript to ensure apples-to-apples comparisons are being made. For example, as currently written, the analysis associated with figures 7 and 9 essentially says to ignore F3 and FM3 because they are for the larger domain. But instead, the authors could do the analysis in two groups: one in which data only within the GOCI domain are presented for all cases, and one which compares the large and small domain results for AMR, AES, F3, FM3 (which is already done now, in section 5.3 and Table 2). Consistent analysis of all results within the GOCI domain should allow the conclusions to be more easily shown. The effect of domain is a separate important facet, and again I really like Table 2 and its discussion. Related, please clarify whether the results and analysis related to Figures 11 and 12 is conducted only within the GOCI domain or if it is mixing results with GOCI and AHI domains. I strongly feel it should be within the GOCI domain.

→ Thank you for your suggestion. We revised all analyses to collocate the retrieved products for the GOCI domain and added the analysis for the AHI domain (excluding GOCI area) in section 5.3.

The postage-stamp graphics and small text in figures 7 and 9 make them nearly illegible. These figures could easily be replaced with simple tables that summarize all the statistics for all the cases (similar to Table 2) within the GOCI domain. I think the key points would clearly emerge, at least for the KORUS-AQ period: within the GOCI domain, all individual satellite products have similar statistics, the ensembles improve the statistics, and the ensemble-mean and MLE techniques appear to produce very similar results. With these new tables, I feel figures 7 and 9 could be omitted.

→ Thank you for your suggestion. We added a new table (Table 2 in revised manuscript) that has replaced these figures.

Though if the authors feel there is some value in the figures, a representative subset could be shown, similar to how figures 4 and 5 have been recast to show AOD for only one example. I would suggest no more than 4 panels, to keep the figures legible (e.g., an AHI example, a GOCI example, F1, FM1). With these new tables, I also feel that Figure 13 can be omitted, as the tables would present the same information in a much more compact and easy to read form. Similarly, figures 8 and 10 could be simplified by only showing representative examples. If the authors prefer to show all cases, I suggest that the 11-panel figures show only the GOCI domain in panels a, b, g,

and k, and that a separate figure be used to show the AHI domain (perhaps with an outline of the GOCI domain) to elaborate on the low values in the extended domain.

→ Thank you for your suggestion. The figure was revised by the GCOS fraction for respective satellite product, F1 and FM1 AODs at each AERONET site. We have added a statement to lines 505-513 along with a Figure 6 to mention improvement of the fused products' accuracy.

I feel that Figure 6 and its discussion are out of place. I think it would make more sense for this analysis to come after the error analysis sections, because the estimated uncertainties of the geostationary satellite products have not been provided in the paper. (Maybe this uncertainty information belongs in Section 2.) It is relevant here; it is not clear how meaningful it is to consider satellite AOD values in the 0.01 to 0.1 range. Also on figure 6, the solid lines (left axis, dAOD) are not useful because the axis range is so large. dAOD rarely exceeds plus/minus 0.5 but the axis range is from -4 to 1. You probably need a separate plot, with appropriate axis range, to meaningfully show dAOD differences. Otherwise, the following statements about reduction of errors are not clearly demonstrated by the figure.

→ Thank you for your advice. The figure was revised to the daily mean AOD during KORUS-AQ period and corrected to show the diurnal variation of hourly AOD (removed dAOD) for May 11-14. Also the figure 9 and discussion was moved after the error analysis section. The importance of low AOD is also mentioned earlier in Section 5.3.

Section 6 should state high-level conclusions regarding the Remarks in Table 1. What can be said about the NRT vs all-available products, about the wider area of AHI, about the missing effect of

→ Thank you for your comments. We added to discussion of Remarks (in Table 1) in Section 6 (lines 754-771).

Technical Corrections

Line 168: Can you specify, what kind of data from JAXA?

→ Thank you for your comment. We added the data type and URL.

Near line 232: Either here or in the Table 1 caption, add a statement that the 4 entries F1-F4 denote the ensemble-mean fusion technique and the 3 entries FM1-FM3 denote the MLE fusion technique.

→ Thank you for your comment. We added each entry information into the Table 1 caption.

Near line 256: Need to comment on the scatter apparent in Fig 1 for the squares (AOD>0.5). It is especially obvious at 2 and 5 UTC for NDVI>0.5, but also apparent as inconsistent temporal patterns between the NDVI bins. Also, there is a typo in the red label of Fig 1, it should be 0.5<NDVI.

→ Sorry for confusion. We revised Figure 1, also we added comments on the scattered results in lines 329-331.

Line 270 and following: Need to state the purpose of Figure 2, I don't understand what its significance is. Is it just a graphical illustration of how the bias is determined, or a means of assessing the degree to which the distributions are Gaussian? Also, the text isn't clear. The curves don't appear linear between plus/minus 1 to me. Though, what is the significance if they are not linear? Also, the multiple statements of excluding data beyond 2 SD seem repetitive.

→ Sorry for confusion. We addressed these issues in section 3.4. The bias was corrected using the Gaussian center value evaluated against AERONET for the evaluation period (Apr. 2018 to Mar. 2019 excluding EMeRGe period), which were applied for the KORUS-AQ and EMeRGe campaigns.

Discussion of Figure 4: An additional statement should be made about Fig 4. The very small values shown everywhere in 4(f) show there is really very little difference between the two ensemble methods. There is only a fairly uniform small offset (bias) apparent.

→ Thank you for your comments. Mean F1 and Mean FM1 appear similar because the RMSEs of AMR, AES, GV1, and GV2 used in this study are mostly similar to 0.1, and the weighting function used for MLE fusion becomes similar. However, Figure 4 has been revised, and Figure 4(f-k) has been removed.

Line 443: I think the word “smaller” should be “larger”

→ Thank you for your comment. However we removed Figure 13.