Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2018-1130-RC3, 2018

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Inconsistency in spatial distributions and temporal trends derived from nine operational global aerosol optical depth products" *by* J. Wei et al.

Anonymous Referee #2

Received and published: 24 December 2018

The authors take 9 satellite aerosol optical depth (AOD) monthly mean data sets, and perform comparisons against each other and AERONET monthly mean data. These come from a variety of satellite instruments and algorithms. They look at similarities in spatial and temporal patterns. This research area is important because understanding aerosol influences on the Earth requires understanding the strengths and limitations of each data set.

This is a pretty big task and it is good to see it being tackled, because as the authors note there has not been a great deal of attention to data set choice in some satellite analyses. However, I think this version of the paper has problems. The statistics and

analysis are very superficial, and the metrics used do not always make sense or are incorrect. For example, autocorrelation and false discovery rate are ignored, a level 2 error metric is used for level 3 analysis. The terminology has errors in some sections (e.g. "validation" when this is not a validation analysis). And in several places the authors omit relevant references and use out of date ones, or instead insert excessive self-citations. There is also a possible wavelength issue with the AVHRR product used. I recommend major revisions and would like to review the revised version. This paper felt to me like the authors just downloaded a bunch of data and ran a bunch of statistical metrics against it, without thinking about what was being done or why. I suggest that when revising, they focus on what science question they are trying to answer, and then figure out the right tools to answer it, and provide a detailed discussion. Otherwise this feels not like a scientific research paper but rather the output of some automated data processing software.

After writing this review, I read the other two comments currently posted on ACPD for this paper. I generally agree with the other reviewers' comments.

My comments in support of my recommendation are as follows:

Line 20, and elsewhere: Operational is not the right word here. It implies something produces as part of routine agency operations while a mission is ongoing. Most of the products do not fit that definition; in fact I think only MODIS and AVHRR do as they are produced with a few hours latency to support assimilation applications. I suggest deleting this word throughout.

Title, line 30, line 35 and elsewhere: Terms like "significant inconsistencies" (or just "inconsistencies" alone), "seriously" are used a lot in this paper. But most of the time they are used as "weasel words", i.e. in a non-specific way which can lead people to get a certain impression which is not necessarily warranted. For example, "inconsistencies". Taken to an extreme, any two data sets will not be identical so are to some extent "inconsistent". The relevant question is, for any particular application, is the level of con-

sistency between them sufficient? For example, if one wants to look at seasonal variations, AOD magnitude might not be as important as the pattern throughout the year. But if one wants to look at radiative effects, magnitude is more important. If one wants to see large-scale features, then a broader swath to improve sampling at the expense of some accuracy might be desirable. The point is that these are all different instruments with different characteristics. We expect them to not be identical. The wording in this paper (these examples and elsewhere) seems designed to send a message that aerosol remote sensing has big problems. In my opinion, that's an overly pessimistic assessment. There are differences but in general the reasons for those are understood. So which data set is best to use for a given study depends on the type of science question you are trying to answer. There is no "best" data set. This recent paper by Sayer et al in JGR (https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2018JD029465) covers some similar ground to the current study, in that part of it compares time series and maps of various over-water satellite AOD data sets. That paper goes into a lot of discussion about the differences between them and why they might be. So although there is a lot of diversity in the over-water AOD, the reasons are generally known. My personal opinion is that over much of the world, differences are probably more due to sampling differences (swath and pixel selection) than algorithm. I suggest the authors refer to that paper in their revised manuscript, and go from describing things as "inconsistent" to try to for example make recommendations as to which data sets might be better or worse suited for different applications. Recommendations like that, with evidence, are more useful than just declaring "inconsistency". Perhaps "comparisons" or "consistency assessment" is a better way to describe the analysis in title and text.

Line 50: I think this should say 20th century, not 19th. I am not aware of any observation networks before the late 20th century. If there are, please provide references. Aerosol science didn't really start until John Aitken in the late 1800s.

Line 118: This should be "Holzer-Popp" not "Holzerpopp". The author's name is double-barrelled.

Line 121: This should be changed to indicate it is the NOAA AVHRR aerosol product. There is also a NASA GISS aerosol product (GACP), which is monthly-only and ocean-only, and a NASA Deep Blue aerosol product, which also covers land but is presently only available for limited time periods (I know 2006-2011 is available). It would be good to clarify what is used and why here. Deep Blue and GISS also provide 550 nm while NOAA AVHRR does not. Perhaps one of those could be added.

Line 125: Authors should state more clearly here that they are using the 0.63 micron AOD (aot1 SDS), as it is important to note that this is different from the 550 nm AOD provided by most other data sets, and would result in offsets dependent on aerosol type. The authors do not seem to mention this later in the paper (e.g. line 179 says the satellites are at 550 nm). Was the AVHRR AOD somehow extrapolated to 550 nm like the others? Or was it left at 630 nm and the wavelength dependence neglected?

Line 139: Authors are missing references for the version 23 algorithm they are using here. Martonchik/Kalashnikova are out of date. The water approach is discussed by Witek et al (2018): https://www.atmos-meas-tech.net/11/429/2018/ The land approach is discussed by Garay et al (2017): https://www.atmos-chem-phys.net/17/5095/2017/ I suggest authors read and cite these papers, since it appears they have been referring to older documents.

Line 155: Sayer et al (2014): https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2014JD(is a more complete reference for the DTB products than Levy et al (2013). It also provides a comparison for DB, DT, and DTB. It will also be useful for the authors' analysis since it provides similar discussion about the level of consistency between the data sets. All the papers cited here are about Collection 6 but I know the MODIS teams and they did not publish papers about Collection 6.1 yet (still in review).

Line 163: I am not sure that the FM acronym for "forward model" is needed here. I don't think it is used later.

Line 166: Somewhere in this section I would add a note to state that this is not a

validation but a comparison, because the authors are using monthly data and not instantaneous data. So there are sampling differences contribution as well as retrieval quality. The authors are not performing a true validation exercise here.

Line 189: The authors insert four self-citations for a one-line equation developed by other people something like 75 years ago. This seems a little excessive. Please remove these citations or replace with ones to the original work by Angstrom.

Line 190: I recommend the authors account for lag 1 month autocorrelation in the time series. This is commonly done in AOD trend analyses as the data can be significantly autocorrelated on these scales (because large-scale systems and seasonal patterns can persist for weeks to months). This will keep the same trend values but affects the estimated uncertainties on the trend. See Weatherhead et al (1998): https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/98JD00995 for examples how to calculate this.

Lines 197-200: I am not sure that "correct trend percentage" makes sense. If a trend is close to 0, you will end up with a lot of apparently "wrong" trends if the sign is wrong, even if the conclusion that there is almost no trend is correct. For example, if you had trends of +0.01 from AERONET and +0.1 from satellite the authors would say this is "correct" even though the difference is huge. But if you had +0.005 from AERONET and -0.005 from satellite the authors would classify it as "incorrect", even though they are both small and probably statistically indistinguishable within trend uncertainties. A further problem is that this makes the implicit assumption that AERONET trends are perfect when of course they also have some measurement uncertainty and sampling uncertainty. I suggest that a better metric would be to report the "consistent trend percentage". This could be calculated by checking whether the satellite and AERONET trends are consistent within each uncertainty or not. This is a more fair and statistically appropriate test. The authors could also report those situations in which the AERONET estimate is too uncertain to be useful. I doubt that five years is enough to estimate a trend robustly in many cases, due to significant annual variability. So quite possibly the

uncertainty on the AERONET estimates even is quite high.

I also wonder if seasonal trends would be better than annual, because we know that aerosol patterns show strong seasonal features (so trends in seasonal behaviour could be masked in an annual trend analysis). The authors need to justify this more strongly.

Line 209: The subscripts are very long. I suggest replacing AOD_RETRIEVAL with AOD_R (for "retrieval") or AOD_S (for "satellite"), and AOD_AERONET with AOD_A. This will make it more readable.

Line 210: Correlation is not useful when the data range is small compared to the uncertainty on the data. You could have a great data set but still have a small correlation. For example, over the open ocean AOD does not change much, so a low correlation is scientifically not much of a problem for most scientific applications, as long as bias and RMSE are low. The authors should note this because a lot of the maps and discussion rely on correlation.

Line 211: This EE is an expected envelope for level 2 error over land only, not for level 3 and not for water. It is not meaningful for level 3 data, and it is misleading to apply it that way. There is at present no error estimate for satellite level 3 products. I suggest the authors remove this quantity because it is misleading. In my view the other statistics are enough. This also requires removing from the discussion later on. Either remove it or create and justify some metric for what is an acceptable EE on the monthly data. My feeling is that a monthly level 3 EE should be smaller than the level 2 one, because some error sources should cancel out.

Line 219, 230, 468, Table 2, Figure 10, and elsewhere: No, this is not a validation, it is a comparison, because you are using monthly mean products and not level 2. Validation requires a ground truth. There is no ground truth for monthly data because there is no instrument sampling continuous monthly data. AERONET is only a validation for level 2 data. The authors should change the wording because it is misleading, and word choice matters. The analysis the authors are doing here is fundamentally different from

the dozens of published level 2 validation papers, and it is important not to muddle the issue.

Line 295: Again, it is not ideal to provide a single self-citation here when these issues have been documented by many algorithm teams for many years.

Line 362: Throughout section 6 the authors talk a lot about trend significance. However something which has been overlooked is that since there is multiple hypothesis testing going on (many data sets and locations are being tested for trends), there could be a significant fraction of false positives. See e.g. Wilks (2006) for more on this: https://journals.ametsoc.org/doi/10.1175/JAM2404.1 So, the authors should make some quantification about the expected false discovery rate. Further, statistical significance is only one factor. Figures 11 and 14 are a prime example of this problem. Scientific significance is another. If you get a trend of 0.001 with an uncertainty of 0.0001, that is statistically significant but scientifically not important because it is so small. But if you get a trend of 0.1 with an uncertainty of 0.1, that is not statistically significant by traditional tests, but is potentially very important, because 0.1 is a large potential trend. The authors here seem to focus on statistical significance and sign rather than actually looking at the numbers. This is quite superficial. I would like to see the whole section reconsidered.

Line 496: “Goddard”, not “Godard”.

Figure 7 (and associated discussion): I do not like annual mean maps in general because AOD patterns and sampling are strongly dependent on season. So in some areas there will be a difference just because data are coming from different months. And in some areas things could look to be in closer agreement than they really are, if biases in different seasons are opposite and can cancel out. Annual mean AOD is also not meaningful for most applications. I would prefer to see this figure and discussion instead as a composite of four sets of seasonal plots. This would be a closer to apples to apples comparison, and also allow an examination of seasonal variability.

Figure 8, 12: Could this be redrawn to show coloured symbols instead of bars? In some cases the bars are overlapping and so it is hard to tell which is which. It can also give misleading impressions. For example, in land ENAM the black and pink are overlapped. I guess black was drawn first and pink second, so pink is on top. So the impression is that black is lower than pink, because we can only see the bottom of black. But in reality, because so much of black is hidden, it probably means that black and pink are very similar. Coloured symbols instead of bars would be clearer and easier to tell.

Figure 9: since this is not a validation but a comparison, it would be better to say “offset” rather than “bias” here. Bias implies an offset with reference to a truth, and we have no truth. Word choice is important.

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-1130, 2018.