Response to reviewer 1 (Andy Sayer)

I'd like to thank Andy for his time and many useful comments. I think the paper has improved in clarity as a result. The on-going discussion on how to calculate annual averages (arithmetic vs geometric) is also an interesting one, and I'm happy to contribute.

The reviewer suggests condensing the paper. Other reviewers have suggested this as well, pointing out the use of supplementary pages. I have decided to move part of the AOT representation discussion (e.g. variations by regions) and the entire AAOT representation discussion to a supplement. That should significantly shorten the main paper, without detracting from the main conclusions. The original AAOT analysis will be available for those with an interest in it.

Page 4 line 7: "sphotometers" - should be sun photometers?

Corrected.

Section 3: This has only one subsection. Could that subheader (3.1) be deleted? Or else another one be added (e.g. for the text summarising the difference between S17 and here)?

Deleted.

Page 7 line 11: Holben (ACP, 2018 https://www.atmos-chem-phys.net/18/655/2018/) is a good reference for the DRAGON campaigns, which could be cited here.

Agreed

Section 4: the evaluation of G5NR is presented mostly in terms of correlation coeffi- cient and regression slope of AERONET vs. G5NR mean and standard deviation of AOT/AAOT. In a sense each site is collapsed down to provide a single data point for the analysis. So this is somewhat different from typical validation analyses where one looks at individual AOT pairings (and in those cases regression is not so appropriate; it is probably fine here, see next paragraph). The reason for this is that G5NR is a nature run so corresponds not to the real (historical observed) world but a realistic world driven by the model. I have used G5NR data before so am familiar with this subtlety, and the author does state it, but I wonder if a less-familiar reader might be confused. I wonder if this point can be hammered-home a bit more with tweaks to working. For example page 7 line 3 says "simulated AOT shows good agreement with the observations" - this might be changed to read "simulated site-mean AOD" to reinforce the point that we are comparing site averages, not individual points, here. Unless I have misunderstood what is being done. That is one example, but the same applied throughout the section.

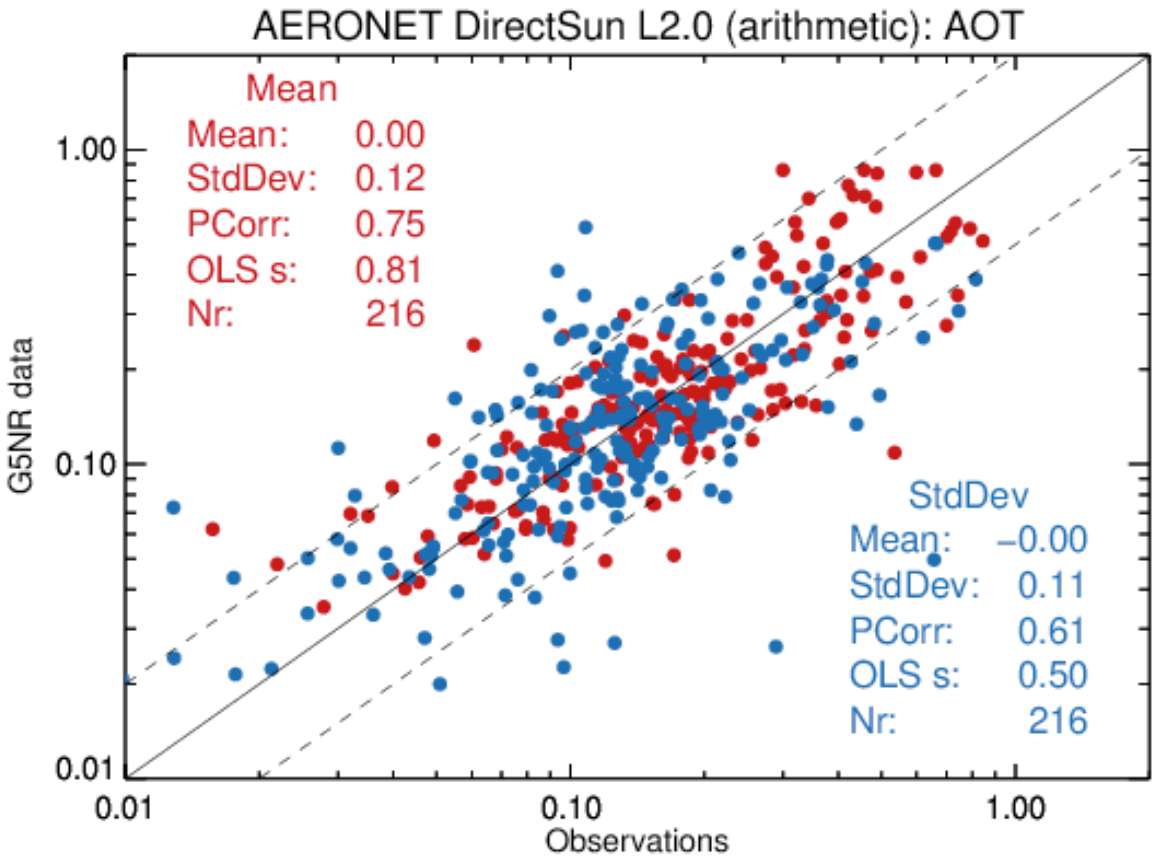
I agree that I can do more to impress upon the reader this is a free run. Very interestingly, yearly AOT per site agrees reasonably well with observations. While in satellite research, it is more common to provide error statistics on daily scales, in model research longer time-scales are more usual. First of all, we want to be able to represent the "base" state of the atmosphere (I do provide additional information in the standard deviation, i.e. variability per site, of AOT). The correlation in these yearly values expresses the ability of G5NR to realistically simulate the spatial distribution of annual AOT (at scales of AERONET separation distances).

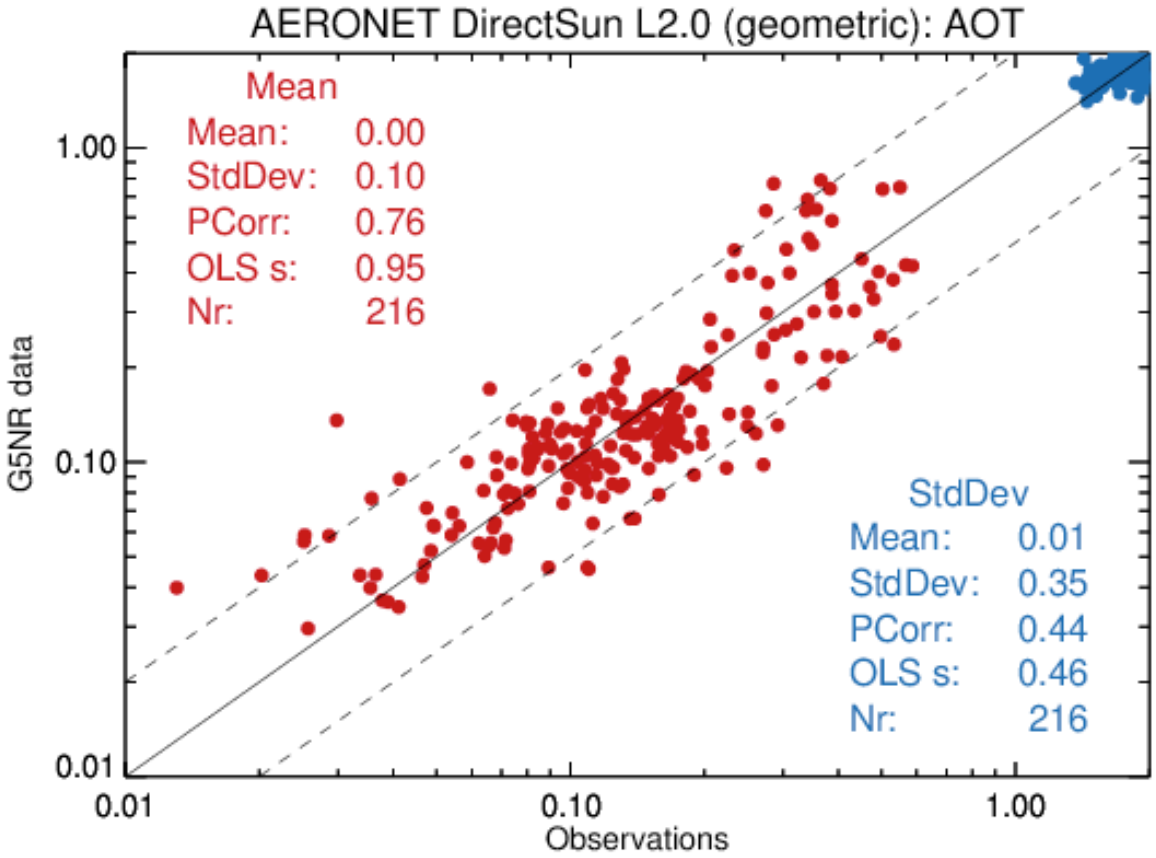
More generally the use of correlation and slope can be a bit problematic for AOT analy- ses, because of the distributions of the data and their error

characteristics. It is probably fine here because we are looking at summary statistics for individual sites, rather than individual points themselves, which is a different application from normal. However, because AOD distributions are skewed (and often close to lognormal on timescales like the year evaluated here – see the Sayer and Knobelspiesse reference mentioned above), I wonder if this analysis and Table 5 might be better presented in terms of geo- metric mean and geometric standard deviation (i.e. in log space). Perhaps the author could do this (doesn't necessarily mean both sets of analysis need to be shown in the paper); if the results are basically the same, great, but if not, it reveals something about limitations of the model simulation.

An interesting idea and easily implemented. I have followed Sayer & Knobelspiesse with interest and suspect we will have many discussions on such issues in upcoming AEROCOM/AEROSAT meetings!

Below I show the evaluation of G5NR, using either arithmetic (as in my paper) or geometric (as advocated by Sayer & Knobelspiesse) means. For definition of geometric mean and standard deviation:
https://en.wikipedia.org/wiki/Geometric_mean and
https://en.wikipedia.org/wiki/Geometric_standard_deviation





Using geometric mean and standard deviation has the following consequences for network statistics (the text in the figure).

- Bias in mean AOT per site hardly changes
- Spread in mean AOT per site decreases by 20%. However, mean AOT per site also decreases by about 20%, so this is not surprising.
- Correlation for mean AOT per site hardly changes
- Regression slope for mean AOT per site improves significantly
- Standard deviation AOT per site now shows rather large values and significantly lower correlation.

If I calculate standard deviation AOT per site from an arithmetic mean over logarithmic AOT (as we discussed off-line), evaluation is still poorer than when using arithmetic mean over AOT.

In short, I see no significant improvements in evaluation statistics when using a geometric mean. The exception would be the regression slope and I think it is worthwhile to explore this further. The use of geometric standard deviation has a negative impact on the correlation and should be used with caution.

Page 7 line 20: it might be worth being clearer here that the AERONET AOT require- ment for level 2 is 0.40 at 440 nm. For an Ångström exponent (AE) of 2 you get to about 0.25 at 550 nm from this. But for dust-dominated columns with an AE around 0.5 you are around 0.35. So the threshold translates to 550 nm differently dependent on aerosol type. As this threshold is mentioned again on page 9, I think it's worth devoting another line or two to the point here. I realise that the author is using 0.25 as a thresh- old on the simulation here (i.e. not using the actual thresholds AERONET applies

in each case), but that will affect the conclusions systematically at e.g. dust-dominated sites (true AERONET sampling will be poorer than the OSSE suggests because the true AERONET threshold for dust will be more like 0.35 than 0.25).

I agree. This was also pointed out by another reviewer. As you say yourself, it essentially means that over dusty sites I present a best case for the representation errors in Inversion L2.0. More importantly, though, is that the brunt of my analysis concerns Inversion *L1.5* and this will not be affected by the threshold.

Page 7 line 21: I think this should be "fewer", not "less" (in both cases), because the observations and sites are countable.

Corrected. I thought it sounded strange but couldn't pinpoint why ☺.

Figure 1: it's not clear what the distinction between solid and dashed lines in the lower panel is here. I know it is pairs of correlation and slope for mean and standard deviation of AOT/AAOT. But I did not see which is which given in the caption or text.

This has been corrected in the caption.

Figure 2: I know there were reviewer and editor comments about number of figures. I think this is one which could potentially be cut (or moved to a supplement) and sum- marised in the text instead, since the main point (if I understand correctly) is that the statistics for the level 2.0 inversion data are not that different from the less-restrictive level 1.5.

Correct. I also feel several figures can be moved to a supplement, Fig. 2 included.

Page 8 line 5-6: I would check in with a member of the AERONET team about this. I don't know what the main uncertainty source leading to AERONET AAOT uncertain- ties (which are driven by SSA uncertainties is). If it is calibration then that would have an air mass factor dependence so could manifest in apparent daily variation (and vi- olate the author's assumption). If it is something like surface albedo then that may be more of a constant uncertainty which might (consistent with the author's assump- tion) not affect daily max vs. min AAOT so much. However in Tom Eck's 2014 paper (https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/jgrd.50500, Figure 4), looking at the variation of SSA at Mongu with day of year, he found different slopes in different years, and attributed this to calibration uncertainties (as the sensor is calibrated be- fore and after each individual deployment, calibration uncertainty is systematic within a year, but random year-to-year). This implies that calibration may be one of the largest contributors, in which case it's possible that the daily variation of SSA (and hence AAOT) is affected (although that paper did not look at SSA diurnal variations). It would probably depend on both the daily variations of SSA and AOT – if AOT varies a lot that may win out over any false signal from SSA. I am not sure whether anyone has looked in great detail but the AERONET team might.

I appreciate your points and I may have been too positive about this. But my line of thinking is that the differencing inherent in a daily MAX – MIN AAOT value mitigates the impact of retrieval errors.

If such errors are constant throughout the day, this is a trivial statement.

If such errors behave entirely random, a yearly average of the difference will not be affected much either

Obviously, correlated but time-varying errors do exist and can be introduced by e.g. a calibration error. But in that case, the error in the difference is unlikely to be larger than the error in individual AAOT.

I've contacted Tom Eck (AERONET) and Oleg Dubovik (Lille U., developer of original Inversion scheme) about this issue and they agreed with my reasoning. Proper research is probably needed to put this on a firmer footing and I'll amend the text accordingly.

Page 8 lines 12-13: another factor is instrument maintenance issues (e.g. cleaning, replacement when it is sent back for cleaning). Even if this is only 1 week per year then that's still up to 2% coverage (or about 1% when accounting for daylight), which is simi- lar to the difference observed at many sites. So I'd say "meteorological differences and site maintenance issues" or something. This is addressed in the following paragraph but relevant for the direct-Sun data discussed here too.

Agreed.

Page 8 line 17: "several times per day" - I believe it is at specific optical air mass factors but I did a quick look and can't find what those are. I want to say it is a maximum of 6 per day. In the newer data they have hybrid scans nearer solar noon which can extend this, but for the year 2006 simulated by G5NR these were not available. So in that sense the newer AERONET data will fill in some of the gap that is in the observation but not predicted by the model.

I was deliberately vague as I don't think this is relevant at this stage. The OSSE overestimates observational coverage, and this can be due to a whole list of reasons. Note that my sensitivity study suggests that this over-estimation of temporal coverage (e.g. Fig 9 and) has no large impact.

Page 8 lines 16-21: One issue is that the inversions require a high degree of azimuthal symmetry (see their QA document at https://aeronet.gsfc.nasa.gov/new_web/Documents/AERONETcriteria_final1.pdf). So for example if an aerosol plume is thicker to one side of the site than the other, then the scene may be rejected. I don't have a good idea how often this happens; the AERONET team might. I wonder if that is one of the larger factors accounting for the overestimation of AERONET inversion coverage. There are a few other things too, e.g. the AOD threshold for AERONET is stricter for dust-dominated scenes than that applied in this study (see earlier comment) which would affect some of the sites in the tropics.

I am aware of these issues. As a matter of fact, because it requires high degrees of azimuthal symmetry, Inversion has a built-in check for *spatial* representation errors. That check will in turn lower temporal coverage at any site. I had no idea how to represent that so left it out of the OSSE. But it is worth discussing this.

Page 8 line 25: I believe style guidelines for the journal require sequential appearance of Figures; here Figure 12 is mentioned for the first time, in between 4 and 5. From context it is clear that Figure 12 should not be shifted back here, but the Copernicus style guide disagrees. Perhaps this sentence could be shifted later in the paper instead (and so call back to this section).

I prefer to leave it as it is. It is sometimes unavoidable that one figure is referenced several times in a paper. I have used the location of the main discussion of a figure in the text to order the figures.

Section 5: I realise that this is framed as relative errors throughout. But many appli- cations require absolute uncertainty, so absolute values are also important. So per- haps some text and/or a table could be introduced, with a summary of what fraction of sites the representation error is smaller than some threshold (perhaps the nominal AERONET AOT uncertainty of 0.01, or the GCOS goal of max[0.03,10%]), for each grid size and time stamp? A large relative sampling uncertainty might be unimportant for a pristine location, for example. Alternatively Figures framed that way could be placed into a Supplement.

Agreed.

Section 5.1, title: I suggest "Representation errors in yearly AOT" to make it clearer up front this is about comparing yearly aggregates colocated in different ways. It will help make the contrast with section 5.2 (monthly) clearer up-front.

Agreed.

Figure 6 caption: "Yeraly" should be "Yearly"

Corrected.

Figures 6, 7 (and dots in 21): can these be regenerated with a different colour bar? The rainbow doesn't print well, emphasises certain parts of the data range but suppresses others, and can't be understood in greyscale or by many colour blind readers. The "viridis" palette is a good alternative, and other op- tions can be found online. Here's a link to an IDL implementation from the CRU: https://crudata.uea.ac.uk/~timo/idl/mkviridis.pro Also, panels are presented as left/right but captions indicate top/bottom, and it would be good to add latitude/longitude labels and/or national borders to this for ease of reference if the reader wants to look up the value for a specific site.

Thanks for the link. Figures will be remade.

Page 9, lines 4-5: yes, it is clear from this Figure that the bias is negative much more of- ten than it is positive. This implies that higher-AOT times are not sampled by AERONET as often as they should be. One explanation is coincidence (plumes systematically avoid them) but I find that unlikely. So, what is the other mechanism? Could this be the clear-sky bias, i.e. AOT is higher near clouds but near-cloud cases are not sampled? I wonder if there is some way to quickly examine this (e.g. rerun part of the analysis with a cloud fraction threshold of 0.9 instead of 0.01, see if the bias in the representation error shrinks)? Ok, reading ahead to page 10, from Figure 13 it looks like it might be the clear-sky bias. Perhaps that figure and text could be moved up a page. This part — quantification of clear-sky bias — is to me quite an important result.

It is the clear-sky bias. My code generates error estimates for individual masking factors (daytime/nighttime, cloudiness, lower AOT threshold) and identification is of the main cause is trivial. I will move this discussion forward.

Page 9 line 10: this is an important point, I'm glad the author highlighted it again in the Conclusions.

I think it may similarly have consequences for AEROCOM model evaluations

Page 9 line 14: I would say “limitation of” rather than “issue with”, to help emphasise this is due to the measurement type rather than being something which was done wrongly.

Agreed.

Page 9 line 22: is -410 m really correct? Which site is 410 m below sea level?

Dead_Sea

Page 9 lines 29-31: the symbol r was previously used for correlation (e.g. prior para- graph), now is being used for Kinne’s rank score. Also, this second use of r does not appear to be stated explicitly in the text. I suggest finding another symbol for the rank score and defining it explicitly in the text. Perhaps capital regular R rather than lower-case italic r.

Kinne uses “r” so I’d like to use it as well. But I will make sure it’s clear this is a different “r” from the rest of the paper.

Page 10 line 1: I would say “typically cannot retrieve aerosol when there are clouds”. CALIOP, for example, can retrieve under some clouds. Other retrievals could be ex- tended to do so (see e.g. Lee JGR 2013 https://doi.org/10.1002/jgrd.50806 for an attempt I was be involved with — I don’t know that this paper needs to be cited or dis- cussed, just providing it here for an example). I suggest the rephrasing because in part this is a sensor issue but in part it is an algorithm issue.

Ok.

Page 10 lines 10-11: I am not sure that I follow this. I agree that it will be true if there is correlation from year to year as well. Which there almost certainly is in many parts of the world. But I think that’s a bit different from the month-to-month correlations here. I think this should be clarified/spelled out a little more clearly.

Note that I am talking about the increase of correlation in 2006 between January and months like November and December (11 or 12 months apart!). Obviously, I can’t prove this is repeated every year but there are good reasons to assume this will happen.

Page 10 line 13: I think the words “radiation records” are missing from the end of the Schwarz paper cited here.

Thanks.

Figure 16: what are the dashed lines here?

Y=2x and x=2y, for convenience. Now explained in caption.

Page 10 line 21: “criterium” should be “criterion”.

Corrected.

Page 10 lines 21-22 and Figures: The impact of the AOT threshold imposed on AAOT representivity is clear. However I am confused because I thought from

Table 2, the AOT threshold was taken as 0.03 for level 1.5 data, and not 0.25 (which was for level 2 data). The text (and Figures) here refer to level 1.5 data, but to the 0.25 threshold. Is there a typo here or have I misunderstood? If the threshold was 0.03, why is the bias so positive? If it was 0.25, why are we discussing level 1.5 data and not level 2 data?

Thanks. It would appear that an earlier edit went wrong. Clearly, the AOT>0.25 statement has no relevance here.

Page 11 line 6: there is a missing Figure reference in this line (appears at ??). From context I think that this should be Figure 18, which seems to fit and is not mentioned elsewhere in the paper.

Corrected.

Page 12, lines 11-12: Thank you for making this list available. I downloaded the file from the DOI linked to the citation and it was clear.

You're welcome. Comments always welcome, also after publication. This will hopefully be an evolving document.

Page 13 lines 20-31: I'd personally split this out as a bulleted list (and perhaps the point about the Wang analysis too), to better drawn attention to these conclusions and recommendations.

Thanks for the suggestion.

Figures 8, 9, 10, 12, 13, 14, 15, 17, 18, 19, 20: I think a note should be added here to state that the colours (and, except for Figure 15, numbering legend) follow Figure 5.

Ok.

As a general question: Is one take away that AERONET and satellites should if possible provide additional hourly products, for intercomparison purposes? Since hourly collo- cation minimises the representation error for longer-term aggregates, making these more readily available might spur users to use them (rather than the current approach which is more or less monthly collocation).

But don't these data already come at hourly or daily resolution? Of course, users seem fond of the monthly L3 products and I am not sure how to change that. Removing monthly L3 data from archives would be my preferred option but I can see that would be unpopular.

Language comment: I think in some places the term "uncertainty" should be used instead of "error". The calculation of representation error via difference between the differently-sampled G5NR simulation is an error. But I think when talking in a larger sense, we are using this representation error (from the OSSE) to estimate the actual representation uncertainty (which we don't know for sure). Also when talking about AERONET inversions, we should be typically talking about the uncertainty in the re- trieval (as the error is not known). I suggest checking individual uses of these terms in the papers.

Ok.