Response to all reviewers

Before detailing my specific response to each reviewer, a few general comments:

- Figures that contain maps have now been remade with colour schemes that are hopefully more readable to colour blind people.
- Several figures have been moved to a supplement (mostly figures to do with the AAOT analysis)
- A few figures have been removed as I considered them too technical for a general audience (e.g. the figure used to discuss the nature of representation errors per site: random vs bias)
- Figure 1 (evaluation of G5NR) has been remade as I realized that I plotted Root Mean Square Differences instead of standard deviations (in blue). This partially changes the figures (results for the site mean are unaffected). Overall conclusions do not change much, except the poor performance of G5NR in temporal variability of AAOT is more pronounced.
- A table with absolute values of representation errors has been added, to compare against measurement error (representation errors are significantly larger).

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,

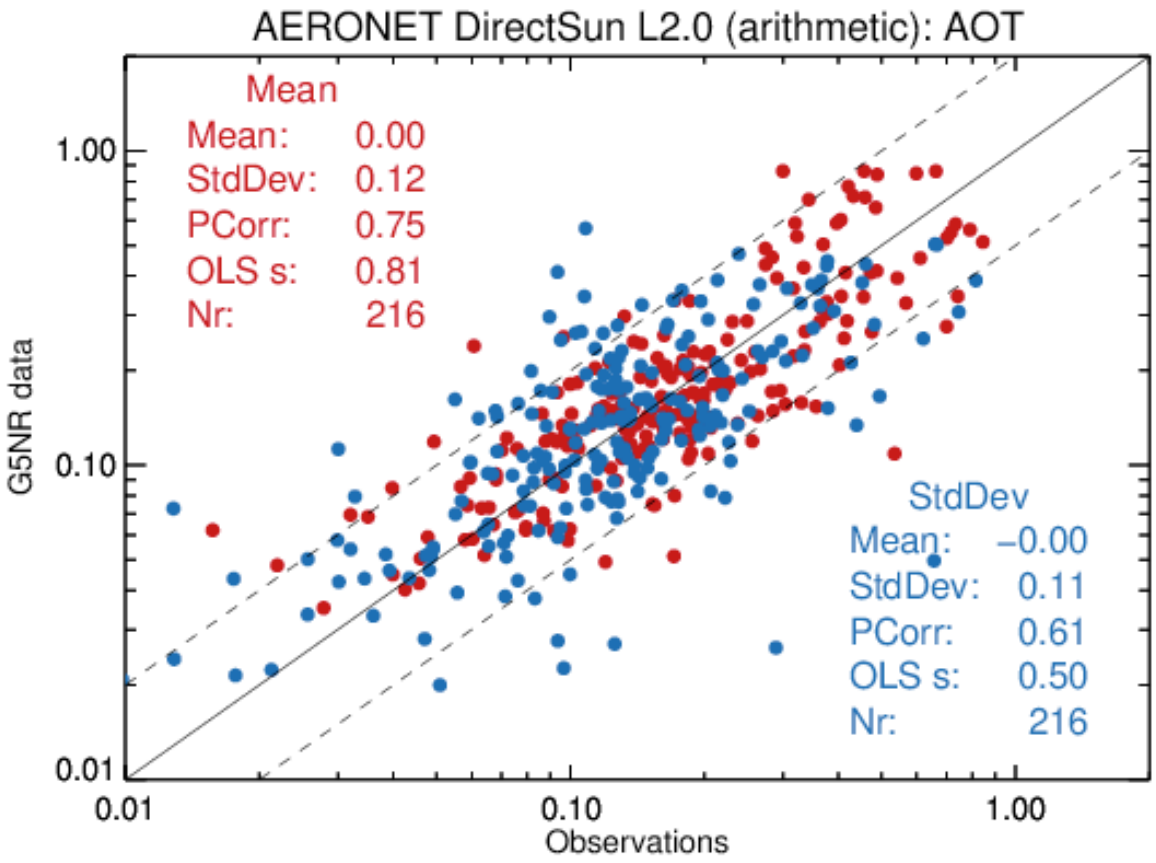
see next paragraph). The reason for this is that G5NR is a nature run so corresponds not to the real (historical observed) world but a realistic world driven by the model. I have used G5NR data before so am familiar with this subtlety, and the author does state it, but I wonder if a less-familiar reader 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.

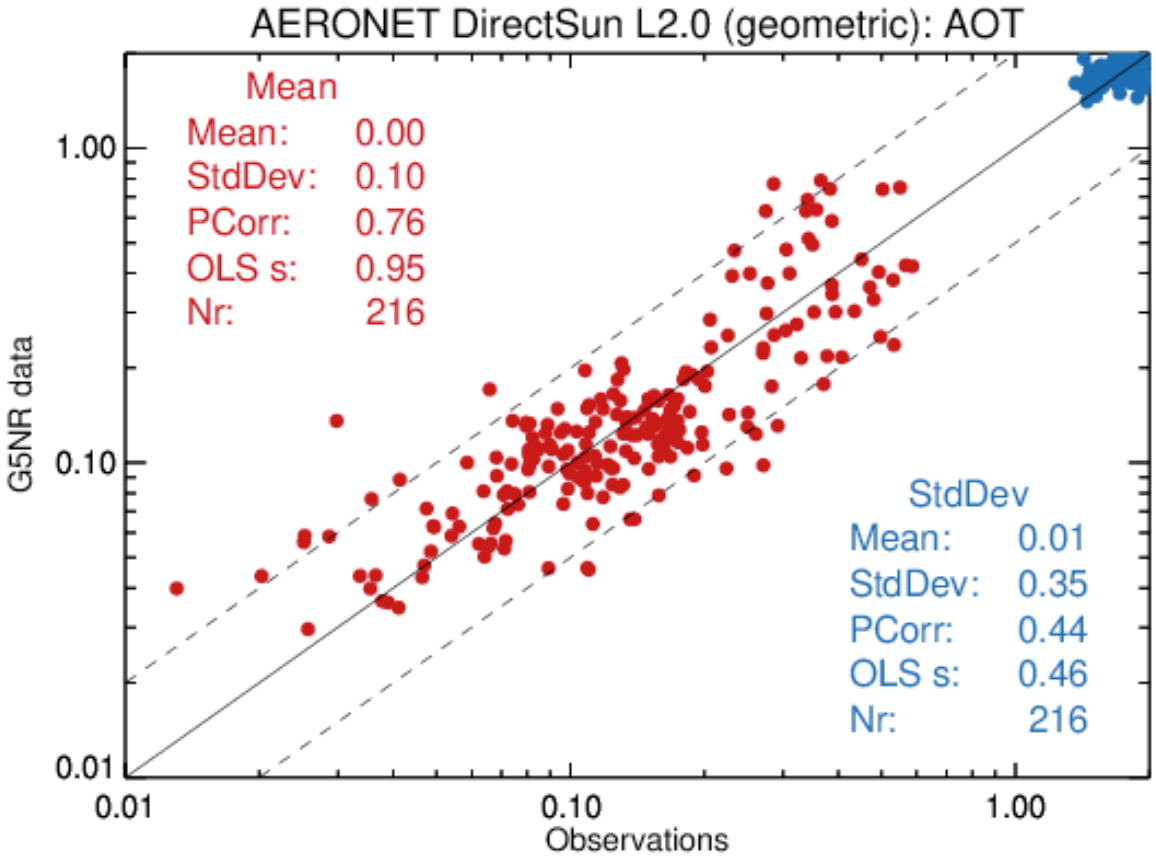
I agree that I can do more to impress upon the reader this is a free run. Very interestingly, yearly AOT per site agrees reasonably well with observations. While in satellite research, it is more common to provide error statistics on daily scales, in model research longer time-scales are more usual. 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


AERONET DirectSun L2.0 (arithmetic): AOT
Mean
Mean: 0.00
StdDev: 0.12
PCorr: 0.75
OLS s: 0.81
Nr: 216
StdDev
Mean: −0.00
StdDev: 0.11
PCorr: 0.61
OLS s: 0.50
Nr: 216
G5NR data
Observations
0.01
0.10
1.00


AERONET DirectSun L2.0 (geometric): AOT
Mean
Mean: 0.00
StdDev: 0.10
PCorr: 0.76
OLS s: 0.95
Nr: 216
StdDev
Mean: 0.01
StdDev: 0.35
PCorr: 0.44
OLS s: 0.46
Nr: 216
G5NR data
Observations
0.01
0.10
1.00

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 of the main cause is easy. 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.

Response to reviewer 2

I'd like to thank the reviewer for their time and many useful comments. I think the paper has improved as a result of their feedback.

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 most of the 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.

[..] These seem to be potentially interesting results which are not discussed in this pa- per: representation errors also vary based on the aerosol type itself [..]

I think there is definitely room for a study on the impact of aerosol species on representation errors. Here and there in my papers I have alluded to this. E.g. in S16a the differences between black carbon, sea salt and sulfate were briefly discussed. It is not so much the species itself but the spatio-temporal distribution of its sources that is the important factor. A proper investigation would be outside the scope of the present paper which already is quite large. For the impact of dust on the G5NR evaluation (which is different from an analysis of representation errros), see e.g. Table 5.

Does Fig 17 have a different collocation protocol for the brown bars than the others?

No, it doesn't but I forgot to update the caption to bring it in line with other figures. Changed now.

p. 3 , Lines 28-29: I believe CERES "cloud fractions" are derived from their collocated MODIS instruments.

The reviewer is correct but I prefer to stick with CERES cloud fraction as this is the form used in Gelaro et al. 2015. They are cloud fractions derived from MODIS, specifically for CERES.

p. 7 Line 5: "correlation (~0.45)". To what does this ~ refer?

It means, "about". I believe this is standard usage: https://en.wikipedia.org/wiki/Tilde#Mathematics .

p. 7 Line 20: as the other reviewers said, this is not strictly true; the L2.0 data have a minimum AOT of 0.4 at 440nm, which here has been interpolated to ~0.25 at 550nm. I'd clarify this point.

The reviewers are of course correct and this point will be clarified.

-Figure 1: which is the solid and which is the dashed line? This should have a caption.

Thank you. Red solid is correlation, red dashed is slope, blue solid is mean, blue dashed is standard deviation. This information has been added to the caption.

Figure 21: the color bar from Fig 7 should be reproduced here; also there should be units added to the BC emissions shading.

I'll try to add a colour bar for the representation errors. The unit for bc emissions is mentioned already in the title.

-throughout the paper, I believe the singular form of "criteria" should be "criterion," not "criterium."

Changed.

-Figures 6 and 7: captions say top/bottom, but should say left/right. Also "yearly"

Changed. Actually, the top/bottom issue is due to different formats used for Discussions and Final publications.

-Figure 23: this figure could benefit from a 1:1 line to guide the eye.

Ok.

And thanks for bringing those typos and misspellings to my attention. They have been corrected.

Response to reviewer 3

I'd like to thank the reviewer for their time and many useful comments. I think the paper has improved in clarity as a result of their feedback.

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 most of the 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.

Sometimes figures are referred by using "Fig." and sometimes "Fig".

All changed to "Fig.".

It would be interesting to see or at least have a comment on the absolute representation errors. This would show if high representation errors mainly correspond to small AOT values only or are there relatively large errors present also in cases with large AOT.

A good point. I started using relative errors in S16b as it allows more easily a comparison across different types of measurement. Also in the current paper, it allows comparison of AOT and AAOT representation errors. But I will include a paragraph on absolute values of these errors.

p.2 l.14 "return times" Would "overpass times" or "revisit times" be more commonly used term to be used here?

Yes, I'll use revisit times.

p.5 l.1 "The maximum cloud-fraction was slightly tuned..." Please clarify what you mean by "slight tuning".

Yes, I can see how this is confusing. I can choose values between 0 and 1 and ended up using 0.01 because the results agree *slightly* better with the observations. As the original text stated, the impact is small. Also, I only explored five different values (0.01, 0.1, 0.5, 0.9, 0.99) so in that sense the tuning was coarse. I have removed 'slight'.

p.6 l.7 "...we will limit our analysis to latitudes below 60°." In Figure 4, there are stations at above 60 degrees.

True, AERONET sites exist at higher latitudes and my data included those as well. For the evaluation of G5NR, I used all sites. For the representation study, I included only sites below 60 latitude (except in Fig 4 & 5). Text now reflects this.

p.7 l.9 For reproduction of the results, please list the sites that were removed from the analysis.

They were only removed from the analysis for one result (Table 5, line Europe*). Throughout the paper they have been used in the analysis of representation errors. I have amended the text to clarify this.

p.7 l.20 Is AOT threshold of 0.25 correct? To my understanding the threshold at 440 nm is 0.4 and it depends on the spectral dependence of AOT (Angstrom Exponent) what it will be at 550 nm. So for me this seems a bit low value for the threshold. Please make sure the reader understand that you have used a "non-standard" value of 0.25 or correct to match the true AERONET threshold (throughout the manuscript, same limit mentioned for example on p.9 l.13).

The reviewer is correct. The text will be modified accordingly. Note that this has almost no impact on the paper as I mostly study Inversion L1.5 data. Only in Fig 9, where a comparison is made between L1.5 and L2.0 will this affect the L2.0 analysis (i.e. representation errors will be underestimated). Note that this issue (AOD@550nm >0.25 instead of AOD@440nm >0.4) will mostly affect dusty stations (for which AOD@550 ~ AOD@440). Since most of my statistics are based on year-averages from stations and dusty stations form a minority, I do not expect very big changes.

p.9 l.22 Altitude of -410 meters, is this correct?

This is correct. These are geopotential altitudes, see also Table 1 & 2.

p.9 l.29 Here notation "r" is used for representation ranking by Kinne et al. (2013). In some parts of the manuscript "r" is used to denote correlation coefficient so there is a conflict here. Please correct throughout the manuscript to remove the possible misunderstandings.

Yes, that is a bit unfortunate. “r” is a common symbol for correlation, which is why I use it. Kinne et al use “r” for their rankings. I will address this specifically when discussing Kinne rankings.

p.13 l.12 "G5NR and the OSSE are evaluated and found to show significant skill." This result was found for AOT, not for AAOT. Please clarify that this statement applies only to AOT to avoid misunderstandings.

I suggest to change this to: “G5NR and the OSSE are evaluated and found to show significant skill in AOT and reasonable skill in AAOT.”

p.18 Figure 1 Bottom row, what are the differences between solid and dashed lines?

Caption has been clarified.

p.21 Figure 9 Please define DS. Also on the upper right corner the text is overlapping with the figure and may be difficult to read.

Caption has been clarified.

p.25 Figure 16 What are the dashed lines?
p.27 Figure 20 "r" is not defined.
p.28 Figure 21 If possible, please add the another colour bar from Fig. 7.

And thanks for the typos etc.

Response to reviewer 4

I'd like to thank the reviewer for their time and many useful comments. I think the paper has improved in clarity as a result of their feedback.

page 3, line 9: It would be good to add a bit more information on the simulation data, notably that it is a free running (not nudged) simulation, possibly also a word on verti- cal resolution and output frequency (hourly or even less?; how 'high-resolution' is the model data with regard to time?).

Some of this information is already in the paper but I agree that it could be stated more prominently. I will modify the text.

page 3, line 24: Replace AOD with AOT, here and throughout the manuscript; likewise for AAOD and AAOT.

Rather, I have changed AOD to AOT to preserve consistency with the many figures in the paper. I know the WMO suggests to use AOD but AOT is often used to mean the same thing. When I checked usage in publications a few years ago, AOT was actually more common than AOD. As long as I am consistent within this paper, I do not expect any confusion to arise. I hope the reviewer finds solace in the fact I have started using AOD in my most recent submissions.

page 4, line 8: What do you mean by "here we will assume a potentially remotely sensed columnar product ... and consider its representation errors"?

I agree that is an awkward sentence. What I meant was: instead of the actual surface measurement, I will assume an AERONET-like columnar measurement of AOT. The sentence has been rephrased.

page 4, line 11: Given that various definitions of "representation error" exist in the literature, it would be helpful if the author could provide the exact definition he uses in this paper (e.g. reference to another paper; formula; description).

Agreed, the references are actually in the paragraph but have been moved up.

page 4, line 14: Here it is said that this work deals mostly with yearly and some monthly averages, yet many figures show hourly data. Please clarify.

Those yearly data can be constructed from data sampled in different ways (see Table 4). The best way (in my opinion, as supported by the paper) is to resample model data to the hours of the observations and then average over a year. This was discussed in p 5, l 4-10. I will take steps to clarify this further.

page 6, line 19: What do you mean by the sing-less error? Absolute error or root- mean-square?

It is unfortunate that "absolute" can mean two different things: 1) with no reference to a baseline; 2) without a sign. Mathematically speaking: $\tau_{obs} - \tau_{area}$

(instead of $\frac{\tau_{obs}-\tau_{area}}{\tau_{area}}$)

or $|\tau_{obs} - \tau_{area}|$. I mean the latter expression (but averaged). It is not an uncommon metric, similar to the standard deviation but it does not suffer as much from out-liers in the data.

page 7, line 4: I assume that by 'correlation' you mean R, not R^2. It may be helpful for the reader to explicitly say so.

I do not know how the reviewer's R is defined but I use the Pearson correlation coefficient (now explicitly mentioned in Sect 3.1),

page 8, line 22: When it is said that G5NR seems capable to realistically simulate the spatial variation of AOT and AAOT, "spatial" here seems to refer to different sites. It is not shown, it seems to me, how realistically G5NR captures the spatial variability of AOT and AAOT around a single site, including the adopted averaging distances between 0.5 and 4 degrees. It may be worthwhile to clarify this point.

I discuss this on p. 7, l. 11 but will repeat it here. The issue is of course there are no datasets available for such evaluation (DRAGON campaigns did not happen until 2012), although parts of W-Europe and the USA have several AERONET sites with distances of less than 100 km.

page 8, line 31: As grid box sizes are reduced, hourly collocation errors are reduced. Could this be because the physical connection (same cause, exchange of signal) be- tween two hourly time series at two distant points decreases with distance? Could the author comment on why the reported finding is (or is not) physically plausible?

This finding is to be expected from first principles: the comparison becomes more and more one of apples and oranges that look remarkably like apples. On the one hand, temporal sampling differences are reduced (by use of hourly protocol). On the other hand, spatial sampling differences are reduced (by decreasing box sizes).

page 9, line 5: Can something be said as to the (physical?) causes of the found east- west (North America) and north-south (Europe) gradient in representativeness?

It appears to be driven by cloudiness which, at least in the model, introduces temporal representation errors when using daily or yearly protocols.

page 9, line 21: Apart from the shorter atmospheric column, could it also matter that high lying mountain sites are often in the 'free troposphere', i.e., (somewhat) decoupled from the sources of (short lived) aerosols in the boundary layer?

This can definitely be part of the explanation for the larger representation errors for mountain sites. However, I would argue this "transport aspect" is part of the "shorter column" explanation?

page 11, line 9: Does it matter here, how missing values are treated when computing the annual mean?

For sure! When using the hourly protocol, missing data in the observational record are also removed from the G5NR data. This does not happen in the yearly protocol, resulting in large representation errors.

page 12, line 23: The author mentions once more the calculated meteorology. Overall, he seems to claim / find that meteorology is not that important for

representativeness. Is this indeed what he means to say? And, if so, how about phenomena like ENSO? Could, for example, the comparatively bad performance of South America be related to the presence / absence of ENSO in the model data?

That is not what I intend to say. Actually, meteorology is a powerful driver of both the temporal sampling of observations and the spatial distribution within an area. In previous papers (S16b and S17), I made an attempt at separating impacts of e.g. daytime/nighttime vs cloudiness and found the latter more important.

page 13, lines 13 and 20: Does this imply that meteorology is not that important for representativeness?

In line 13, I was talking about the evaluation of G5NR and not about the representation errors. In line 20, I am talking about representation errors. I believe these strong monthly correlations to be partly driven by meteorology (see also Sect 5.2 and Fig 15). However, it is difficult (maybe even impossible with the current datasets) to disentangle e.g. impacts of source distribution and wind advection. See also my answer to the previous remark by the reviewer.

page 13, line 24: It is not clear where the error of typically 20% globally comes from, I do not see this in the main text of the paper.

See e.g. Fig 5 which shows collocation errors for different boxes and protocols. For the yearly protocol, the mean sign-less error varies between 22-23%. It's important to realise that this is not a global bias: some sites will underestimate their area's average and others will over-estimate their area's average. The term "globally" has been removed and a reference to Fig 5 inserted.

Figure 1: One may add in the caption what the different line-styles in the lower row mean.

Agreed.

Figure 5: Any idea why there is an overall bias towards negative values? It seems unlikely that the (few) high lying GAW sites (and their shorter atmospheric column) alone can serve as an explanation.

Correct, negative biases arise from cloudy parts in the site's representative area: these tend to have higher AOT than the clear part (that include the site). An explanation will be added.

Figure 6: Any idea what the (physical?) reason is behind the found spatial gradients?

Cloudiness, as also explained after the reviewer's comment "page 9, line 5: Can something be said as to the (physical?) causes of the found east- west (North America) and north-south (Europe) gradient in representativeness? "

Figure 8: Any idea why Europe is so good and South America rather bad? Geography? ENSO? Number of sites? Other?

ENSO possibly. For sure a strong seasonal cycle in cloudiness that makes observations much less likely during SH autumn compared to SH winter season. This may be a quirk in the G5NR simulation, although I see something similar in the AERONET observations. Note how it is the yearly protocol (brown bar,

Fig 8) that is affected inordinately. i.e. this is driven by temporal sampling. The spatial representativeness of sites in Europe and S-America does not differ much.

Figure 12: Maybe refer in the caption to table 6 (explanation of r). Also, the figure seems to suggest that there is no connection between "r" from Kinne et al. and the relative representation error from this paper; the bars in the plot look pretty much the same for "all", "r=0", "r=1", and even "r>1" for yearly data. Please comment.

Actually, the text that refers to this Figure has more explanation. For 4 degrees, there seems to be little impact from “r”, but at 1 degree higher “r”’s result in smaller representation errors. I.e. the Kinne rankings agree with my results (at least statistically). But an importand but also subtle finding is that this is only true when using the hourly protocol; Kinne et al. did not consider temporal sampling of observations in their representation rankings.

Figure 16: What are the dashed lines?

Y = 2 x and y = x /2. Now explained in caption.

page 1, line 16: due *to* methodological choices
page 2, line 14: remove S16b
page 10, line 20: "for for" should read "for"
page 11, line 6: "Fig.??" should be properly referenced

Thanks for pointing out these typos and oversights.

Site representativity of AERONET and GAW remotely sensed AOT and AAOT observations

Nick A.J. Schutgens[1]

[1]Department of Earth Science, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, the Netherlands

Correspondence: Nick Schutgens (n.a.j.schutgens@vu.nl)

Abstract. Remote sensing observations from the AERONET (AErosol RObotic NETwork) and GAW (Global Atmosphere Watch) networks are intermittent in time and have a limited field-of-view. A global high-resolution simulation (GEOS5 Nature Run) is used to conduct an Observing System Simulation Experiment (OSSE) for AERONET and GAW observations of AOT (Aerosol Optical Thickness) and AAOT (Absorbing Aerosol Optical Thickness) and estimate the spatio-temporal representativity of individual sites for larger areas (from 0.5^o to 4^o in size).

GEOS5 NR and the OSSE are evaluated and shown to have sufficient skill, although daily AAOT variability is significantly underestimated while the frequency of AAOT observations is over-estimated (both resulting in an underestimation of temporal representativity errors in AAOT).

Yearly representation errors are provided for a host of scenarios: varying grid-box size, temporal collocation protocols, and site altitudes are explored. Monthly representation errors are shown to correlate strongly throughout the year, with a pronounced annual cycle. The collocation protocol for AEROCOM (AEROsol Comparisons between Observations and Models) model evaluation (using daily data) is shown to be sub-optimal and the use of hourly data is advocated instead. A previous subjective ranking of site *spatial* representativity (Kinne et al., 2013) is analysed and a new objective ranking proposed. Several sites are shown to have yearly representation errors in excess of 40%.

Lastly, a recent suggestion (Wang et al., 2018) that AERONET observations of AAOT suffer a positive representation bias of 30% globally is analysed and evidence is provided that this bias is likely an overestimate (the current paper finds 4%) due methodological choices.

1 Introduction

As the temporal sampling of observations is often intermittent and their field-of-view limited, the ability of observations to represent the weather or climate system is negatively affected (Nappo et al., 1982). This adverse effect can be described through a representation error, which allows comparison to e.g. observational errors or model errors.

Representation errors have been receiving more attention recently, in a variety of fields: solar surface radiation (Hakuba et al., 2014b, a; Schwarz et al., 2017, 2018), sea surface temperatures (Bulgin et al., 2016), trace gases (Sofieva et al., 2014; Coldewey-Egbers et al., 2015; Lin et al., 2015; Boersma et al., 2016), water vapour (Diedrich et al., 2016), cloud susceptibility (Ma et al., 2018) and even climate data (Cavanaugh and Shen, 2015; Director and Bornn, 2015). In the field of aerosol, most work has been on the representativity of satellite measurements (Kaufman et al., 2000; Smirnov, 2002; Remer et al., 2006; Levy et al., 2009; Colarco et al., 2010; Sayer et al., 2010; Colarco et al., 2014; Geogdzhayev et al., 2014), either using satellite data or model data. A new development is the use of local spatially relatively dense measurement networks (Shi et al., 2018; Virtanen et al., 2018).

As ~~aerosol~~ aerosols are known to vary over short time and spatial scales (Anderson et al., 2003; Kovacs, 2006; Santese et al., 2007; Shinozuka and Redemann, 2011; Weigum et al., 2012; Schutgens et al., 2013), aerosol studies are likely to experience large representation errors. Indeed, ~~Schutgens et al. (2016a) (S16a~~ Schutgens et al. (2016b) (S16b hereafter) showed that representation errors due to temporal sampling in both satellite and AERONET observations were of similar magnitude as actual model errors and often larger than observational errors. Similarly, ~~Schutgens et al. (2016b) (S16b~~ Schutgens et al. (2016a) (S16a hereafter) showed that the

narrow field-of-view of in-situ measurements could lead to large differences from area averages (monthly RMS differences of $10-80\%$ for 201×210 km^2, depending on the type of measurement and the location of the site). Recently, Schutgens et al. (2017) (hereafter S17) considered the combined impact of spatio-temporal sampling on the representativeness of remote sensing data (both satellite and ground-based). They provide representation ~~error~~ uncertainty estimates and optimal strategies when dealing with different observing systems (ground networks, polar orbiting satellites with varying ~~return~~ revisit times, or geo-stationary satellites). ~~S16b~~

In this paper, a global one-year high-resolution simulation of the atmosphere (GEOS5 Nature Run) is used to conduct an Observing System Simulation Experiment to estimate representation errors for remote sensing measurements of aerosol optical thickness (and its absorptive counterpart) as observed by the global networks AERONET and GAW. In ~~S16b~~ S16a and S17, regional high-resolution simulations covering a month were used to study representation errors. This prevented an analysis of such errors world-wide and on longer time-scales. In addition, the limited spatio-temporal domains made evaluation of the high-resolution simulation difficult. These issues are addressed in the current study. Note that the current paper does not replace previous work (which also considers satellite, in-situ and flight measurements) but extends it. In addition, the current study allows us to evaluate a recent suggestion by Wang et al. (2018) that representation errors in AERONET AAOT observations are positively biased (by $\sim30\%$) which would help to explain the observed underestimation of AAOT in global models (Bond et al., 2013).

Representation errors are not only determined by observational sampling but also by how these observations are put to use. If observations are used to evaluate models, different protocols (or strategies) exist to temporally collocate model data and observations. For instance, within AEROCOM, an oft-used strategy is daily collocation: daily averages of observations are collocated with daily model data. The different sampling of model and observations throughout the day are ignored (e.g. most remote sensing observations only observe a small part of the diurnal cycle). In contrast, hourly collocation uses hourly model data that is collocated with hourly averages of observations. S17 showed that in the case of remote sensing observations daily collocation allows significantly larger representation errors than hourly collocation. A third protocol would be yearly collocation which is seldom used these days in model evaluation as it yields large representation errors (~~S16a~~S16b). However, if remote sensing observations are used to construct a yearly climatology, effectively a yearly collocation protocol is used.

In data assimilation the representation error is often (but not always) thought to include effects from incorrectly modelled sub-grid processes. In this paper, the representation error is purely thought of as resulting from the different sampling by observations and models.

Section 2 describes the high-resolution simulation data and AERONET observations used in this study. The OSSE for estimating representation errors is briefly explained in Sect. 3 but more details can be found in S17. An evaluation of the high-resolution simulation with a particular focus on its use in an OSSE is given in Sect. 4. While the simulation shows deviations from AERONET observations, the agreement is deemed sufficient to study representation errors. Representation errors in AERONET AOT & AAOT are studied in Sect. 5. A ranking of AERONET sites in terms of their representativity is given in Sect. 6. As may be expected, the paper finishes with a summary of the conclusions (Sect. 7).

2 Data

2.1 GEOS-5 Nature Run

The GEOS-5 Nature Run (G5NR here-after) is a 2-year global, non-hydrostatic simulation from June 2005 to May 2007 at a 0.0625^o grid-resolution ($\sim$ 7 km near the equator). Not just a simulation of standard meteorological parameters (wind, temperature, moisture, surface pressure), G5NR includes tracers for common aerosol species (dust, seasalt, sulfate, black and organic carbon) and several trace gases: O_3, CO and CO_2. The simulation is driven by prescribed sea-surface temperature and sea-ice, daily volcanic and biomass burning emissions, as well as monthly high-resolution inventories of anthropogenic sources (Putman et al., 2014). As it is a nature run (i.e. no meteorological nudging), the meteorology in G5NR can deviate substantially from the actual weather in 2006.

Aerosol in GEOS-5 are calculated using the Goddard Chemistry, Aerosol, Radiation, and Transport (GOCART) module (Chin et al., 2002) that uses 15 tracers to describe externally mixed species of organic carbon, black carbon, sulphate, sea-salt and dust. Biomass burning emissions are obtained from QFED (Quick Fire Emissions Dataset) (Suarez et al., 2013) with a diurnal cycle imposed online. Anthropogenic emissions of organic and black carbon use EDGAR-HTAP (Emissions Database for Global Atmospheric Research-Hemispheric Transport of Air Pollution) emissions (Janssens-maenhout et al., 2012) which were rescaled to match AEROCOM Phase II emissions. Non-shipping anthropogenic SO_2 emissions come from EDGAR v4.1.

Evaluation of G5NR (Gelaro et al., 2015) against NASA/GMAO MERRA (Modern-Era Retrospective analysis for Research and Applications) Aerosol Reanalysis (da Silva et al., 2012) suggest that global organic carbon, black carbon and sulphate AOT are underestimated by $30-40\%$ while dust AOT is overestimated by $\sim50\%$. Global sea-salt AOT is similar to MERRA within 10%. ~~(Note that~~ Hence, Castellanos et al. (2019) derived global rescaling factors for aerosol speciated ~~AOD~~ AOT in G5NR ~~. How such scaling~~

~~factors will affect AAOD is unknown. True scaling factors~~ but these are not used in the current study (true scaling factors are unlikely to be global, ~~and representation errors in this paper are relative anyway. In this paper the original, i.e. not rescaled, model data will be used~~it is unclear what to do about AAOT and the focus here is on relative errors anyway). Comparison with AEROCOM models shows that G5NR sulphate life-times are quite low (at 2.7 days) while the other species fairly agree with the AEROCOM multi-model mean. ~~Clouds in~~ G5NR ~~show~~ shows reasonable cloud fractions compared to CERES-SSF (Clouds and the Earth's Radiant Energy System-Single Scanner Footprints), although in the equatorial/sub-tropical region (30S-30N), G5NR has a deficit of partially cloudy scenes. In addition there are too few clouds off western continental coasts and the southern branch of the ITCZ is too strong. CALIOP (Cloud-Aerosol Lidar with Orthogonal Polarization) data suggests G5NR cloud fraction are too low, especially over equatorial/sub-tropical lands in the Northern Hemisphere, and too high in the northern polar region.

For this study, the following hourly G5NR data for 2006 were obtained: see Table 1.

Table 1. G5NR data used in this study

short name	description
totexttau	aerosol total column extinction at 550 nm
totscatau	aerosol total column scattering at 550 nm
swtdn	TOA* downward short-wave radiation
cldtot	total cloud area fraction
phis	surface geopotential height
bceman	monthly anthropogenic burning BC emissions
bcembb	monthly biomass burning BC emissions

*: Top Of Atmosphere

2.2 AERONET observations & geolocations

AERONET data were obtained from `https://aeronet.gsfc.nasa.gov`. For 2006, AOT from Direct Sun Version 3 L2.0 and AOT & AAOT from Inversion Version 2 L1.5 and L2.0 were logarithmically interpolated to values at 550 nm and averaged over an hour. For all years starting in 1992, geolocation data were obtained for all sites (1144 in total).

The DirectSun dataset contains only AOT (at multiple wavelengths). These observations are based on direct transmission measurements of solar light and have high accuracy of ± 0.01 (Eck et al., 1999; Schmid et al., 1999). The Inversion dataset contains both AOT and AAOT (at multiple wavelengths) and these observations are based on measurements of scattered solar light from multiple directions. This inversion uses radiative transfer calculations (Dubovik and King, 2000) and yields larger errors than the DirectSun measurements. In particular, Dubovik et al. (2000) showed that Single Scattering Albedo (SSA) errors decrease with increasing AOT and estimated SSA errors of ± 0.03 for water-soluble aerosol at AOT at 440 nm ≥ 0.2 although for dust and biomass burning aerosol higher AOT at 440 nm ≥ 0.5 were needed. Consequently, one important distinction between Inversion L1.5 and L2.0 data is a minimum threshold of AOT at 440 nm ≥ 0.4 used in the latter (improved cloud screening is another distinction). Inversion L2.0 is a subset of the L1.5 dataset.

In the current study, only AOT at 550 nm is used and the Inversion L2.0 AOT at 440 nm criterion is adapted to AOT at 550 nm ≥ 0.25. This is the minimum value of AOT at 550 nm present in actual Inversion L2.0 data, but also corresponds to AOT at 440 nm = 0.4 for small particles (Ångström exponent = 2.1). As a result, the OSSE in this paper is rather lenient when it comes to selecting valid observations similar to Inversion L2.0.

2.3 GAW geolocations

GAW geolocation data were obtained from NILU (Norwegian Institute for Air Research). Two networks were used: the GAW-AOT network which comprises 29 sun-tracking ~~sphotometers~~ sun photometers that measure AOT; and the GAW-ABS network which comprises 81 ~~surface-based filter instruments . While~~ filter instruments that measure surface properties. The real GAW-ABS network is not capable ~~to provide~~ of measuring a columnar (A)AOT ~~measurements, but~~ here we will assume ~~a potential remotely sensed columnar product~~it does, similar to AERONET~~(A)AOT~~, and consider its representation errors.

3 Method: analysis of representation errors

The representation error is defined as the difference between a perfect observation (i.e. no observational error) and a truth value (area average), see also S16a and S17. Here, a self-consistent high resolution simulation will be used to generate both observation and truth ~~(~~in a so-called ~~OSSE), as was first described in S16b and extended in S17~~Observing Systems Simulation Experiment. The representation error may refer to instantaneous values or time averages. This work concerns itself mostly with yearly averages (and some monthly averages). For instantaneous and daily error values, see ~~S16b~~ S16a and S17. The mapping from G5NR data to the data used in this study is given in Table 2.

Perfect observations are generated from the high-resolution simulation by choosing the data at the location of an AERONET or GAW site and ~~sub-sample~~ sub-sampling those data in time according to certain conditions for solar zenith angles (SZA), cloud-fraction and AOT. Table 3 lists the threshold conditions for which observations will

Table 2. Mapping from G5NR data to data used in this study

G5NR	this study	units
totexttau	AOT	
totextau-totscatau	AAOT	
$\frac{180}{\pi}\arccos(\mathrm{swtdn}/1367)$	SZA	degrees
cldtot	cloud fraction	
phis/9.81	geopotential altitude	m
bceman+bcembb	BC emissions	kg/m^2 s

be possible. Values for SZA and AOT are inferred from real AERONET data files. The maximum cloud-fraction was ~~slightly~~ tuned to obtain similar temporal coverage of observations as real AERONET data (see Sect. 4 and Fig. 3 but the impact of tuning is small).

Table 3. Conditions for valid AERONET observations as simulated in this study

source	maximum SZA	maximum cloud-fraction	minimum AOT
DirectSun L2.0	80^o	0.01	0.0
Inversion L1.5	80^o	0.01	0.03
Inversion L2.0	80^o	0.01	0.25

The truth is generated from the high-resolution simulation by averaging AOT and AAOT over a large area (0.5^o to 4^o grid-boxes) and further averaging in time. Here we should distinguish three different protocols depending on how one intends to use the observations, see Table 4. In the case of a gridded climatology derived from observations, the truth should be an average over a continuous long-term time range (say a year). In the case of model evaluation, it is possible to resample model data to the times of the observations. E.g. within the AEROCOM community, a daily collocation protocol is often used, where daily model data is used for days with observations only (irrespective of the temporal sampling of those observations throughout the day). To assess representation errors in this case, the truth needs to be sampled accordingly to days with observations before yearly averages are determined. The same protocols were also explored in S17.

The current methodology differs slightly from S17 in that:

1. a different model is used to construct the OSSE,
2. previously, SZA was assumed to be sufficiently high for a fixed fraction of the day (10 hours). In the current work, SZA is calculated from downward-welling TOA SW radiation and will vary with geo-location and time-of-day,

Table 4. Collocation protocols

collocation protocol	purpose
yearly	gridded climatology
daily	model evaluation (AEROCOM)
hourly	model evaluation

3. previously, the truth was generated for grid-boxes centered on the observations. In the current work, those grid-boxes are assumed regularly spaced from 0^o to 360^o longitude and -90^o to 90^o latitude. The AERONET and GAW sites can be located anywhere within those grid-boxes (at their real geo-location),
4. previously, the high-resolution simulation had a constant grid-size of (about) 10 km. In the current work, the grid-size varies but has a constant angular size of 0.0625^o ($\sim$ 7 km at the equator).

The last point implies that the simulation grid-box used for the observation decreases towards zero as we approach the poles. Since this is clearly undesirable (field-of-view will remain on the order of several kilometers), we will limit our analysis of *representation errors* to latitudes below 60^o. The exceptions are the Figures 5 and S4.

Our methodology allows separation of the factors that determine the representation error: spatial extent of the grid-box, and observational intermittency due to low SZA, high cloud-fraction or low AOT. We will not present such *causal analysis* in this paper (see S17 instead) but will refer to it to explain results.

3.1 ~~Statistical parameters~~

To show the distributions of representation errors, box-whisker plots using the 2, 9, 25, 75, 91 and 98% quantiles will be used in this paper. For a normal distribution, these quantiles will be equally spaced. Any skewness or extended wings in a distribution will be readily visible. In addition to quantiles, the mean error and the mean sign-less error will be provided. The mean sign-less error ~~(or mean absolute error)~~ is deemed more relevant than the standard deviation as 1) it includes biases; 2) the errors are seldom normally distributed, and a standard deviation is very sensitive to larger errors ("out-liers"). For a normal distribution with a mean of zero and a standard deviation of one, the mean sign-less error is ~ 0.8. The correlation used in this paper is the Pearson correlation coefficient that assesses linear relationships. Regression slopes were calculated with a robust Ordinary Least Squares regressor (OLS bisector from the IDL `sixlin` function, Isobe et al. (1990)). This regressor is recommended when

there is no proper understanding of the errors in the independent variable, see also Pitkänen et al. (2016).

4 Evaluation of G5NR and OSSE

In this section, G5NR is evaluated with real AERONET observations of AOT and AAOT, with special focus on its usefulness in an OSSE. As G5NR generates its own meteorology that deviates from 2006, one might expect differences between simulation and observations. Simulated data were nevertheless collocated to the time of the observations (within the hour) to ensure the same temporal sampling throughout the days, the months and the year.

The mean and standard deviation ~~in~~ of AOT and AAOT per site are shown in Fig. 1, top row. In general, simulated site-mean AOT shows good agreement with the observations with correlations around 0.75 and slopes around 0.84. Simulated site-mean AAOT does not agree as nicely with the observations but there is still correlation (~~~ 0.45~~0.48) (the evaluation of AAOT will ~~of course~~ be affected by large measurement errors). The agreement in standard deviation suggests that simulated and observed AOT and AAOT show similar temporal variation. But the global agreement also suggests that the simulation captures spatial variation rather well. This is also true on shorter length scales, as an analysis by region shows in Table 5. Europe appears to be the exception but this is mostly due to a few southern sites. ~~Removing them from the analysis, significantly increases correlation~~ As the table shows: without those sites, correlation increases significantly. This may be related to the overestimation of dust and underestimation of carbonaceous & sulphate aerosol in G5NR (Gelaro et al., 2015), which will affect north-south gradients in AOT in Europe. DRAGON (Distributed Regional Aerosol Gridded Observation Networks, see Holben et al. (2018)) campaigns might allow evaluation of the spatial distribution of simulated AOT at even smaller length-scales (10's of kilometers) but ~~they did not start until 2010.~~ are not available for 2006.

Table 5. Correlation in modelled and observed yearly ~~site~~ site-mean AOT

region	nr	correlation
World	216	0.75
Europe	55	0.26
Europe*	26	0.68
Africa	32	0.86
Asia	34	0.82
N. America	49	0.81
S. America	13	0.91

*: southern AERONET sites removed from analysis

The top row of Fig. 1 was created using only sites that provide a minimum of 100 real observation throughout 2006. The lower row shows how this ~~criterium~~ criterion affects results. As the minimum number of observations per site increases, so do the correlations, probably due to a reduction in statistical noise (partly due to different simulated and actual meteorologies). But the overall bias also increases. This ~~criterium~~ criterion selects for sites with lower cloudiness (higher number of observations) until predominantly northern African and Saudi Arabian sites are left for a minimum of 500 observations per site. The increase in bias is thus likely due to the overestimation of dust AOT that was mentioned earlier.

Note that AAOT is here evaluated with L1.5 data. The L2.0 data have a minimum AOT threshold ~~of ~ 0.25~~ which results in ~~less observations and less~~ fewer observations and fewer available sites overall. Although L1.5 is considered a less reliable product, the evaluation with L2.0 (which now uses a minimum of 30 observations per site) yields a similar but slightly poorer result for G5NR, see Fig. ~~??~~S1, and over a shorter range of values.

Figure 2 shows mean values per site for the daily difference in maximum and minimum AOT. Again, good agreement for simulated AOT is seen but AAOT compares rather poorly. However, ~~it's~~ its correlation is still above 0.6 and it is clear that the simulation *underestimates* daily AAOT variation. ~~AAOT measurement errors are not expected to have a big impact~~ The impact of AAOT measurement error on daily variation ~~(which is the~~ is likely reduced as the variation is a difference between two measurements (pers. comm. with T. Eck and O. Dubovik).

Figure 3 ~~evaluates the OSSE and shows~~ shows the temporal coverage (or frequency of observation) per site as a function of latitude. G5NR's simulated coverage is calculated using the ~~limitations~~ conditions described in Table 3 (and explained ~~later~~ in Sec. 3). This coverage would be 100% if observations are available 24 hours a day, 365 days a year. In practice it cannot be higher than 50% due to the day-night cycle, and will be less due to cloudiness or low AOT.

The bimodal structure that is visible in both the simulation and observations is due to SZA variation (which reduces coverage towards the poles) and cloudiness (which reduces coverage near the equator). Simulated and real coverage per site are not expected to agree well due to meteorological differences and down times from site maintenance. Still, the results suggests that the OSSE predicts similar frequency of Direct Sun observations as actually observed.

However, the OSSE also simulates more Inversion observations in the Northern hemisphere than actually occur. This suggests there are ~~limiting~~ additional factors in observational coverage that are not accounted for in Table 3. One factor is that real Inversion measurements are ~~simply~~ attempted less frequently (several times per day) than Direct Sun measurements (several times per hour). Other factors may include inversion failure at low SZA (real observations show that In-

version data generally have larger SZA than DirectSun data even though Inversion data is generally closer to the equator) and overestimation of dust AOT in G5NR (largest overestimates of coverage occur for Sahara ~~nd~~ and Saudi Arabia sites). Yet another issue is that successful inversion requires a high degree of azimuthal symmetry in the measurements. In essence, this is a built-in check on the magnitude of *spatial* representation errors which will lower temporal coverage of the observations but is not considered in the OSSE due to lack of information. Finally, instrument malfunction & maintenance are not taken in to account~~, which will explain some of the discrepancy~~.

In all, it seems that G5NR can realistically simulate spatial and temporal variation in AOT and AAOT, ~~although there~~ at least on the scales accessible by the available observations. There is some underestimation of daily AOT variation and significant underestimation of daily AAOT variation. G5NR can also be used to fairly realistically simulate frequency of observation (temporal coverage), although it will overestimate this for the Inversion products in the Northern Hemisphere. ~~Further evidence for G5NR's applicability in an OSSE is given in Fig. 9 where it is shown that the present study agrees with an earlier analysis by Kinne et al. (2013) on the most representative AERONET sites.~~

5 Results

5.1 Representation errors in yearly AOT

Figure 4 shows yearly representation errors for AERONET DirectSun L2.0 AOT observations ~~, for three different~~ as a function of model grid-box size, for the three collocation protocols (see Table 4)~~, as a function of model grid-box size. Hourly collocation yields the smallest representation errors, and this is more pronounced for smaller grid-box sizes.~~ . As grid-box size changes from 4^o to 0.5^o, ~~hourly collocation errors~~ errors for hourly collocation are more than halved from 13% to 5% while those for daily collocation change only from 17% to 12%. ~~By construction, hourly collocation errors become zero for a grid-box size equal to 0.0625^o (the resolution of G5NR).~~ In contrast, ~~yearly collocation errors~~ errors for yearly collocation ($\sim 22\%$) are dominated by temporal sampling and do not depend much on grid-box size. Smaller representation errors for hourly collocation can also be seen in a regional analysis, see Fig. S2. The hourly collocation is especially beneficial when using the Inversion L2.0 AOT product, which allows large representation errors due to the condition of a minimum AOT at 440 nm ($>= 0.4$) for valid observations, see Fig. S3, even though it results in a global 9% bias.

The impact of collocation protocol can also be shown through the total number of sites that yield errors larger than, say, 10%: 821 (yearly), 653 (daily), 235 (hourly) out of 1108 AERONET stations in total, for a grid-box of $1^o \times 1^o$. ~~Also~~ In addition to larger representation errors in general, the yearly and daily collocation protocols ~~yield significant bias~~ also allow significant biases across the AERONET network. Regionally~~that bias translates into~~ , spatial patterns with east-west or north-south gradients in the representation errors exist, see Fig. 5 and ~~??~~Fig. S4. Such patterns are absent or at least much reduced for hourly collocation. ~~Representation errors for different regions are shown in~~

The biases in regional and global distributions of representation errors for yearly and daily collocations are strongly affected by cloudiness. Higher humidity in the cloudy part of a grid-box increases AOT through wet growth The area averages used to calculate representation errors have been derived for the *entire* grid-box (all-sky), both clear and cloudy parts. Representation errors for clear-sky parts of grid-boxes are lower for the yearly and daily collocation protocols, see Fig. **??**. 6. In certain situations, it seems more realistic to use only the clear part of the grid-box in calculating representation errors: e.g. when the grid-box average stands in for an aggregated satellite product. In this paper, focus will be on the all-sky representation error.

Table 6 shows absolute values of the yearly representation errors for different collocation protocols (yearly and hourly) and grid-box sizes. The statistical metrics provided are the mean of the sign-less representation error over all AERONET sites, and the 90% quantile of the sign-less representation error (an indication of the large representation errors possible for some sites). Using absolute values allows a comparison with the AERONET AOT measurement error of 0.01 (Eck et al., 1999; Schmid et al., 1999). This is the error for individual measurements, and not that of a yearly average which is likely to be much smaller. Clearly, representation errors are larger than measurement errors.

Results so far suggest that the daily collocation is a significant improvement ~~from~~ over the yearly collocation. This is in contrast to S17 (Fig. 7) where the representation errors for daily and *monthly* collocation were found to be similar. ~~Further analysis of the data suggests that the~~ The absence of diurnal (anthropogenic) emission profiles in G5NR may cause underestimation of ~~daily collocation errors .~~ representation errors for the daily collocation in the current study.

~~Representation errors for AOT do not differ much for the Direct Sun L2.0 and Inversion L1.5 products, see Fig. **??**. However, the condition of a minimal AOT ($>= 0.25$) for valid observations causes large but unsurprising errors for the Inversion L2.0 product. This issue with the Inversion L2.0 data is well-known but the current analysis may be the first realistic estimate of incurred errors . Figure **??** also shows results for two sensitivity studies where observational coverage in the Northern Hemisphere was artificially lowered (see discussion in last paragraph of Sect. 4) but this has no clear impact as temporal coverage is quite low anyway.~~

Table 6. Absolute representation errors for AERONET sites

metric	protocol	4^o	2^o	1^o	0.5^o	0.0625^o
mean	yearly	0.043	0.042	0.042	0.042	0.044
	hourly	0.021	0.015	0.011	0.008	0.000
90 %	yearly	0.086	0.079	0.082	0.083	0.086
	hourly	0.052	0.033	0.029	0.017	0.000

It is interesting to compare the representation errors of two different networks, AERONET and GAW. AERONET was not designed with representativity in mind but the GAW network was. Nevertheless, Fig. 7 suggests that GAW sites exhibit slightly larger representation errors than AERONET. In particular, GAW error statistics are strongly skewed to negative values. In the G5NR OSSE, GAW sites are located at higher altitudes and more often on isolated mountains than AERONET sites ~~. A~~ (G5NR site altitudes correlate very well with real altitudes, $R = 0.98$, but tend to underestimate by 28 m on average, with a standard deviation of 171 m). A look at yearly representation errors for the hourly collocation reveals a systematic altitude dependence, see Fig. 8. A high altitude site on an isolated mountain will observe a shorter atmospheric column than the surrounding grid-box ~~which~~ (most of which is at lower altitudes) which will cause a negative representation error~~, see Fig. 8. Actual site altitudes vary from -410 to 5320 m. G5NR site altitudes correlate very well ($r = 0.98$) but tend to underestimate by 28 m on average, with a random error of 171 m.~~. Note that AERONET sites do not show this dependence on altitude for 1^o grid-boxes, probably because they are located more often on mountains surrounded by similar mountaineous terrain.

~~Previous work by Kinne et al. (2013) ranked AERONET~~ Finally, a comparison is made with a previous study into AERONET representation errors (Kinne et al., 2013). Using a range score r, see Table 7, they ranked sites according to their representativity ~~, see Table 7~~ for larger domains. This ranking is subjective in that it is non-quantitative, based on personal knowledge of the sites and only defines representativity in broad terms. The ~~ranking is~~ range scores are only available for sites that had at least 5 months of data before 2008. Using the methodology of this paper, representation errors were calculated for all sites of a certain ~~ranking~~range score, see Fig. 9. For large grid boxes of 4^o (~ 450 km near equator)~~)~~, the impact of ~~ranking~~ the range score on representation error is quite small. While there is a visually arresting change in the error distribution for $r > 1$ (wide flanks are changed into a broader center), the mean sign-less error barely changes. This rather weak dependence on range score suggests that Kinne et al. (2013) overestimated the size of the domains ($>= 500$ km for $r > 1$) for which their sites were representative. On the other hand, for a grid-box of 1^o a substantial reduction in representation error can be seen for $r > 1$ sites. However, this only occurs for the hourly collocation: Kinne et al. (2013) did not consider the temporal sampling of the observations which causes large representation errors. ~~A new~~ An alternative ranking of representativity will be introduced in Sect. 6.

~~Sofar the area averages used to calculate representation errors have been derived for the entiregrid-box, both the clear and cloudy parts. Under certain circumstances, it may be more realistic to use only the clear part. Examples are the evaluation of aggregated satellite products with AERONET (like AERONET, satellites can not observed aerosol when there are clouds), or the evaluation of certain models that explicitly calculate clear-sky AOT (usually by estimating clear-sky humidity from grid-box averaged humidity). Representation errors for clear-sky parts of grid-boxes are improved for the yearly and daily collocation protocols, see Fig. 6.~~

5.2 Representation errors ~~at~~ in monthly ~~time-scales~~AOT

Surprisingly, monthly representation errors are not that much larger than yearly errors, see Fig.10. If monthly errors for the same site were independent and random, one would expect them to be $\sim \sqrt{12} \approx 3.5$ larger than yearly errors but that is ~~clearly~~ not the case. As a matter of fact, monthly errors are strongly correlated from month to month, throughout the year, see Fig. 11. The increase in correlation with January after September, is probably due to yearly cycles in meteorology and emissions and very likely to be a realistic aspect of representation errors. The implication of this is that multi-year averages may not reduce representation errors as strongly as one would hope.

This analysis also provokes the question whether representation errors (per site) should be seen as mostly biases or random errors ~~with strong correlations~~ (see also Schwarz et al. (2018)). ~~Our preliminary~~ Preliminary analysis suggests that ~~at the monthly scale,~~ both cases can occur. ~~Figure ?? shows both maximum and minimum monthly errors by site as a function of yearly error.Many~~ I.e. some sites show large variations ~~in monthly representation errors, but significantly reduced yearly errors, suggesting that the~~

Table 7. Range scores for AERONET sites in (Kinne et al., 2013)

range score	spatial domain	number of sites	comments
0	100 km	120	includes mountainous sites
1	300 km	106	
2	500 km	28	
3	900 km	6	

errors are essentially random. However, some sites also show very similar monthly maxima and minima, and yearly errors, suggesting that these errors are better interpreted as biases. (including sign changes) in representation error from month to month, and as a consequence a strongly reduced yearly representation error. Here monthly representation errors may be interpreted as mostly random. Other sites show monthly representation errors with not much variation and as a consequence yearly representation errors are similar to the monthly errors. There the representation error is better characterised as a bias. A proper analysis of this would require significantly longer time-series of data than are currently available. Further discusion of this can be found in Sect. 6.

5.3 Representation errors in AAOT

Representation The discussion of representation errors for Inversion L1.5 AAOT product are shown in will be shorter than that for DirectSun L2.0 AOT, as the main conclusion is identical: the hourly collocation yields smaller representation errors than the other protocols, see Fig. 12. As for AOT, representation errors decrease with decreasing grid-box sizes, although the decrease is small for for yearly collocation. Sgnificant positive biases can be seen for all protocols and large grid-box sizes. These biases are partly due to the AOT > 0.25 criterium for valid observations, which translates into an AAOT $= (1 - \mathrm{SSA})\mathrm{AOT} > 0.025$ criterium for SSA = 0.9. However, other reasons for Note also that representation errors in AAOT are of a similar magnitude as for AOT. One obvious difference is that AAOT representation errors tend to be positively biased while the AOT errors were negatively biased. While the latter was due to cloudiness as discussed before, the positive bias are the proximity of AERONET sitesto sources of absorbing aerosol and the impact of orography (e.g. see Schutgens et al. (2017) and Sect. 6. Unsurprisingly the hourly collocation protocol shows the smallest positive bias and reduces it faster for decreasing grid-box size.for AAOT is more difficult to explain. It appears that a combination of conditions (location of the sites, necessity of day-light, clear skies and a minimum AOT of 0.03) together conspire to create these positive biases. Only over the Amazon can a simple explanation be found: the clear sky condition prevents many observations outside the biomass burning season, explaining large positive biases for yearly collocation (see also Fig. S5, discussed later).

Strikingly, daily collocation yields very similar errors as hourly collocation. This is very Even more than for AOT, representation errors for AAOT are very similar for the daily and hourly collocations. As discussed before, this is likely due to a limitation in the OSSEthe absence of diurnal (anthropogenic) emissions profiles. The daily variation of AAOT is strongly underestimated by G5NR (see Sect 4 and Fig. 2), possibly due to an absence of diurnal anthropogenic emission profiles.

Regionally, there is some variation in representation errors but not a lot, see For completeness' sake, an analysis of AAOT representation errors for different regions (Fig. S5), different products (Fig. **??**.The exception is for the yearly collocation protocol which allows significant biases for sites in South America and Africa.This is related to the AOT criterium for valid observations and the dominant influence of episodic biomass burning for these two continents: outside the burning season much less observations are made. Consequently the observations will favour the absorbing biomass burning aerosol.S6), different networks (Fig. S7) and different range scores by Kinne et al. (Fig. S8) are given in the supplement. Overall the conclusions are very similar to those for AOT.

A comparison between AERONET and GAW, Fig. **??**, shows error distributions that are positively skewed for AERONET and negatively skewed for GAW-AOT. The smaller bias for GAW than AERONET The similarity in general behaviour of representation errors for AOT and AAOT should not be taken to mean that these errors are identical per site. As discussed in Sect. 6, f representation errors for AOT and AAOT or individual sites can be very different. Ultimately this is due to a balancing of the positive bias due to the AOT criterium for valid observations and the negative bias due to site altitude (see also Fig 7 and its discussion).

An analysis of the impact of the site rankings by Kinne et al. (2013), shows similar results for AAOT as for AOT, see Fig **??**the different sources of AOT and AAOT which leads to different spatio-temporal distributions in the atmosphere.

5.4 Comparison to recent results from Wang et al. '18

Recently Wang et al. (2018) suggested that the observed underestimation of AAOT by AEROCOM models (Bond et al., 2013) may be due to *spatial* representation errors. Their analysis found that AERONET Inversion L1.5 AAOT representation errors exhibit a global bias of 30% for $2^o \times 2^o$ model grid-boxes, which would help explain the aforementioned underestimation by the global models. As AERONET sites need to be serviceable, they are often found near roads and urban build-up, i.e. near sources of absorbing aerosol. Compared to the larger area of global model grid-boxes, these sites would quite naturally observe larger AAOT. Thus, Wang et al. (2018) concluded that at least part of the underestimation of modelled AAOT is an artefact, created by the location of the AERONET sites.

Wang et al.'s idea is quite persuasive and indeed one can see evidence of such positive representation errors in Fig. 13 where sites in major cities like London, Paris, Madrid and Barcelona clearly exhibit positive representation errors. (For another example, see Fig. 3b in S17 concerning surface black carbon concentrations). But Wang's study found such biases for the majority of AERONET sites, not just a few located in big cities. As a matter of fact, the current study shows no evidence of this global bias of 30%. Instead it finds a global bias of only 9%, dominated by a few sites with large positive representation errors (median bias over all sites: 4%).

Wang et al. (2018) performed an analysis very much like the one in this study with one crucial difference. As they did not have a global simulation at high resolution like G5NR, they downscaled results from a standard global simulation at $2.5^o \times 1.27^o$ resolution. The downscaling was accomplished with the help of a high-resolution ($0.1^o \times 0.1^o$) black carbon emission map (Wang et al., 2016). It is possible to simulate this procedure using the high-resolution G5NR black carbon emission maps and AAOT simulations (the AAOT simulation was first coarsened over $2^o \times 2^o$) and explain the different results in Wang et al. (2018) and the current study.

Figure 14 shows AAOT *spatial* representation errors as estimated by the current study and by Wang's methodology as simulated with G5NR data. A global bias of 25%, not very different from the original 30% mentioned in Wang et al. (2018), is found for the Wang analysis whose representation errors yield a strongly skewed distribution over all sites. In contrast, the present study yields a more symmetric distribution with a much smaller bias. Unlike in the Wang analysis this bias is dominated by just a few sites with large positive representation errors.

The analysis above is a self-consistent evaluation of Wang's methodology. Using high-resolution black carbon emission data to downscale coarse model AAOT fields ignores redistribution of absorbing aerosol due to small scale (at and below the coarse model's grid-box) advective and turbulent transport as well as removal by local precipitation (Wang et al. were aware of this limitation but could not assess its impact). It also ignores local orography and the contribution of absorbing dust to AAOT. The result is that there is very little correlation between representation errors as estimated by the two methods, see Fig. 15. As a matter of fact, representation errors from the current study do not show a systematic dependence on emission distributions, unlike the representation errors from Wang's methodology.

6 A ranking of representativity for the AERONET sites

A ranking of AERONET and GAW sites in terms of their *spatial* representativity for AOT and AAOT can be found at Schutgens (2019). Only sites below 60^o latitude are considered, and temporal sampling of observations is ignored. The latter was done for two reasons: 1) as discussed in Sect 2 and 4, temporal sampling of observations is considered less accurately modelled by the OSSE than spatial variability; 2) both S17 and the current study show that once hourly collocation is used, the remaining representation error is similar although slightly larger than the spatial representation error.

Relative representation errors are classed according to ~~bin boundaries 0%~~ bins: 0-5% (rank 1), 5-10% (rank 2), 10-20% (rank 3), 20-40% (rank 4), ~~5% , 10% , 20% ,~~ 40% and ~~up. Using a~~ higher (rank 5). The accuracy of this ranking depends of course on the skill of G5NR and the OSSE, but also on statistical noise due to the use of a single year of data. The latter source of uncertainty was assessed using a block bootstrap method ~~on~~ (Efron, 1979) on the time-series per site~~, the uncertainty in yearly representation errors was assessed~~. Typically more than 85% of all resampled time-series yield a representation error in the same class as the original time-series. For large grid-boxes (4^o) and small errors ($< 10\%$), this may drop down to 66% of the resampled time-series. ~~In any case,~~ For those resampled time-series that yielded a different ranking, this ranking was only off by 1. It then seems that statistical noise does not prevent a robust classification of yearly relative spatial representation errors~~can be classed robustly. Of course,~~. The impact of G5NR and ~~the OSSE are not perfect, which will introduce an uncertainty into the ranking that can not currently~~ OSSE skill on the classification can currently not be assessed.

Compared to the subjective ranking by Kinne et al. (2013), the new ranking is objective because the rank is related to a well-defined representation error that is quantified bottom-up from known emission sources and calculated meteorology. That in itself is of course no guarantee for accuracy.

Inspection of the rankings turns up several interesting points. Analysis in the previous sections determined a few "rules" for the behaviour of representation errors (e.g. errors decrease ~~as does~~ when the grid-box size decreases) but these can easily be "broken" for specific sites: a smaller grid-box may actually lead to larger representation errors (e.g. AOE_Baotou, Ascension_Island, Aras_de_los_Olmos), monthly errors may be substantially larger than yearly errors

(e.g. ARM-Darwin, BORDEAUX). Also, representation errors for AOT and AAOT may be very different: Bayfordbury shows small yearly representation errors for AOT but large errors for AAOT, while Mace_Head shows the opposite.

7 Conclusions

Remote sensing observations from the AERONET and GAW networks are intermittent in time and have a limited field-of-view. Consequently such observations have limited ability to represent ~~AOT or AAOT~~ (Absorbing) Aerosol Optical Thickness, or (A)AOT, over larger areas. The resulting spatio-temporal *representation* error is here analysed using a high-resolution simulation of global aerosol (GEOS5 Nature Run, ~ 7 km resolution near equator). Using G5NR, an ~~OSSE~~ Observing System Simulation Experiment (OSSE) was constructed that simulates the frequency of AERONET observations taking ~~SZA~~Solar Zenith Angle, cloud fraction and AOT values into account.

This work extends previous work on temporal representation with global low-resolution models (Schutgens et al., 2016b) to spatio-temporal representation. It also extends previous work on spatio-temporal representation with regional high-resolution simulations (Schutgens et al., 2016a, 2017) to the global domain. The current work is more limited in scope than the previous studies and only considers ground-based remote sensing observations. For satellite remote sensing, see Schutgens et al. (2016b) and Schutgens et al. (2017). For in-situ measurements, see Schutgens et al. (2016a) ~~(and Schutgens et al. (2017))~~and Schutgens et al. (2017).

G5NR and the OSSE are evaluated and found to show significant skill in AOT and reasonable skill in AAOT. AERONET mean AOT per site, as well as yearly and daily variability were estimated ~~correctly~~ quite correctly, usually within a factor less than $2\times$. Considering that G5NR generates its own meteorology, G5NR AOT correlated very well ($~~r \approx 0.75~~ R \approx 0.75$) with the observations. Similarly, the OSSE was surprisingly good at simulating the overall pattern of observational coverage (frequency of AOT observation). Results were not as good for AAOT but still ~~impressive~~acceptable. Yearly AAOT variability was slightly underestimated while daily AAOT variability was severely underestimated. The latter is possibly related to the absence of diurnal anthropogenic emission profiles in G5NR. For representativity studies that take diurnal variations into account, see Schutgens et al. (2016a, 2017). In addition, the OSSE tended to overestimate the frequency of AAOT observations per site (although this was shown to have no impact on representation errors).

Both yearly and monthly representation errors are provided for observations from ground sites that attempt to represent larger areas (from 0.5^o to 4^o in size). The monthly representation errors are shown to be strongly correlated throughout the year. For some sites this is an expression of a bias but that is not universally the case. In any case, monthly representation errors can not be treated as independent and this has (negative) consequences for the reduction of representation errors in multi-year averages. Other conclusions are: 1) AERONET derived climatologies allow for substantial representation errors (yearly collocation allows errors of typically 20%~~globally~~, see Fig. 4); 2) AEROCOM evaluation protocol is sub-optimal (daily collocation ~~allows errors of typically~~ can show errors of 25% in coherent regional patterns). Instead hourly collocation ~~was~~ is advocated. Also, the representativity of AERONET and GAW sites was shown to be not very different, although AERONET sites seem to be more affected by nearby sources while GAW sites seem more affected by their altitude. Finally, a subjective ranking (Kinne et al., 2013) of the *spatial* representativity of sites was analysed and shown to broadly agree with the current study, although it appears to overestimate represented spatial domain sizes and judges several sites as less representative than the current analysis. A new objective ranking is also presented.

Spatial representation errors have been used to reconcile observations and global simulations of AAOT. Bond et al. (2013) showed that global models tend to significantly underestimate AAOT but Wang et al. (2018) suggested that AERONET AAOT observations may suffer from a global 30% representation bias. In contrast, the current analysis finds a much smaller bias of 9% which is more-over strongly influenced by a few sites with large positive representation errors due to their proximity to black carbon sources. Judiciously excluding those sites significantly reduces the bias even further (4%). The large positive representation errors found by Wang et al. are shown to be due to methodological choices that limit the realism of their OSSE.

Several questions remain and seem interesting for follow-up studies: 1) how can we evaluate the representativity rankings?; 2) how do OSSE errors affect estimated representation errors?; 3) how will diurnal emission profiles impact results?; 4) can representation errors at any site be decomposed in a bias and random error (possibly with temporal correlations over several months)?; 5) what are representation errors like in multi-year averages?

Code and data availability. G5NR data can be obtained from `https://gmao.gsfc.nasa.gov/global_mesoscale/7km-G5NR/data_access`, AERONET data can be obtained from `https://aeronet.gsfc.nasa.gov`. Analysis code was written in IDL and is available from the author upon request.

Author contributions. NS designed the experiments, carried them out and prepared the manuscript.

Competing interests. No competing interests are present

Acknowledgements. NS thanks the NASA Global Modelling and Assimilation Office team that conducted the GEOS-5 Nature Run simulation, in particular Arlindo da Silva and Ravi Govindaraju for help in obtaining the G5NR data. NS thanks the PI(s) and Co-I(s) and their staff for establishing and maintaining the many AERONET sites used in this investigation. NS thanks Ann Mari Fjaeraa (NILU) for providing GAW-AOT and GAW-ABS geolocation data. NS is also grateful to Rong Wang, Björn H. Samset and Gunnar Myhre for valuable discussions. The figures in this paper were prepared using David W. Fanning's Coyote Library for IDL. This work is part of the Vici research programme with project number 016.160.324, which is (partly) financed by the Dutch Research Council (NWO).

References

Anderson, T. E., Charlson, R. J., Winker, D. M., Ogren, J. A., and Holmen, K.: Mesoscale Variations of Tropospheric Aerosols, J. Atmospheric Sciences, 60, 119–136, 2003.

Boersma, K. F., Vinken, G. C. M., and Eskes, H. J.: Representativeness errors in comparing chemistry transport and chemistry climate models with satellite UV-Vis tropospheric column retrievals, Geoscientific Model Development, 9, 875–898, https://doi.org/10.5194/gmd-9-875-2016, 2016.

Bond, T. C., Doherty, S. J., Fahey, D. W., Forster, P. M., Berntsen, T., Deangelo, B. J., Flanner, M. G., Ghan, S., K??rcher, B., Koch, D., Kinne, S., Kondo, Y., Quinn, P. K., Sarofim, M. C., Schultz, M. G., Schulz, M., Venkataraman, C., Zhang, H., Zhang, S., Bellouin, N., Guttikunda, S. K., Hopke, P. K., Jacobson, M. Z., Kaiser, J. W., Klimont, Z., Lohmann, U., Schwarz, J. P., Shindell, D., Storelvmo, T., Warren, S. G., and Zender, C. S.: Bounding the role of black carbon in the climate system: A scientific assessment, Journal of Geophysical Research Atmospheres, 118, 5380–5552, https://doi.org/10.1002/jgrd.50171, 2013.

Bulgin, C. E., Embury, O., and Merchant, C. J.: Sampling uncertainty in gridded sea surface temperature products and Advanced Very High Resolution Radiometer (AVHRR) Global Area Coverage (GAC) data, Remote Sensing of Environment, 177, 287–294, https://doi.org/10.1016/j.rse.2016.02.021, http://dx.doi.org/10.1016/j.rse.2016.02.021, 2016.

Castellanos, P., da Silva, A. M., Darmenov, A. S., Buchard, V., Govindaraju, R. C., Ciren, P., and Kondragunta, S.: A Geostationary Instrument Simulator for Aerosol Observing System Simulation Experiments, Atmosphere, 10, https://doi.org/10.3390/atmos10010002, 2019.

Cavanaugh, N. R. and Shen, S. S. P.: The effects of gridding algorithms on the statistical moments and their trends of daily surface air temperature, Journal of Climate, 28, 9188–9205, https://doi.org/10.1175/JCLI-D-14-00668.1, 2015.

Chin, M., GInoux, P., KInne, S., Torres, O., Holben, B. N., Duncan, B. N., Martin, R. V., Logan, J. A., Higurashi, A., and Nakajima, T.: Tropospheric Aerosol Optical Thickness from the GOCART Model and Comparisons with Satellite and Sun Photometer Measurements, Journal of the Atmospheric Sciences, 59, 461–483, 2002.

Colarco, P., Silva, A., Chin, M., and Diehl, T.: Online simulations of global aerosol distributions in the NASA GEOS - 4 model and comparisons to satellite and ground - based aerosol optical depth, Journal of Geophysical Research : Atmospheres, 115, https://doi.org/10.1029/2009JD012820, 2010.

Colarco, P. R., Kahn, R. A., Remer, L. A., and Levy, R. C.: Impact of satellite viewing-swath width on global and regional aerosol optical thickness statistics and trends, Atmospheric Measurement Techniques, 7, 2313–2335, https://doi.org/10.5194/amt-7-2313-2014, 2014.

Coldewey-Egbers, M., Loyola, D. G., Koukouli, M., Balis, D., Lambert, J. C., Verhoelst, T., Granville, J., Van Roozendael, M., Lerot, C., Spurr, R., Frith, S. M., and Zehner, C.: The GOME-type Total Ozone Essential Climate Variable (GTO-ECV) data record from the ESA Climate Change Initiative, Atmospheric Measurement Techniques, 8, 3923–3940, https://doi.org/10.5194/amt-8-3923-2015, 2015.

da Silva, A., Colarco, P. Darmenov, A., Buchard-Marchant, V., Randles, C., and Govinaradju, R.: Overview of the MERRA Aerosol Reanalysis: Toward an Integrated Earth System Analysis, in: 4th WCRP International Conference on Reanalyses, 2012.

Diedrich, H., Wittchen, F., Preusker, R., and Fischer, J.: Representativeness of total column water vapour retrievals from instruments on polar orbiting satellites, Atmospheric Chemistry and Physics, 16, 8331–8339, https://doi.org/10.5194/acp-16-8331-2016, 2016.

Director, H. and Bornn, L.: Connecting point-level and gridded moments in the analysis of climate data, Journal of Climate, 28, 3496–3510, https://doi.org/10.1175/JCLI-D-14-00571.1, 2015.

Dubovik, O. and King, M. D.: A flexible inversion algorithm for retrieval of aerosol optical properties from Sun and sky radiance measurements, Journal of Geophysical Research, 105, 20 673, https://doi.org/10.1029/2000JD900282, http://doi.wiley.com/10.1029/2000JD900282, 2000.

Dubovik, O., Smirnov, A., Holben, B. N., King, M. D., Kaufman, Y. J., Eck, T. F., and Slutsker, I.: Accuracy assessments of aerosol optical properties retrieved from Aerosol Robotic Network (AERONET) Sun and sky radiance measurements, Journal of Geophysical Research, 105, 9791–9806, https://doi.org/10.1029/2000JD900040, http://doi.wiley.com/10.1029/2000JD900040, 2000.

Eck, T. F., Holben, B. N., Reid, J. S., Smirnov, A., O'Neill, N. T., Slutsker, I., and Kinne, S.: Wavelength dependence of the optical depth of biomass burning, urban, and desert dust aerosols, J. Geophysical Research, 104, 31 333–31 349, 1999.

Efron, B.: Bootstrap methods: another look at the jackknife, The annals of Statistics, 7, 1—-26, 1979.

Gelaro, R., Putman, W. M., Pawson, S., Draper, C., Molod, A., Norris, P. M., Ott, L., Privé, N., Reale, O., Achuthavarier, D., Bosilovich, M., Buchard, V., Chao, W., Coy, L., Cullather, R., Silva, A., Darmenov, A., and Errico, R. M.: Evaluation of the 7-km GEOS-5 Nature Run, Tech. Rep. NASA / TM – 2014-104606, NASA, 2015.

Geogdzhayev, I., Cairns, B., Mishchenko, M. I., Tsigaridis, K., and van Noije, T.: Model-based estimation of sampling-caused uncertainty in aerosol remote sensing for climate research applications, Quarterly Journal of the Royal Meteorological Society, 140, 2353–2363, https://doi.org/10.1002/qj.2305, http://doi.wiley.com/10.1002/qj.2305, 2014.

Hakuba, M. Z., Folini, D., Sanchez-lorenzo, A., and Wild, M.: Spatial representativeness of ground-based solar radiation measurements - extension to the full Meteosat disk,

Journal of Geophysical Research: Atmospheres, 16, 50 673, https://doi.org/10.1002/jgrd.50673., 2014a.

Hakuba, M. Z., Folini, D., and Wild, M.: Solar absorption over Europe from collocated surface and satellite observations, Journal of Geophysical Research: Atmospheres, pp. 3420–3437, https://doi.org/10.1002/2013JD021421, 2014b.

Holben, B. N., Kim, J., Sano, I., Mukai, S., Eck, T. F., Giles, D. M., Schafer, J. S., Sinyuk, A., Slutsker, I., Smirnov, A., Sorokin, M., Anderson, B. E., Che, H., Choi, M., Crawford, J. H., Ferrare, R. A., Garay, M. J., Jeong, U., Kim, M., Kim, W., Knox, N., Li, Z., Lim, H. S., Liu, Y., Maring, H., Nakata, M., Pickering, K. E., Piketh, S., Redemann, J., Reid, J. S., Salinas, S., Seo, S., Tan, F., Tripathi, S. N., Toon, O. B., and Xiao, Q.: An overview of mesoscale aerosol processes , comparisons , and validation studies from DRAGON networks, Atmospheric Chemistry and Physics, 18, 18 655– 18 671, 2018.

Isobe, T., Feigelson, E. D., Akritas, M. G., and Babu, G. J.: Linear regression in Astronomy I, The Astrophysical Journal, 364, 104–113, 1990.

Janssens-maenhout, G., Dentener, F., Aardenne, J. V., Monni, S., Pagliari, V., Orlandini, L., Klimont, Z., Kurokawa, J.-i., Akimoto, H., Ohara, T., Wankmüller, R., Battye, B., Grano, D., Zuber, A., and Keating, T.: EDGAR-HTAP : a harmonized gridded air pollution emission dataset based on national inventories, Tech. rep., JRC, Institute for Environment and Sustainability, https://doi.org/10.2788/14102, 2012.

Kaufman, Y. J., Holben, B. N., Tanre, D., Slutsker, I., Smimov, A., and Eck, T. F.: Will aerosol measurements from Terra and Aqua polar orbiting satellites represent the daily aerosol abundance and properties?, Geophysical Research Letters, 27, 3861–3864, 2000.

Kinne, S., O'Donnel, D., Stier, P., Kloster, S., Zhang, K., Schmidt, H., Rast, S., Giorgetta, M., Eck, T. F., and Stevens, B.: MAC-v1: A new global aerosol climatology for climate studies, Journal of Advances in Modeling Earth Systems, 5, 704–740, https://doi.org/10.1002/jame.20035, http://doi.wiley.com/10.1002/jame.20035, 2013.

Kovacs, T.: Comparing MODIS and AERONET aerosol optical depth at varying separation distances to assess ground-based validation strategies for spaceborne lidar, Journal of Geophysical Research, 111, D24 203, https://doi.org/10.1029/2006JD007349, http://www.agu.org/pubs/crossref/2006/2006JD007349.shtml, 2006.

Levy, R. C., Leptoukh, G. G., Kahn, R., Zubko, V., Gopalan, A., and Remer, L. a.: A critical look at deriving monthly aerosol optical depth from satellite data, IEEE Transactions on Geoscience and Remote Sensing, 47, 2942–2956, https://doi.org/10.1109/TGRS.2009.2013842, 2009.

Lin, M., Horowitz, W., Cooper, O. R., Tarasick, D., Conley, S., Iraci, L. T., Johnson, B., Leblanc, T., Petropavlovskikh, I., and Yates, E. L.: Revisiting the evidence of increasing springtime ozone mixing ratios in the free troposphere over Western North America, Geophysical Research Letters, 42, https://doi.org/10.1002/2015GL065311.Received, 2015.

Ma, P.-l., Rasch, P. J., Chepfer, H., Winker, D. M., and Ghan, S. J.: Observational constraint on cloud susceptibility weakened by aerosol retrieval limitations, Nature Communications, 9, https://doi.org/10.1038/s41467-018-05028-4, 2018.

Nappo, C., Caneill, J., Furman, R., Gifford, F., Kaimal, J., Kramer, M., Lockhart, T., Pendergast, M., RA, P., Randerson, D., Shreffler, J., and Wyngaard, J.: The workshop on the representativeness of meteolorogical observations, June 1981, Boulder, Colo, Bulletin of the American Meteorological Society, 63, 761–764, 1982.

Pitkänen, M. R. A., Mikkonen, S., Lehtinen, K. E. J., Lipponen, A., and Arola, A.: Artificial bias typically neglected in comparisons of uncertain atmospheric data, Geophysical Research Letters, 43, 10,003–10,011, https://doi.org/10.1002/2016GL070852, http://doi.wiley.com/10.1002/2016GL070852, 2016.

Putman, W., da Silva, A., Ott, L., and Darmenov, A.: Global Modeling and Assimilation Office Model Configuration for the 7-km GEOS-5 Nature Run , Ganymed Release (non-hydrostatic 7km Global Mesoscale Simulation), Tech. Rep. GMAO Office Note No 5, NASA, http://gmao.gsfc.nasa.gov/pubs/office{_}notes, 2014.

Remer, L., Kaufman, Y., and Kleidman, R.: Comparison of Three Years of Terra and Aqua MODIS Aerosol Optical Thickness Over the Global Oceans, IEEE Geoscience and Remote Sensing Letters, 3, 537–540, https://doi.org/10.1109/LGRS.2006.879562, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1715312, 2006.

Santese, M., De Tomasi, F., and Perrone, M. R.: AERONET versus MODIS aerosol parameters at different spatial resolutions over southeast Italy, Journal of Geophysical Research, 112, D10 214, https://doi.org/10.1029/2006JD007742, http://www.agu.org/pubs/crossref/2007/2006JD007742.shtml, 2007.

Sayer, A. M., Thomas, G. E., Palmer, P. I., and Grainger, R. G.: Some implications of sampling choices on comparisons between satellite and model aerosol optical depth fields, Atmospheric Chemistry and Physics, 10, 10 705–10 716, https://doi.org/10.5194/acp-10-10705-2010, http://www.atmos-chem-phys.net/10/10705/2010/, 2010.

Schmid, B., Michalsky, J., Halthore, R., Beauharnois, M., Harnson, L., Livingston, J., Russell, P., Holben, B., Eck, T., and Smirnov, A.: Comparison of Aerosol Optical Depth from Four Solar Radiometers During the Fall 1997 ARM Intensive Observation Period, Geophysical Research Letters, 26, 2725–2728, 1999.

Schutgens, N.: Representativeness of AERONET and GAW aerosol observation sites, https://doi.org/10411/XDZD4A, https://hdl.handle.net/10411/XDZD4A, 2019.

Schutgens, N., Gryspeerdt, E., Weigum, N., Tsyro, S., Goto, D., Schulz, M., and Stier, P.: Will a perfect model agree with perfect observations? The impact of spatial sampling, Atmospheric Chemistry and Physics Discussions, 16, https://doi.org/10.5194/acp-2015-973, http://www.atmos-chem-phys-discuss.net/acp-2015-973/, 2016a.

Schutgens, N., Partridge, D. G., and Stier, P.: The importance of temporal collocation for the evaluation of aerosol models with observations, Atmospheric Chemistry and Physics, 16, 1065–1079, https://doi.org/10.5194/acp-16-1065-2016, 2016b.

Schutgens, N., Tsyro, S., Gryspeerdt, E., Goto, D., Weigum, N., Schulz, M., and Stier, P.: On the spatio-temporal representativeness of observations, Atmospheric Chemistry and Physics, 17, 9761–9780, https://doi.org/10.5194/acp-2017-149, https://www.atmos-chem-phys-discuss.net/acp-2017-149/, 2017.

Schutgens, N. J., Nakata, M., and Nakajima, T.: Validation and empirical correction of MODIS AOT and AE over ocean, Atmospheric Measurement Techniques, 6, 2455–2475, https://doi.org/10.5194/amt-6-2455-2013, http://www.atmos-meas-tech.net/6/2455/2013/, 2013.

Schwarz, M., Folini, D., Hakuba, M. Z., and Wild, M.: Spatial representativeness of surface-measured variations of downward solar radiation, Journal of Geophysical Research: Atmospheres, 122, 13 319–13 337, https://doi.org/10.1002/2017JD027261, http://doi.wiley.com/10.1002/2017JD027261, 2017.

Schwarz, M., Follini, D., Hakuba, M., and Wild, M.: From Point to Area : Worldwide Assessment of the Representativeness of Monthly Surface Solar Radiation Records, Journal of Geophysical Research : Atmospheres, 123, 13 857–13 874, https://doi.org/10.1029/2018JD029169, 2018.

Shi, X., Zhao, C., Jiang, J. H., Wang, C., Yang, X., and Yung, Y. L.: Spatial Representativeness of PM 2.5 Concentrations Obtained Using Reduced Number of Network Stations, Journal of Geophysical Research: Atmospheres, 123, 3145–3158, https://doi.org/10.1002/2017JD027913, http://doi.wiley.com/10.1002/2017JD027913, 2018.

Shinozuka, Y. and Redemann, J.: Horizontal variability of aerosol optical depth observed during the ARCTAS airborne experiment, Atmospheric Chemistry and Physics, 11, 8489–8495, https://doi.org/10.5194/acp-11-8489-2011, 2011.

Smirnov, A.: Diurnal variability of aerosol optical depth observed at AERONET (Aerosol Robotic Network) sites, Geophysical Research Letters, 29, 2115, https://doi.org/10.1029/2002GL016305, http://doi.wiley.com/10.1029/2002GL016305, 2002.

Sofieva, V. F., Kalakoski, N., Päivärinta, S. M., Tamminen, J., Kyrölä, E., Laine, M., and Froidevaux, L.: On sampling uncertainty of satellite ozone profile measurements, Atmospheric Measurement Techniques, 7, 1891–1900, https://doi.org/10.5194/amt-7-1891-2014, 2014.

Suarez, M. J., Darmenov, A. S., and da Silva, A.: The Quick Fire Emissions Dataset (QFED) - Documentation of versions 2 . 1 , 2 . 2 and 2 . 4, Tech. rep., NASA, 2013.

Virtanen, T. H., Kolmonen, P., Sogacheva, L., Rodríguez, E., Saponaro, G., and Leeuw, G. D.: Collocation mismatch uncertainties in satellite aerosol retrieval validation, Atmospheric Measurement Techniques, 11, 925–938, 2018.

Wang, R., Balkanski, Y., Boucher, O., Ciais, P., Schuster, G. L., Chevallier, F., Samset, B. H., Liu, J., Piao, S., Valari, M., and Tao, S.: Estimation of global black carbon direct radiative forcing and its uncertainty constrained by observations, Journal of Geophysical Research, 121, 5948–5971, https://doi.org/10.1002/2015JD024326, 2016.

Wang, R., Andrews, E., Balkanski, Y., Boucher, O., Myhre, G., Samset, B., Schulz, M., Schuster, G. L., Valari, M., and Tao, S.: Geophysical Research Letters, Geophysical Research Letters, 45, 2106–2114, https://doi.org/10.1002/2017GL076817, 2018.

Weigum, N. M., Stier, P., Schwarz, J. P., Fahey, D. W., and Spackman, J. R.: Scales of variability of black carbon plumes over the Pacific Ocean, Geophysical Research Letters, 39, https://doi.org/10.1029/2012GL052127, http://doi.wiley.com/10.1029/2012GL052127, 2012.

~~Evaluation of the G5NR simulation of AAOT with L2.0 AERONET data. Each dot represents statistics for a single AERONET site (with at least 30 observations in 2006); the mean value is shown in red and the standard deviation in blue. The coloured text summarizes the statistics over all data points in the figure.~~

~~Yearly representation errors for AOT from DirectSun L2.0 AERONET in different regions and a model grid-box size of 1^o.~~

~~Yearly representation errors for AOT from AERONET for different products and a model grid-box size of 1^o. The $*$ refers to Inversion products with artificially lowered temporal coverage in the Northern Hemisphere (see Sect. 4).~~

~~Yearly representation errors for AOT from from DirectSun L2.0 AERONET using all-sky or clear sky conditions and model grid-box size of 4^o (left) or 1^o (right).~~

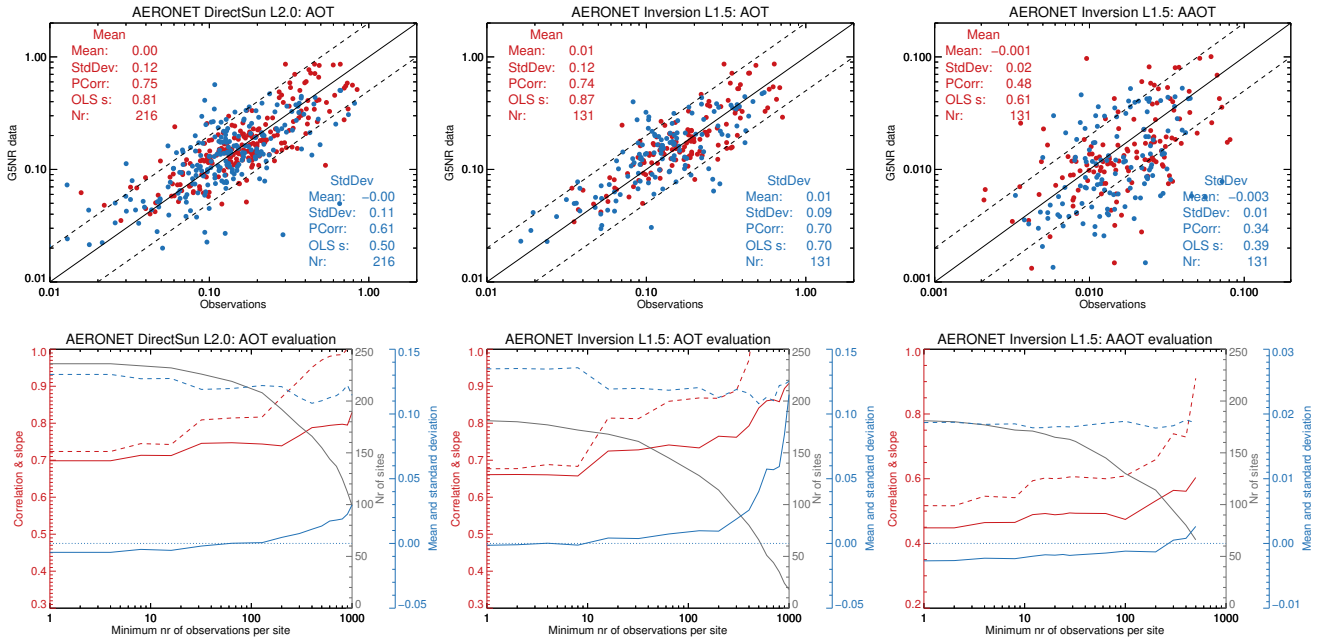


Figure 1. Evaluation of the G5NR simulation of AOT and AAOT with AERONET data. The top row shows evaluation against three different datasets. Each dot represents ~~statistics~~ the yearly mean or standard deviation for a single AERONET site (with at least 100 observations in 2006); the mean value is shown in red and the standard deviation in blue. The coloured text summarizes the statistics over all data points in the figure. In the bottom row, the impact of the minimum required number of observations per site on those summary statistics (for means) is shown. Colours relate lines to axes and have different meaning than in the top row. Red solid is correlation, red dashed is slope, blue solid is mean, and blue dashed is standard deviation. In all figures, hourly G5NR model data was collocated in time & space with AERONET observations before calculating site statistics.

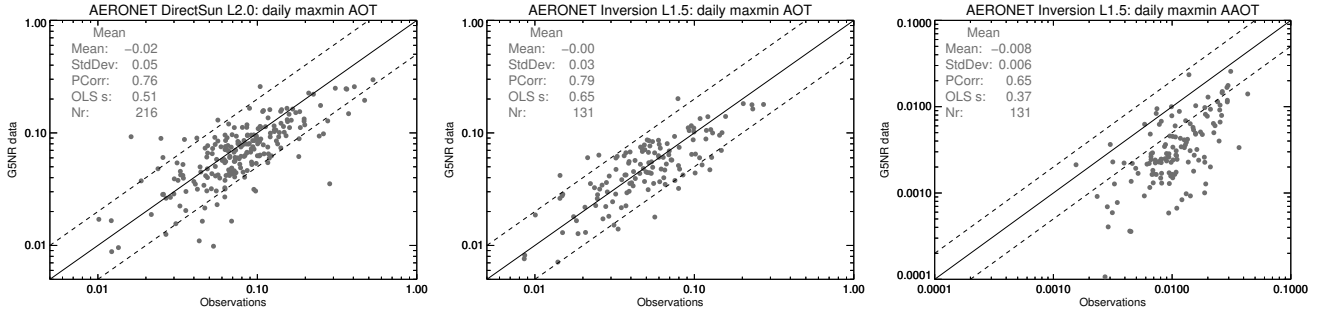


Figure 2. Evaluation of the G5NR simulation of ~~daily variation in~~ AOT and AAOT with AERONET data. Each dot represents ~~statistics~~ the yearly average of daily variation (maximum minus minimum value) for a single AERONET site (with at least 100 observations in 2006). The grey text summarizes the statistics over all data points in the figure. In all figures, hourly G5NR model data was collocated in time & space with AERONET observations before calculating site statistics.

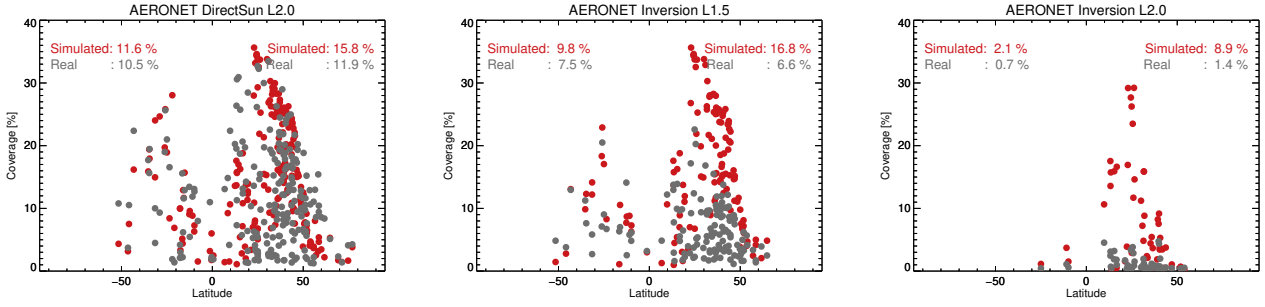


Figure 3. Evaluation of the ~~observational~~ temporal coverage predicted by the OSSE with AERONET observations. Each dot represents ~~statistics~~ temporal coverage (or frequency of observation) for a single AERONET site (with at least 100 observations in 2006, at least 30 observations for Inversion L2.0). The grey dots are real AERONET data, the red dots are simulated by the methodology described in Sec. 3. The numbers in the graph are temporal coverages estimated by hemisphere.

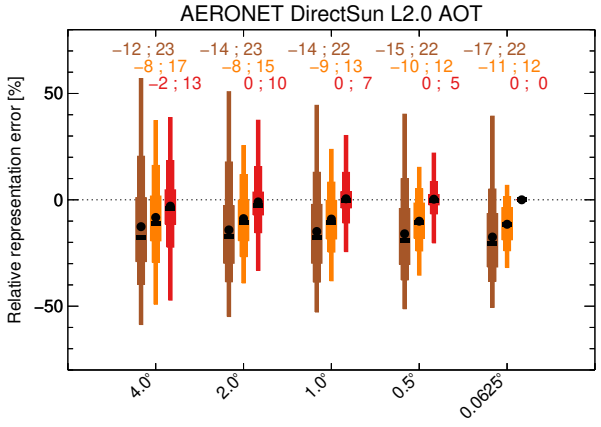


Figure 4. Yearly representation errors for AOT from DirectSun L2.0 AERONET for different model grid-box sizes. The colours indicate different collocation protocols: yearly (brown), daily (orange) and hourly (red). Numbers on top are mean of the errors and mean of the sign-less errors.

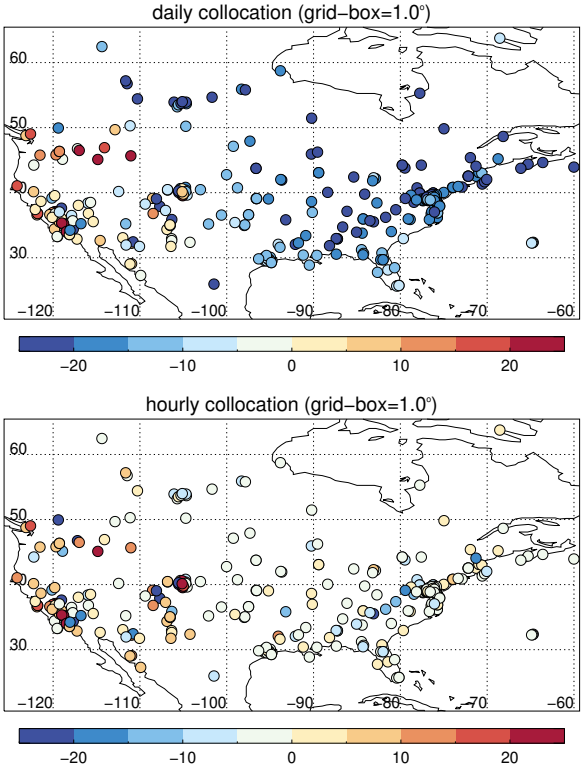


Figure 5. ~~Yeraly~~ Yearly representation errors for AOT from DirectSun L2.0 AERONET in Northern America, for two different collocation protocols (top: daily; bottom: hourly) and a model grid-box size of 1^o.

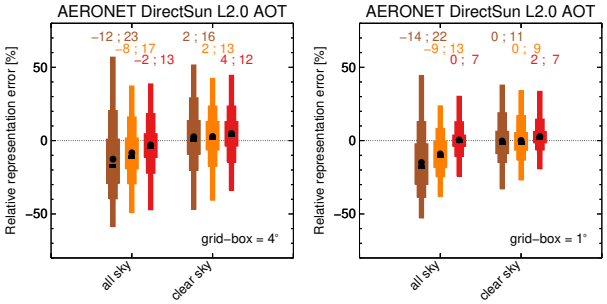


Figure 6. Yearly representation errors for AOT from from DirectSun L2.0 AERONET ~~in Europe, for two~~ using all-sky or clear sky conditions and model grid-box size of 4^o (left) or 1^o (right). The colours indicate different collocation protocols~~(top~~: yearly (brown), daily ~~; bottom:~~ (orange) and hourly (red). Numbers on top are mean of the errors and ~~a model grid-box size~~ mean of ~~1^o~~ the sign-less errors.

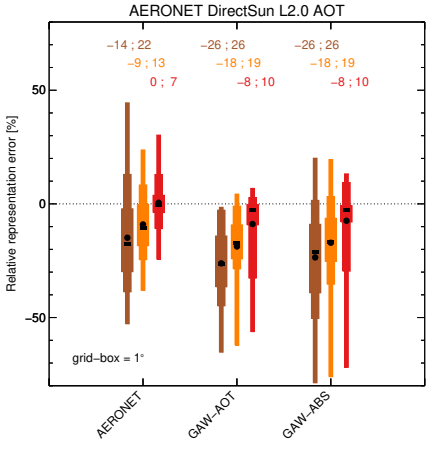


Figure 7. Yearly representation errors for AOT from Direct Sun L2.0 AERONET and GAW and a model grid-box size of 1^o. The colours indicate different collocation protocols: yearly (brown), daily (orange) and hourly (red). Numbers on top are mean of the errors and mean of the sign-less errors.

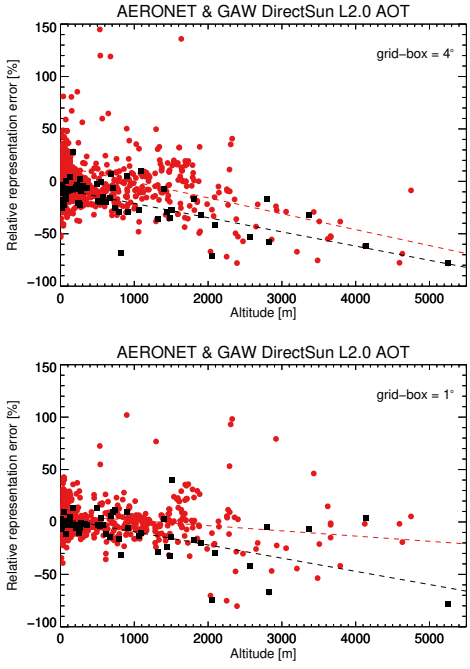


Figure 8. Yearly representation errors for AOT from Direct Sun L2.0 AERONET (red circles) and GAW (black squares) as a function of site altitude, for a model grid-box size of either 4^o or 1^o; using hourly collocation.

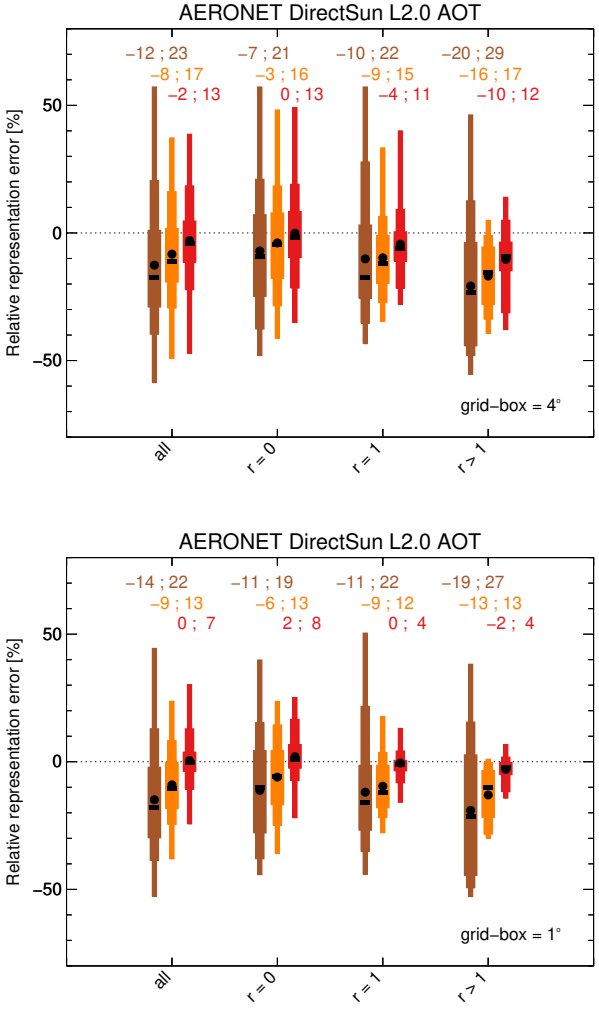


Figure 9. Yearly representation errors for AOT from DirectSun L2.0 AERONET for different ~~representation rankings~~ range scores r by Kinne et al. (2013), for a model grid-box size of either 4^o or 1^o. The colours indicate different collocation protocols: yearly (brown), daily (orange) and hourly (red). Numbers on top are mean of the errors and mean of the sign-less errors.

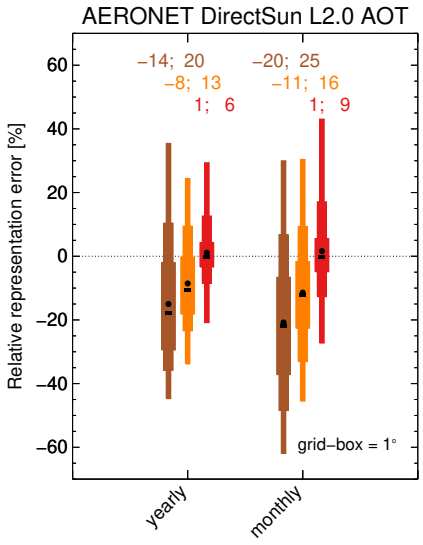


Figure 10. Yearly and monthly representation errors for AOT DirectSun L2.0 AERONET, for a model grid-box size of 1^o. The colours indicate different collocation protocols: yearly (brown), daily (orange) and hourly (red). Numbers on top are mean of the errors and mean of the sign-less errors.

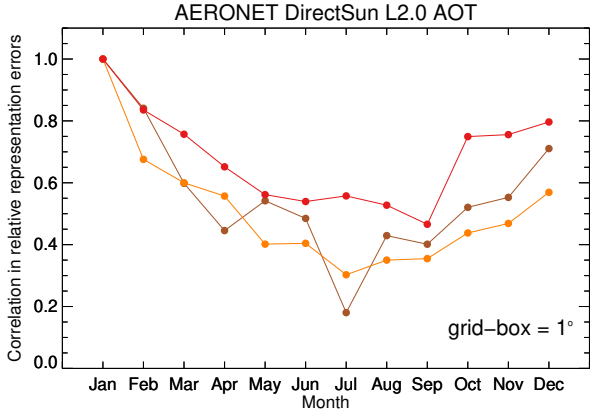


Figure 11. Correlation in monthly representation errors with errors for January, for AOT DirectSun L2.0 AERONET, for a model grid-box size of 1^o. The colours indicate different collocation protocols: yearly (brown), daily (orange) and hourly (red).

~~Maximum (red) & minimum (blue) monthly representation errors versus yearly representation errors, for AOT DirectSun L2.0 AERONET, for a model grid-box size of 1^o and hourly collocation~~

1210

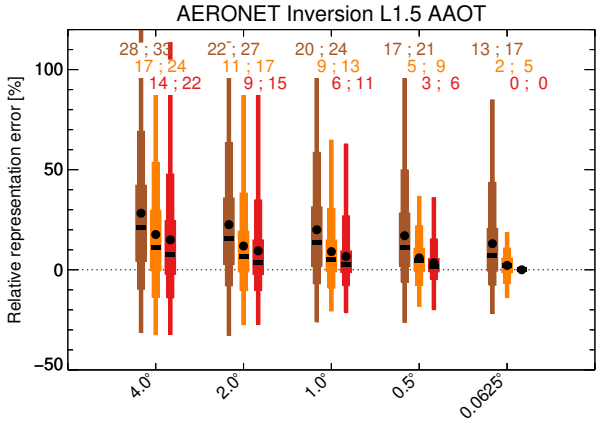


Figure 12. Yearly representation errors for AAOT from Inversion L1.5 AERONET for different model grid-box sizes. The colours indicate different collocation protocols: ~~no colocation~~ yearly (brown), daily ~~collocation~~ (orange) and hourly ~~collocation~~ (red). Numbers on top are mean of the errors and mean of the sign-less errors.

~~Yearly representation errors for AAOT from Inversion L1.5 AERONET in different regions and a model grid-box size of 1^o.~~

~~Yearly representation errors for AAOT from Inversion L1.5 AERONET and GAW and a model grid-box size of 1^o.~~

~~Yearly representation errors for AAOT from Inversion L1.5 AERONET for different representation rankings by Kinne et al. (2013), for a model grid-box size of 1^o.~~

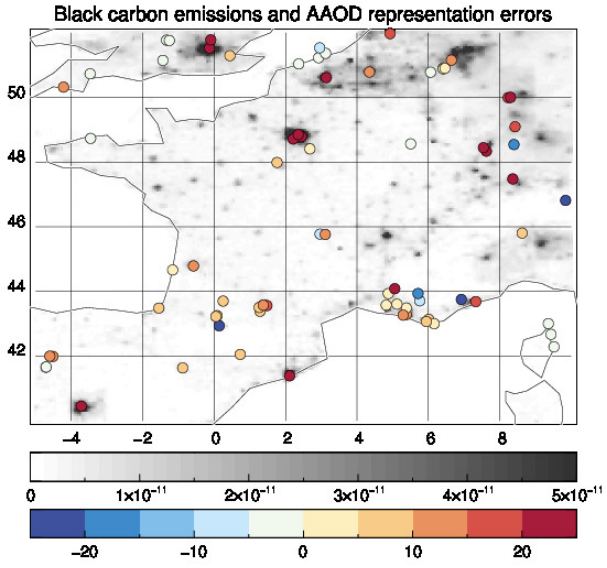


Figure 13. Black carbon emissions over France, Europe, with the representation errors in AAOT from Inversion L1.5 AERONET super-imposed. The ~~representation errors use the same colour bar as in Fig. ??~~ top colourbar (white-black) represents emissions ([kg/m^2s]), and ~~runs from -25%~~ the bottom colourbar (~~blue~~blue-red) ~~to $+25\%$~~ represents relative representation errors (~~red~~[%]). Only *spatial* representation errors are shown, i.e. the temporal sampling of observations is ignored.

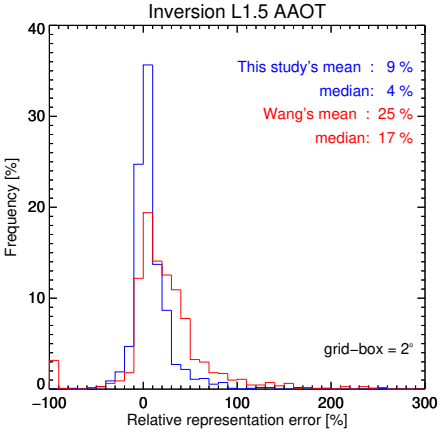


Figure 14. Yearly representation errors for AAOT from Inversion L1.5 AERONET as estimated in this paper or using the methodology from Wang et al. (2018) and a model grid-box size of 2^o. The representation error shown is the spatial representation error (Schutgens et al., 2017), i.e. temporal sampling of observations is ignored.

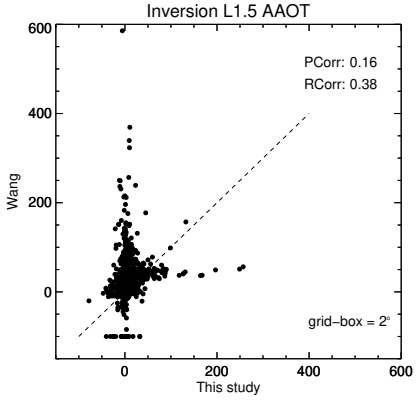


Figure 15. Comparison of yearly representation errors for AAOT from Inversion L1.5 AERONET as estimated in this paper or using the methodology from Wang et al. (2018) and a model grid-box size of 2^o. The representation error shown is the spatial representation error (Schutgens et al., 2017), i.e. temporal sampling of observations is ignored. Also shown are the Pearson linear correlation (PCorr) and rank correlation (RCorr) between the data. The dashed line shows $y = x$.