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.