The authors thank the reviewer Dr Andrew Sayer for carefully reading the manuscript and helping us improve it through many useful comments and questions.

1. General comments and recommendation

The main comment of note is the use of 6-hour vs. 24-hour comparisons, and how it would be very useful to be able to state to what extent 6-hour data are better than 24-hour data for these types of comparison (it could help set a new standard approach).

We will address this comment in more detail below, when answering to specific comments. A higher frequency of model output (or conversely a finer granularity in defining the time of an observation) is never a bad thing, although it may cause issues with data storage and processing. If the question is about an optimal frequency (leaving aside the definition of 'optimal'), the obvious but probably also slightly disappointing answer is that it will depend on the application.

2. Specific comments

General: I think that the Copernicus style guide requires acronyms defined in the Abstract (e.g. AOT, AE, SSA) to be defined in the main text at first use as well. Similarly MODIS, AERONET, AEROCOM and others are presented without definition.

Acronyms are now defined as suggested by the reviewer and as requested by the ACP style guide.

P26193, L20 onwards: Another study which might be worth mentioning here is Smirnov et al. (2002), which looked at the diurnal variability of AOT from AERONET sites. So this was similar in principle to the Kaufman study cited, although shows the breakdown of results from individual sites directly.

We have added a mention of the Smirnov et al 2002 results. Their results provide a bit more context for the Kaufman et al 2000 paper, which we felt was a bit optimistic regarding diurnal cycles in aerosol (Smrinov et al do find diurnal cycles of 40% in AOT).

P26194, L7-12: You might add that it is more common for satellite-satellite comparisons to use common spatial sampling on a daily basis already (e.g. the various comparisons between SeaWiFS and other data presented in Sayer et al., 2012), since in that case both data sets have temporal sampling limitations. So these issues are known about in the satellite community, it's more that the modelling community haven't started accounting for these sampling issues as often yet. (And hopefully the present study will contribute to this changing.)

We agree with the reviewer. Actually our summary (p. 26204, lines 14-16: "We'd like to point out that the practice of temporal collocation of datasets is very normal in the remote sensing and data assimilation communities, but less so in the modelling community. ") already contains a strong statement on this practice. We prefer not to mention this in the paragraph suggested by the reviewer as that paragraph is very specifically on model comparisons.

P26197, L1: I would suggest Levy et al. (2007, 2010) as better references for the MODIS Collection 5 aerosol products, since Remer et al. (2006) (mistakenly given as 2005 in the paper) was an analysis of Collection 4, not Collection 5, data. The Levy papers provide some algorithm/validation details for Collection 5 (although focused

onland; there isn't an ocean C5-focused paper to my knowledge).

Thanks for spotting our error in suggesting the Remer et al 2005 paper deals with Coll 5. We still refer to this paper as it discusses the over-ocean retrieval (but indicate this is for Coll 3 and 4) and point to Levy et al 2007 for the Coll 5 over land retrieval. The references here are intended to point to retrieval discussions, not so much validation studies.

P26197, L24-27: I agree that it's a good idea to be using these bias-corrected data sets for probing these sampling effects on the models, and in that sense 6-hourly output would seem to make sense, since that's the temporal output of the NRL product. However, a large number of model comparisons are done just with the standard MODIS products, which are provided on a daily (or longer) basis, and so these comparisons are made on a daily/monthly/yearly basis. So, my question: is it feasible to add a discussion of how much the results change if the temporal compositing of the remote sensing data is not 6-hourly but 24-hourly? I think that this would be helpful. If it turns out that things don't change much, then this suggests that people who are currently doing monthly/yearly comparisons may be ok if they just go to daily remote sensing products instead (which are readily available). On the other hand if 6-hour and 24-hour results are different, then this is useful information since it means that us satellite data providers should really consider providing 6-hourly level 3 output (as opposed to the current daily/8-day/monthly now) as a standard, because these 6-hourly products are not as highly used (i.e. although the NRL and AORI bias-corrected data sets exist, a lot of people still go to the un-bias-corrected NASA standard products). As is noted later in the manuscript, standard practice for AEROCOM output is also daily rather than 6-hourly output, so this would be a change which the modelling teams would have to adopt as well. So the answer to the 6-hour vs. 24-hour question could be in my view an important finding and action item coming out of the manuscript, if the analysis could be extended to answer it.

We want to thank the reviewer for raising a very interesting point. Our original manuscript gave an answer to these questions on p. 26204, line 14 to p 26205, l 2, and especially from line 25 onward. However, prompted by the reviewer's comments we revisited this issue and realized our original assessment had been too optimistic.

In the new manuscript, we now show plots (new Fig. 15 and discussion in third but last paragraph of Sect 5.5) of the temporal sampling errors when yearly averages are constructed from daily data. In that case, errors are much reduced from the case where a straight yearly average of the model is used. However, we still find non-negligible sampling errors of typically 7-17% (depending on model). More-over these errors seem to correlate with the modelled diurnal cycles that we analysed and found to be under-estimated compared to AERONET data. Please see the attached PDF for some figures.

P26202, L7-11: Clouds as well as snow/ice are important in winter.

We have added a reference to this in the text.

P26202, L18 onwards: This paragraph indicates an artificial contrast between land and ocean can be created in AE data, as a result of a minimum AOT threshold in the AORI bias-correction and filtering algorithm. I think that it should be made clear that this is a specific feature of the AORI data set (from the current text it is implicit if one is familiar with the literature and MODIS products, but a less familiar reader may infer incorrectly that this is a feature in the standard MODIS products, which is not the case). Therefore this result may not be transferable if one were using the standard

MODIS products, or an alternative bias-correction technique which used a different (or no) AOT threshold.

Also of relevance to this section is that the latest MODIS Collection 6 Dark Target products do not provide AE over land, only over ocean (Levy et al., 2013). This was a result of analyses (e.g. Levy et al., 2010) which indicated that there was little skill in this parameter. So this result about the artificial contrast is not transferable to the most current version of the MODIS data used. (NRL and AORI bias corrections have not yet been made available for Collection 6, so it's reasonable to use the older Collection 5 data, although the caveat could be made in the paper that the results will change a bit once these bias-correction data sets are updated). Note also that MODIS Deep Blue (Hsu et al., 2013), not used in the present study, DOES provide AE over land and appears to have some skill (Sayer et al., 2013). So this information is still available from standard MODIS Deep Blue products over land (albeit not in the Dark Target products which are used in the current versions of the NRL/AORI products).

The reviewer is right that this concerns the AORI dataset. We believe that the effect will be seen in other datasets as well. First, the AORI dataset only considers retrievals over ocean (see Fig. 9). The only distinction is made for retrievals in known outflow regions and retrievals over the deep ocean. No over-land retrievals are included in this AE analysis. Second, the minimum AOT value that causes this contrast can be found from the official Coll. 5 data. We refer to Fig. 11 in Schutgens et al. AMT 2013 which strongly suggests that AE calculated from official MODIS AOT products agrees much better with AERONET for AOT> 0.06.

The one caveat here is that Schutgens et al 2013 calculated AE from official Coll 5 MODIS AOT at 860 and 470nm (to bring it better in line with AERONET AE from 870 and 440nm), instead of using the official MODIS AE product. As the reviewer knows, a dependence of AE quality on AOT may be expected.

We have added a paragraph to discuss this.

P26203, Section 5.4: The plots accompanying this (Figures 10 and 11) are interesting in that the errors, as one would expect, are smallest when the observational data coverage is high. I wonder if the authors could add a map or some discussion showing what the observational data coverage is for the data sets used. This way if one wanted to put a filter of, say, 50% coverage when considering where to compare, we could get an idea for how much of the world this would throw away, and where this would happen. I did something similar to this in the paper Sayer et al. (2010), which the authors cite earlier in their manuscript, although this was fraction of grid box filled rather than fraction of time series filled, and the observational data I used then had poorer temporal sampling. I realise that the point of the present study is to encourage people to use coincident temporal sampling, but the point of my suggestion here is that seeing where these low-coverage areas are most likely to occur can also provide guidance for understanding results from older studies where such temporal collocation was not performed.

The problem is, of course, that coverage is only one predictor of sampling error and by no means is it conclusive. More-over, for AERONET spatial coverage is sparse such that a map might not be very useful. For MODIS NRL Aqua AOT, we have provided a yearly coverage map in Fig. 7. See also the discussion on p 26202, lines 4-6.

P26203, L18: The minimum AERONET AOT threshold for a successful Level 2.0 inversion with SSA is 0.4 (at 440 nm), not 0.2 as stated here. I'd suggest adding a

reference to Dubovik et al. (2000) here, together with a bit more discussion about this topic.

The reviewer is right that 0.4 is often quoted as a minimum AOT for successful retrievals. However, the data files contain retrievals for AOT starting at 0.2 and (where possible) we have not noted significant differences between SSA retrievals for 0.2 < AOT < 0.4 and those above 0.4.

Note that a higher minimum AOT will of course lead to larger temporal sampling errors.

P26204, L22: This discussion and associated figure could also be used to argue for model and satellite output being provided at 6-hourly, rather than daily, time scales (see my prior main comment). This is an important figure since it basically shows a longitudinal dependence of error from the definition of a 'day' relative to UTC.

We agree, although it will depend on the time-scales that researchers will be interested in (see our response to the reviewer's previous comment).

P26205, L6-8: I'd like the authors to expand a bit on this comment in the Conclusions. If sampling errors on daily data are larger than observational errors, then that suggests that the first step in improving the utility of observational data sets is not to reduce observational errors but to improve observational coverage. Although cloud/surface issues will preclude complete coverage, this could be used to make the argument for a multi-sensor bias-corrected AOT data set a priority. I believe that NRL either have or are creating a combined MODIS Terra/Aqua data set (not sure if MISR is being included as well or not, but this would fill some gaps in MODIS Terra Sun glint holes over the ocean). It (the whole paper in fact) is also an argument for the development of AOT data sets from geostationary sensors (particularly the second-generation such as GOES-R, Himawari-8, MTG) as well as DSCOVR/EPIC at the L1 point to get at diurnal variability. It'd be good to have some more discussion of these possibilities, to help point to the need for them as a complement to our existing polar orbit sensor data records.

While we agree with the reviewer there seem to be two important caveats. One, for MODIS satellite observations daily sampling errors are smaller than observational errors: compare Fig 12 (MODIS) and 15 (AERONET) for AOT errors. So for satellite data, better retrieval algorithms/better quality control seems to be the priority. Two, temporal sampling errors in model evaluation can be simply addressed through high-frequency output (e.g. 3 or 6-hourly). This merely requires more storage space.

However, we believe that a better understanding of aerosol processes (and their evaluation in models) will require analysis of time-series and not just yearly averages. That would argue for space-borne observing systems that have a better temporal resolution than the current polar-orbiting satellites. As Fig. 4 suggests, being able to observe daily variation in AOT from space would offer great opportunities to evaluate models.

While data assimilation is an interesting technique for filling gaps due to temporal sampling of observations, it is important to note that the most often used observational datasets (MODIS Aqua and Terra and MISR as well as CALIOP) do not differ much in their observing times. Also, it is currently unclear whether global models have any ability in representing diurnal cycles (see Fig. 4), let alone daily variation.

A new paragraph has been added to the Conclusions to discuss some of these issues.

Table 2: From the referring text (P26201) and caption, I'm not certain what the different rows in this table refer to. Why isn't there just one number per model? Can the text and caption be clarified?

The table itself shows δ_o / δ_m, the ratio of daily variation in observations and model. This ratio is itself a yearly and (sparse) global average (based on AERONET). The different rows refer to a selection criterion based on the value of daily δ_o (i.e. observed daily variation). What we see is that as observed daily variation increases, so does the ratio: models are really poor at representing observed cases of strong daily variation. The δ_o threshold values are based on AERONET AOT retrieval uncertainties (± 0.01). If we assume two measurements during a day to be independent, their difference will have an error of $\sqrt{(2 \times 0.01^2)}$. We have added a bit more text to clarify.

Figures, general: Can we have a map of the NRL and AORI global AOT fields, to see how similar they are? (Or at least a brief discussion in the text about this.)

That would take this paper into the realm of an intercomparison or even validation of these datasets which is not our purpose. Note that the AORI dataset uses different QA criteria and so its temporal sampling is different from NRL. Yearly averaged AORI AOT and AE can be seen in Fig. 15 & 16 in Schutgens et al. AMT 2013. Unpublished evaluation of both AORI and NRL data against AERONET and Maritime Aerosol Network measurements suggest that AORI is slightly better, especially for low AOT (where NRL has a positive bias).

Note that we very briefly discussed differences between NRL and AORI in lines 10-13 of p 26197.

Figures 1,2, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15: Can font size on labels and legends be increased?

There is a trade-off here between a space for the actual figure (of which we have a lot) and space for the labels and legends. Since we already have a lot of figures (and consequently more pages than is usual) for this paper, we prefer to be economical. We have also seen that in the final ACP format the figures (and their labels) come-out somewhat larger than in the ACPD format.

Figure 2 (and other AOT maps later e.g. Figure 7): For the panels showing AOT, green is not necessarily an intuitive color scale for showing AOT variations: something based on red/brown might be better (e.g. the reverse of the scale used in Figure 13). Also, a discrete colour scale (e.g. 10 levels from 0-0.5) might be easier to pick out the exact AOT at different locations from the different models. At present for example it's hard to resolve some of the shades. Also, it might make sense to put the AOT and AOT standard deviation on the same range (currently AOT is 0-0.5 while standard deviation is 0-1); again, with different scales, it's hard to compare, and the overall impression from just looking at the hues is that the AOT standard deviation is small where from the bottom row we can tell this is actually often not the case.

We understand that every-one has their preferred esthetics for figures. While red/brown may be a standard colour scale amongst some of our colleagues it is by no means the norm. While AOT and its standard deviation in Fig 2 do not have the same scale (to better represent these two different fields), the lowest row shows their ratio. As can be seen, there are substantial differences between the models. Regarding more discrete colourbars, we have experimented with 10 levels but saw no noticeable

improvement (our subjective judgment, arguably). Some example plots are shown below: