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.

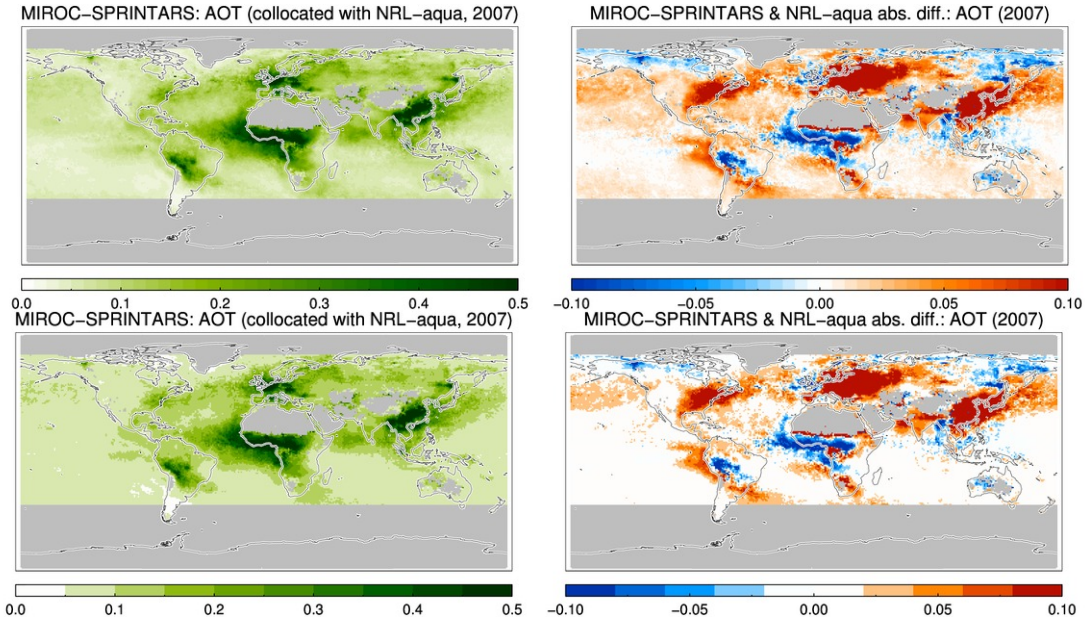
Figures 1,2, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15: Can font size on labels and legends be increased?

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:



Manuscript prepared for Atmos. Chem. Phys.
with version 2014/09/16 7.15 Copernicus papers of the LaTeX class copernicus.cls.
Date: 8 January 2016

The importance of temporal collocation for the evaluation of aerosol models with observations

N.A.J. Schutgens[1], **D.G. Partridge**[2,3], **and P. Stier**[1]

[1]Department of Physics, University of Oxford, Parks road, OX1 3PU, England
[2]Department of Environmental Science and Analytical Chemistry, Stockholm University, Stockholm, Sweden
[3]Bert Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden

Correspondence to: N.A.J. Schutgens
(schutgens@physics.ox.ac.uk)

Abstract. It is often implicitly assumed that over suitably long periods the mean of observations and models should be comparable, even if they have different temporal sampling. We assess the errors incurred due to ignoring temporal sampling and show they are of similar magnitude as (but smaller than) actual model errors (20–60%).

Using temporal sampling from remote sensing datasets~~(~~, the satellite imager MODIS (MODerate resolution Imaging Spectroradiometer) and the ground-based sun photometer network AERONET ~~)~~ (AErosol Robotic NETwork), and three different global aerosol models, we compare annual and monthly averages of full model data to sampled model data. Our results show that sampling errors as large as 100% in AOT (Aerosol Optical Thickness), 0.4 in AE (Ångström Exponent) and 0.05 in SSA (Single Scattering Albedo) are possible. Even in daily averages, sampling errors can be significant. More-over these sampling errors are often correlated over long distances giving rise to artificial contrasts between pristine and polluted events and regions. Additionally, we provide evidence that suggests that models will underestimate these errors. To prevent sampling errors, model data should be temporally collocated to the observations before any analysis is made.

We also discuss how this work has consequences for in-situ measurements (e.g. aircraft campaigns or surface measurements) in model evaluation.

Although this study is framed in the context of model evaluation, it has a clear and direct relevance to climatologies derived from observational datasets.

1 Introduction

In the past decades, the role of atmospheric aerosol in the Earth's climate and biosphere has become clearer. Aerosols can change the global radiation budget directly (Angstrom, 1962) and indirectly (Twomey, 1974; Albrecht, 1989). They can affect the temperature structure of the atmosphere (Hansen et al., 1997; Lohmann and Feichter, 2005) and may have consequences for the hydrological cycle (Lohmann and Feichter, 1997). Dust aerosols may transport nutrients for the biosphere over long distances (Vink and Measures, 2001; McTainsh and Strong, 2007; Maher et al., 2010; Lequy et al., 2012) and anthropogenic aerosols can pose health hazards for humans (Dockery et al., 1993; Brunekreef and Holgate, 2002; Ezzati et al., 2002; Smith et al., 2009; Beelen et al., 2013).

One approach to explore the role of aerosol, is through (global) models. A lot of effort has gone into evaluating global aerosol models against observations (Kinne et al., 2006; Schulz et al., 2006; Textor et al., 2006, 2007; Huneeus et al., 2011; Koch et al., 2009; Quaas et al., 2009; Koffi et al., 2012) in the framework of the AEROCOM (AEROsol Comparisons between Observations and Models) community (http://aerocom.met.no). Although this has allowed the community to identify weaknesses in models (and has driven further model development), less effort has gone into developing best practices when performing such evaluation.

When evaluating models with observations, one problem is the different temporal sampling of the datasets. While models can in principle provide data at any time, observations may not be available for much of a day or even extended periods (up to several weeks) due to e.g. satellite overpass times, cloudiness, low-light conditions or other unfavourable

circumstances as well as instrument malfunction or maintenance. Although limited temporal sampling is clearly an issue for remote sensing datasets, in-situ data may likewise be affected. This temporal sampling issue is however ignored in many studies where it is implicitly assumed that suitably long averaging periods (months or years) will put model and observational datasets on equal footing.

In this paper, we will study errors due to temporal sampling errors by applying temporal samplings from actual observational datasets to global model data. The temporal evolution of aerosol has received some interest before: Anderson et al. (2003) used hourly nephelometer measurements of dry extinction to study time-scales of aerosol evolution over one polluted and one remote site. They found time-scales from 2 to 48 hours, with aerosol often becoming uncorrelated over a day and a half. Kaufman et al. (2000) used AERONET ~~AOT~~ (AErosol Robotic NETwork) AOT (Aerosol Optical Thickness) data to assess whether a single daily overpass from either Terra or Aqua satellite would allow an accurate estimate of daily-averaged AOT. They concluded this was indeed possible, suggesting that aerosol diurnal cycles are small. **?**, also using AERONET AOT, showed that diurnal cycles of 10 to 40% existed near polluted and biomass burning sources while diurnal cycles were insignificant over ocean. Remer et al. (2006) put an upper limit of 2.5% on regional AOT changes over three hours by comparing MODIS (MODerate resolution Imaging Spectroradiometer) Aqua and Terra aerosol observations over the remote oceans. Geogdzhayev et al. (2014) studied the ability of MODIS Aqua and Terra to assess regional and monthly AOT using model data and MODIS L2 temporal samplings. They found differences in monthly, regional AOT of 0.01-0.02 over ocean and of 0.03-0.09 over land. The model data consisted of daily-averages so no investigation of diurnal cycles was possible.

The previous studies highlight aspects of the temporal evolution of aerosol but always with the purpose of deriving climatologies from observational datasets. Temporal sampling of aerosol observations in the context of model evaluation has received little interest, although Sayer et al. (2010) pointed out that differences in AOT of up to 0.1 in monthly and annual regional averages were possible when comparing model data to AATSR (Advanced Along-Track Scanning Radiometer) satellite observations.

Global models are routinely compared to a host of remote sensing data, from satellites and ground-sites. Within AEROCOM, standard practice is to use daily averaged model data in these comparisons but in the literature monthly or yearly averages are used as well. With regard to temporal sampling issues, it seems that many questions are still left unanswered: what magnitude of temporal sampling error is possible on daily, monthly and yearly time-scales, for individual grid points? Do these errors exhibit distinct spatial patterns? Do aerosol diurnal cycles impact temporal sampling errors? Can models represent aerosol temporal variability well enough to be used to assess sampling errors? Are intensive aerosol properties like AE ~~and SSA~~ (Ångström Exponent) and SSA (Single Scattering Albedo) less affected than extensive aerosol properties like AOT? What controls the size of these sampling errors? And finally, how do temporal sampling errors compare to model errors and observational errors? In this paper, we make a first attempt at answering these questions.

In Sec. 2, we give a brief overview of the models that will be used, followed by a similarly brief overview of the observational datasets in Sect. 3. Our methodology is explained in more detail in Sect. 4, but the bulk of our paper is dedicated to the results in Sect. 5. We will start by describing simulated observations, their temporal variability and how they compare to actual observations in Sect. 5.2. Next, we will discuss the spatial patterns in temporal sampling errors. A fuller explanation of what controls temporal sampling errors is given in Sect. 5.4. These errors are compared to observational errors and models errors in Sect 5.5. The paper is wrapped up with a summary (Sect. 6) of our results.

2 Models

Three global aerosol models~~(~~, ECHAM-HAM, HadGEM-UKCA and MIROC-SPRINTARS, see also Table 1~~)~~, are used in this paper to provide time-varying aerosol fields.

The global aerosol model ECHAM-HAM consists of the aerosol module HAM (Stier et al., 2005, 2007; Zhang et al., 2012) coupled to the atmospheric general circulation model ECHAM (Roeckner et al., 2003, 2006). It solves the prognostic equations for vorticity, divergence, surface pressure and temperature using spherical harmonics with triangular truncation. Aerosols are advected with a flux-form semi-Lagrangian transport scheme (Lin and Rood, 1996) on a Gaussian grid. The aerosol module HAM calculates the global evolution of five aerosol species: ~~sulfate~~sulphate, particulate organic matter, black carbon, sea salt and dust. These species are the constituents of both internally and externally mixed aerosol particles whose size distribution is represented by 7 uni-modal log-normal distributions called modes. These 7 modes describe four size classes (nucleation, Aitken, accumulation and coarse) and two hygroscopic classes (hydrophobic and hydrophilic). HAM uses the two-moment M7 aerosol microphysics scheme (Vignati et al., 2004).

The global aerosol model HadGEM-UKCA (Bellouin et al., 2013) uses the UKCA aerosol & chemistry schemes with the third generation of the Hadley Centre Global Environmental Model (Hewitt et al., 2011) developed at the UK Met Office. This circulation model is non-hydrostatic and uses a semi-Langangian transport scheme. UKCA calculates the evolution of five aerosol species, ~~sulfate~~sulphate, particulate organic matter, black carbon, sea salt and dust, in both internally and externally mixed particles. The aerosol scheme in UKCA is based on the Global Model of Aerosol Processes

(GLOMAP-mode, Mann et al. (2010)) which is similar to the M7 framework. The main exception is that dust is currently calculated separately using 6 size bins. UKCA hence only considers 5 modes.

The global aerosol model MIROC-SPRINTARS uses the aerosol module SPRINTARS (Takemura et al., 2000, 2002; Takemura, 2005) in conjunction with the atmospheric general circulation model MIROC. SPRINTARS calculates the global evolution of three externally mixed aerosol species: ~~sulfate~~sulphate, sea salt and dust; and two species that can be either internally or externally mixed: organic matter and black carbon. In contrast to HAM and UKCA, SPRINTARS uses a single moment scheme and carries only mass as a prognostic variable. The size distributions of ~~sulfate~~sulphate, organic matter and black carbon are fixed but for sea salt and dust a bin scheme is used.

All three models were run for the year 2007, with nudged meteorology, prescribed sea surface temperatures and appropriate emission inventories (see Table 1).

3 Observations

We used the temporal sampling from several remote sensing datasets, both satellite and ground sites. The satellite sensor is the MODIS wide-swath (~ 2330 km) multi-channel imager onboard the Aqua satellite (local equator crossing time 1:30PM, repeated view period: 1 to 2 days depending on latitude). The ground sites constitute the AERONET network of sun-photometers that measure solar transmittance with nominally a 15^{min} time-resolution.

We chose not to use the original observations but datasets that have undergone further quality checks and spatio-temporal aggregation. It seems reasonable to study temporal sampling issues for datasets that have been optimised for actual model evaluations.

The satellite remote sensing data are the MODIS Aqua L3 NRL (Naval Research Laboratory) data (Zhang and Reid, 2006; Shi et al., 2011; Hyer et al., 2011) for 2007. These are based on official MODIS Aqua L2 Coll. 5 data (see Remer et al. (2005) for a discussion of the very similar over-ocean retrieval for Coll. 4 and ? for Coll. 5 over-land retrievals) that have been subjected to further quality checks, empirical corrections and spatial aggregation. The data comprises of AOT and its error estimate for 1° by 1° grid boxes at a 6-hourly resolution. We have analysed MODIS Terra as well but since it does not alter our conclusions, MODIS Terra observations will not be used in this paper.

We also use the MODIS Aqua L3 AORI (Atmosphere and Ocean Research Institute) data (Schutgens et al., 2013) for 2007. These are based on MODIS Aqua L2 data that have been subjected to further quality checks, empirical corrections and spatial aggregation. The AORI data differ from the NRL data as they use different empirical corrections and include both AOT and AE. Unlike the NRL data, AORI data is available only over ocean. In a preliminary analysis against Maritime Aerosol Network observations, we found that the AORI data suffered less from the high bias in AOT at low AOT values than the NRL data. The AORI AOT and AE data and its error estimates are available for 1° by 1° grid boxes at a 6-hourly resolution. MODIS Terra data will be ignored, again because it does not alter our conclusions.

Not applying the extra quality checks in the AORI product almost doubles the number of observations. While this does reduce the temporal sampling error somewhat (primarily in yearly averages), it does not alter our conclusions fundamentally. This also suggests that using Coll. 6 instead of Coll. 5 will have a minimal impact.

The other two remote sensing datasets are ground-based: AERONET Direct Sun lev 2.0 (Holben et al., 1998; Eck et al., 1999; Schmid et al., 1999) and AERONET Version 2 Inversions lev 2.0 (Dubovik and King, 2000; Dubovik et al., 2000) for 2007. Observations at each site were averaged over 6 hours, each 6 hours. AERONET data used in this study are AOT, AE and SSA.

As the MODIS-NRL product is provided on a regular 6-hourly grid, all other observations were similarly put on a regular 6-hourly grid. Using a 3-hourly grid for AERONET hardly affected our results. Moreover it was just as likely to (slightly) increase sampling errors as (slightly) decrease them.

4 Method

The 3-hourly model data were linearly interpolated to the locations of actual observations. These model values v were then sampled at the times of the actual observations to generate simulated observations.

The straight average

$$\bar{v} = N^{-1} \sum_{i=1}^{i=N} v_i, \tag{1}$$

(where i represents time and N the number of values over a day, a month or a year) is the average normally produced by models and often used in model evaluations.

The sampled average

$$\tilde{v} = \left(\sum_{i=1}^{i=N} f_i \right)^{-1} \sum_{i=1}^{i=N} f_i v_i, \tag{2}$$

is a model-simulated observation (daily, monthly or yearly mean) where f_i represent the observational sampling (taken from actual observational datasets):

$$f_i = \begin{cases} 0 & \text{if } \textit{no} \text{ observation present at time } i \\ 1 & \text{if observation present at time } i \end{cases} \tag{3}$$

The impact of not collocating model data with observations is now given by the error

$$\epsilon = \bar{v} - \widetilde{v}, \tag{4}$$

where we consider $\widetilde{v}$ the reference value. After all, actual observations are (hopefully) a proxy for the truth and used to assess models. Note that this error can be rewritten as

$$\epsilon = N^{-1} \sum_{i=1}^{i=N} \left(1 - \left(\frac{\sum_{j=1}^{j=N} f_j}{N} \right)^{-1} f_i \right) v_i. \tag{5}$$

The temporal sampling error depends on the statistics of the time-series v_i as well as the statistics of observations times f_i and in particular the amount of observational temporal coverage

$$C = \frac{\sum_{j=1}^{j=N} f_j}{N}. \tag{6}$$

5 Results

5.1 Does temporal sampling matter?

Fig. 1 shows yearly and zonally averaged AOT from both observations (MODIS-NRL Aqua) and models (either collocated with the observations or not). The difference between the straight model average (solid red line) and the sampled model average (dotted red line) shows the impact sampling has on model averages. The difference between the sampled model average (dotted red line) and observations (dotted black line) is due to model errors (and observational errors). Clearly, sampling errors can be as big as model errors. We also see that resampling the model improves agreement with some prominent features in the observations.

5.2 Temporal evolution of modelled AOT

To better understand the temporal sampling errors, we first need to consider the models' temporal evolution. Figure 2 shows the mean, standard deviation and their ratio (relative standard deviation) of the modelled AOT time-series. All three models agree in global mean AOT patterns (elevated AOT near natural sources like the Sahara & Southern Ocean and anthropogenic sources like Europe & East Asia), but there are substantial differences: HadGEM-UKCA shows lower mean AOT over the Sahara, ECHAM-HAM shows lower AOT over Europe and MIROC-SPRINTARS shows lower AOT over the Southern Ocean. Similarly, there are differences in standard deviation: HadGEM-UKCA has its largest standard deviation in AOT over East-Asia, ECHAM-HAM over the Sahara and East-Asia and MIROC-SPRINTARS over East-Asia and Eastern Europe. The relative standard deviation (standard deviation divided by the mean) also differs among these models, even though all three had their meteorology nudged to reanalysis data, suggesting there are fundamental differences in their time evolution due to the aerosol modelling itself. In particular, HadGEM-UKCA shows globally lower relative standard deviation than the other two models, while ECHAM-HAM and MIROC-SPRINTARS differ in the spatial patterns of the relative standard deviation. Regionally, relative standard deviation can easily differ by more than a factor of 2 among these three models.

AOT time series in our models are characterised by power-law power spectra with an exponent close to Kolmogorov's -5/3 constant, see Fig. 3 as an example. Sometimes this power spectrum flattens out into white noise at large periods due to synoptic-scale weather systems. In any case, AOT variability is largest for the longest periods. Diurnal cycles are not very strong for these models and their global patterns and magnitudes are very different (Fig. 4). As the modelled diurnal cycles are quite uncertain, so will our (later) conclusions regarding diurnal sampling errors.

We present a comparison of the (collocated) models against actual AOT observations as Taylor plots in Fig. 5. A Taylor plot (Taylor, 2001) graphically represents the variability of the model data and its correlation with observations and is especially suited to analyse the models' temporal evolution. Here we show them for four different regions: the entire Earth, the Amazon, East-Asia and Western-Africa. Mostly the models underestimate observed variability (the exception is MIROC-SPRINTARS which often overestimates it). Note that observational error has only a small impact on these statistics.

Assessing the model's diurnal cycle from remote sensing observations that require daylight is difficult, but a simple analysis that is appropriate to this paper's focus is to consider daily variation. This daily variation is not necessarily the result of a diurnal cycle but may also be caused by e.g. synoptic weather patterns. We compare the difference δ between maximum and minimum AOT over a day for all AERONET sites for both observations δ_o and models δ_m. Fig. 6 shows the median δ_o/δ_m for each AERONET site throughout the year, while Table 2 presents the median across all sites (in this case, medians provide a more conservative and lower estimate than means). Over a period of a day, observations tend to show more variability than the models. ~~Sensitivity studies (not shown~~ Various sensitivity studies (e.g. δ_o/δ_m as a function of δ_o as shown in Table 2) suggest that this is not due to observational error. It is likely due to at least two factors: first an underestimation of modelled AOT over the continents that limits absolute daily variation; secondly, modelled AOT time-series show very strong auto-correlations (> 0.8) over 6 hours, further reducing potential daily variation. Consequently, we suggest that these models underestimate daily variation in AOT.

Similarly, there are substantial differences in the time-series of AE and SSA among the three models (not shown). In particular, ECHAM-HAM shows a larger standard devi-

ation in AE and SSA over the continents than the other two models, while HadGEM-UKCA shows a very small standard deviation in SSA over the oceans.

5.3 Spatial patterns in temporal sampling errors

Yearly averaged MODIS-NRL Aqua AOT and its temporal coverage are shown in Fig. 7. Coverage is especially low over regions with high cloudiness and land areas with strong variations in surface albedo. In Fig. 8, we show yearly averages of the modelled AOT collocated with MODIS-NRL Aqua. The relative sampling error (Eq. 4) when using straight model averages is also shown. There are several things to note. First, temporal sampling errors can easily cause AOT to be over or under-estimated by as much as 50%. Secondly, these temporal sampling errors show substantial spatial correlation: entire regions show similar errors greatly reducing the possibility of reducing temporal sampling errors through spatial averaging (see e.g. Fig. 1). Thirdly, different models predict different error patterns. Finally, we point out that observational coverage only partly explains sampling errors. For instance, Fig. 7 suggests that India is far better sampled (temporally) than the Bay of Bengal, yet relative errors are larger in India, at least in the ECHAM-HAM and MIROC-SPRINTARS experiments.

Obviously, the temporal sampling error in yearly averages may be strongly impacted by several months of no observations at all. Over land, snow cover (combined with cloud cover) may preclude meaningful observations and e.g. Siberia is not observed at all during the Northern Hemisphere winter. Even over ocean, high latitudes may not be observed during part of the year due to low SZA (Solar Zenith Angle) or ice cover. We have inspected global maps of this temporal sampling error for monthly means, but found that errors can be just as large as for yearly means, due to the 1 to 2-day period between repeated views by MODIS and the possibility of extended cloud cover. In particular, for almost every location on Earth observed by MODIS, we found at least one month when temporal sampling errors exceeded 100%. A more in-depth analysis of the impact of averaging periods will follow later (Sect. 5.5).

Lastly, the absolute temporal sampling error in yearly averaged AE from the MODIS-AORI dataset is shown in Fig. 9. The observational coverage of AE is less than that of AOT, due to more stringent quality checks (not shown). Sampling errors rarely exceed ± 0.3. However, a very distinct spatial pattern can be seen for all three models: near coast-lines, especially in polluted outflow regions, the error tends to be strongly negative while over the remote oceans it tends to be strongly positive. If one does not properly colocate AE model data with observations, this would create an artificial contrast up to ~ 0.6 in AE between the continents and the remote oceans. The reason is that AE observations require a minimum AOT to be successful (see (Schutgens et al., 2013)). In areas with polluted outflows, low AOT correspond to the seasalt background with low AE and observations will mostly sample the outflow with high AE, leading to negative sampling errors. Conversely, over the remote oceans high wind-speeds lead to high AOT and low AE (larger particles) that will be more often observed than the lower AOT and higher AE (smaller particles) that occur at low wind-speeds.

The last paragraph suggests that minimum AOT requirements in retrievals can have large impacts on temporal sampling errors. While this was shown for the AORI dataset, already the official MODIS Coll. 5 product shows a strong dependence of the quality of AE retrievals on AOT values, as might be expected. See also Fig. 11 in Schutgens et al. (2013).

5.4 Drivers of temporal sampling errors

In Fig. 10, we show temporal sampling errors for monthly averaged AERONET AOT (similar plots can be shown for MODIS). Sampling errors increase with the standard deviation of the time-series, but decrease with increasing observational temporal coverage. Still, even for relatively high coverage large sampling errors are possible. We surmise that the clear positive bias is due to cases of high relative humidity that will tend to 1) prevent observations due to cloudiness; 2) increase model AOT due to wet growth. Randomising the observation times (instead of using actual AERONET times) in this analysis removes this bias and also reduces the spread in the errors by 15 to 45% (depending on the model). Thus it appears that correlations between the time-series and the observation times (see Eq. 5) also ~~affects~~ affect these sampling errors errors.

Similar figures for monthly-averaged AERONET AE and SSA are shown in Fig. 11. They confirm previous results and mostly serve to show the magnitude of the sampling errors that can possibly occur. The positive bias for SSA is likely due to ~~the minimum AOT (0.2)required for a proper inversion~~a minimum AOT in the AERONET Inversion datafiles (all inverted SSA values have AOT ≥ 0.2). As a result, the sampled model data will mostly sample pollution or dust events with low SSA while the straight model data also includes the aerosol background state with higher SSA.

5.5 Temporal sampling errors in context

To put these sampling errors in context, we will compare them to model errors and observational errors. The model errors are defined as the difference of collocated model values from actual observations. Strictly speaking this "model error" includes observational errors as well but unfortunately they can not be separated. However, estimates of this observational error's standard deviation are part of our observational datasets (we will assume this error distribution to be an unbiased Gaussian). The model error (minus the observa-

tional error) is the signal we are interested in when evaluating models.

Figure 12 shows a comparison of the above errors for yearly, monthly and daily averages of MODIS data (AOT and AE). For both AOT and AE, we see that observational errors and model errors decrease as averaging periods increase. In contrast, sampling errors behave differently, with the largest errors typically occurring for monthly averages and the smallest errors for daily averages. The models differ both in the sampling errors and how they compare to the model errors. But a few conclusions seem obvious: 1) for yearly and monthly averages, sampling errors will contribute considerably (typically 30-60%) to model errors (defined as the difference between a model and observations) if no collocation is performed; 2) sampling errors in AOT are comparable to or larger than observational errors for monthly or yearly averages; 3) for daily averages, these sampling errors appear to be smaller.

This last point is important as standard practice for AEROCOM models is to save data as daily averages. Doing so would incur sampling errors of at most 25% of the model error. There are a few caveats however. We have already pointed out that models seem to under-estimate AOT variability, especially on daily time-scales (see Sect. 5.2). Thus we are likely to underestimate daily sampling errors. Furthermore, the AEROCOM daily averages are defined according to UTC. This means that e.g. satellite observations (made close to local noon time) near the date-line are not as representative of the model day as observations near the Greenwich meridian. The consequence of this is shown in Fig. 13, where correlations between daily model averages and simulated observations clearly drop off towards the date-line. A box-whisker plot as in Fig. 12 also hides the fact that in large of parts of the world daily sampling errors are in magnitude at least 30% of the daily model errors (see Fig. 14). ~~Obviously, the sampling errors introduced by comparing~~ It is interesting to consider what temporal sampling errors remain if an annual average is constructed from daily model averages ~~to observationswill be strongly reduced when producing e. g. yearly averages from these datasets (not shown) , but any~~ and observations, excluding those days for which no observations exist. Figure 15 shows that this does indeed reduce sampling errors compared to the case of straight annual averages (as in Fig. 8) but we still find typical root mean square errors of 12, 7 and 17% for ECHAM-HAM, HadGEM-UKCA and MIROC-SPRINTARS. Interestingly, such a procedure results in mostly positive errors because daily minimum AOT (as modelled) occurs near the time of observation. Note also that the error patterns correlate somewhat with the diurnal cycles in these models, see Fig. 4.

Any analysis of model data ~~looking beyond a yearly average~~ attempting to look at shorter time-scales (e.g. scatter plots, time-series) will be ~~affected negatively~~ further negatively affected when using daily data. To prevent this, ~~and any sampling error discussed in this paper,~~ models need to be resampled to the observation times.

Figure 16 shows a comparison of the above errors for yearly, monthly and daily averages of AERONET data (AOT, AE and SSA). The overall picture is much the same as Fig. 12: model errors and observational errors decrease as averaging increases, while sampling errors are the largest for either yearly or monthly averages. Note that in the case of AERONET AOT, already for daily averages, sampling errors are larger than observational errors.

6 Summary

Although model data and observations usually have very different temporal sampling, researchers often assume this does not affect their analyses. Consequently, monthly or yearly means are compared without first temporally collocating the model data with the observations. We have assessed potential errors resulting from this practice and shown them to be significant. 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.

Our analysis is based on the temporal sampling of several oft used remote sensing datasets (MODIS, AERONET) and data from three different models: ECHAM-HAM, HadGEM-UKCA and MIROC-SPRINTARS. We define the temporal sampling error as the difference in yearly, monthly or daily means of the full model dataset and the sampled model datasets. Three different models were used as we found that current models differ in their temporal evolution even though yearly mean global patterns are fairly similar and meteorological fields were nudged to reanalysis data. Although this study is framed in the context of model evaluation, it has a clear and direct relevance to climatologies derived from observational datasets.

We find that temporal sampling errors in yearly and monthly means can amount to 100% in AOT, 0.4 in AE and 0.05 in SSA. In addition, these sampling errors are comparable to model errors (30–60% for MODIS, 20–90% for AERONET). While daily averages incur smaller sampling errors, we argue (based on observations) that our models very likely underestimate daily variation and hence sampling errors. In addition, model daily averages are defined through UTC while remote sensing observations are only possible during local daylight. This introduces a marked longitudinal dependence in temporal sampling errors. ~~Such sampling errors~~ Annual averages constructed from daily averages still exhibit temporal sampling errors although at reduced levels (we estimate typical errors of 7-17%).

Temporal sampling errors will affect model evaluation of not only AOT, AE and SSA but also derived properties like direct aerosol radiative forcing, if data is not properly collocated with the observations.

Our analysis should also provide caution to researchers using in-situ data from e.g. ground sites. While such data often has nominally high measurement frequencies (hours), significant gaps (days to weeks) in temporal coverage may be present. Much will depend on the observational sampling and the nature of the observed time-series. For instance, PM2.5 filter measurements by EMEP [1] (European Monitoring and Evaluation Network, http://www.emep.int) are time-integrated measurements that can probably be compared to models without temporal sampling issues. But several EUSAAR [1] (European Supersites for Atmospheric Aerosol Research, http://www.eusaar.net) stations have significant intermittency in their observations. Flight campaign data obviously require temporal collocation to make any sense in a model evaluation context. Also, unlike remote sensed column properties like AOT, in-situ measurements will be affected by vertical transport. This may possibly exacerbate any temporal sampling issues.

As a corollary to our study, we have shown that the temporal evolution of models can be very different (although yearly means are quite similar). We suggest that more effort should be made to understand modelled and observed time-series. It is likely that such time-series contain finger prints of individual aerosol processes and provide new ways to evaluate models. In particular we point out the large differences we found in diurnal cycles (see Fig. 4) and the research opportunities that may lie in observational datasets (e.g. from geostationary satellites) that resolve this diurnal cycle.

Finally, if there is one error in the field of aerosol modelling over which we have direct control and that can be confidently eliminated, it is the error due to incongruous temporal sampling of model data and observations. Not doing so barely saves time or resources but risks compromising any model evaluation. For instance, we have shown that the latitudinal pattern of yearly, zonal averages of collocated model AOT agrees better with MODIS observations than the normal yearly, zonal model average. The collocation of model data with observations can either be done on-line through observation simulators or off-line through sampling of high frequency model output. A generic tool to support the latter operation is freely available as the Community Intercomparison Suite (http://www.cistools.net).

Acknowledgements. This work was supported by the Natural Environmental Research Council grant nr NE/J024252/1 (Global Aerosol Synthesis And Science Project). Computational resources for the ECHAM-HAM runs were made available by Deutsches Klimarechenzentrum (DKRZ) through support from the Bundesministerium für Bildung und Forschung (BMBF). The ECHAM-HAMMOZ model is developed by a consortium composed of ETH Zurich, Max Planck Institut für Meteorologie, Forschungszentrum Jülich, University of Oxford, the Finnish Meteorological Institute and the Leibniz Institute for Tropospheric Research, and managed by the Center for Climate Systems Modeling (C2SM) at ETH Zurich. P. Stier would like to acknowledge funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) ERC project ACCLAIM (grant agreement no. FP7- 280025). HadGEM-UKCA was run on the ARCHER UK National Supercomputing Service (http://www.archer.ac.uk). The development of the UKCA model (www.ukca.ac.uk) was supported by the UK's Natural Environment Research Council (NERC) through the NERC Centres for Atmospheric Science (NCAS) initiative. MIROC-SPRINTARS was run on the SX-9 supercomputer at NIES (CGER) in Japan. The figures in this paper were prepared using David W. Fanning's coyote library for IDL. The authors thank an anonymous reviewer and in particular Dr Andrew Sayer for useful comments that helped improve the manuscript.

[1] European Monitoring and Evaluation Network, http://www.emep.int

[1] European Supersites for Atmospheric Aerosol Research, http://www.eusaar.net

References

Albrecht, B. A.: Aerosols, cloud microphysics, and fractional cloudiness, Science, 245, 1227–1230, 1989.

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.

Angstrom, B. A.: Atmospheric turbidity , global illumination and planetary albedo of the earth, Tellus, XIV, 435–450, 1962.

Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z. J., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Fischer, P., Nieuwenhuijsen, M., Vineis, P., Xun, W. W., Katsouyanni, K., Dimakopoulou, K., Oudin, A., Forsberg, B., Modig, L., Havulinna, A. S., Lanki, T., Turunen, A., Oftedal, B., Nystad, W., Nafstad, P., De Faire, U., Pedersen, N. L., Östenson, C.-G., Fratiglioni, L., Penell, J., Korek, M., Pershagen, G., Eriksen, K. T., Overvad, K., Ellermann, T., Eeftens, M., Peeters, P. H., Meliefste, K., Wang, M., Bueno-de Mesquita, B., Sugiri, D., Krämer, U., Heinrich, J., de Hoogh, K., Key, T., Peters, A., Hampel, R., Concin, H., Nagel, G., Ineichen, A., Schaffner, E., Probst-Hensch, N., Künzli, N., Schindler, C., Schikowski, T., Adam, M., Phuleria, H., Vilier, A., Clavel-Chapelon, F., Declercq, C., Grioni, S., Krogh, V., Tsai, M.-Y., Ricceri, F., Sacerdote, C., Galassi, C., Migliore, E., Ranzi, A., Cesaroni, G., Badaloni, C., Forastiere, F., Tamayo, I., Amiano, P., Dorronsoro, M., Katsoulis, M., Trichopoulou, A., Brunekreef, B., and Hoek, G.: Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project, The Lancet, 6736, 1–11, doi:10.1016/S0140-6736(13)62158-3, http://linkinghub.elsevier.com/retrieve/pii/S0140673613621583, 2013.

Bellouin, N., Mann, G. W., Woodhouse, M. T., Johnson, C., Carslaw, K. S., and Dalvi, M.: Impact of the modal aerosol scheme GLOMAP-mode on aerosol forcing in the hadley centre global environmental model, Atmospheric Chemistry and Physics, 13, 3027–3044, doi:10.5194/acp-13-3027-2013, 2013.

Brunekreef, B. and Holgate, S. T.: Air pollution and health., Lancet, 360, 1233–42, doi:10.1016/S0140-6736(02)11274-8, http://www.ncbi.nlm.nih.gov/pubmed/12401268, 2002.

Dockery, D., Pope, A., Xu, X., Spengler, J., Ware, J., Fay, M., Ferris, B., and Speizer, F.: An association between air pollution and mortality in six U.S. cities, The New England Journal of Medicine, 329, 1753–1759, 1993.

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, doi: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, doi: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.

Ezzati, M., Lopez, A. D., Rodgers, A., Vander Hoorn, S., and Murray, C. J. L.: Selected major risk factors and global and regional burden of disease., Lancet, 360, 1347–60, doi:10.1016/S0140-6736(02)11403-6, http://www.ncbi.nlm.nih.gov/pubmed/12423980, 2002.

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, doi:10.1002/qj.2305, http://doi.wiley.com/10.1002/qj.2305, 2014.

Hansen, J., Sato, M., and Ruedy, R.: Radiative forcing and climate response, Journal of Geophysical Research, 102, 6831–6864, 1997.

Hewitt, H. T., Copsey, D., Culverwell, I. D., Harris, C. M., Hill, R. S. R., Keen, a. B., McLaren, a. J., and Hunke, E. C.: Design and implementation of the infrastructure of HadGEM3: the next-generation Met Office climate modelling system, Geoscientific Model Development, 4, 223–253, doi:10.5194/gmd-4-223-2011, http://www.geosci-model-dev.net/4/223/2011/, 2011.

Holben, B. N., Slutsker, T. I. E. I., Tar, D., Buis, J. P., Setxerj, I. I. A., Reagan, A., J, Y., Nakajima, T., Lavenu, F., Vemte, E., Jankowiak, I., and Smirnozjt, A.: AERONET-A Federated Instrument Network and Data Archive for Aerosol Characterization, Remote Sensing of Environment, 66, 1–16, 1998.

Huneeus, N., Schulz, M., Balkanski, Y., Griesfeller, J., Prospero, J., Kinne, S., Bauer, S., Boucher, O., Chin, M., Dentener, F., Diehl, T., Easter, R., Fillmore, D., Ghan, S., Ginoux, P., Grini, A., Horowitz, L., Koch, D., Krol, M. C., Landing, W., Liu, X., Mahowald, N., Miller, R., Morcrette, J.-J., Myhre, G., Penner, J., Perlwitz, J., Stier, P., Takemura, T., and Zender, C. S.: Global dust model intercomparison in AeroCom phase I, Atmospheric Chemistry and Physics, 11, 7781–7816, doi:10.5194/acp-11-7781-2011, http://www.atmos-chem-phys.net/11/7781/2011/, 2011.

Hyer, E. J., Reid, J. S., and Zhang, J.: An over-land aerosol optical depth data set for data assimilation by filtering , correction , and aggregation of MODIS Collection 5 optical depth retrievals, Atmos. Meas. Tech., 4, 379–408, doi:10.5194/amt-4-379-2011, 2011.

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., Schulz, M., Textor, C., Guibert, S., Balkanski, Y., Bauer, S. E., Berntsen, T., Berglen, T. F., and Boucher, O.: An AeroCom initial assessment – optical properties in aerosol component modules of global models, Atmospheric Chemistry and Physics, 6, 1815–1834, 2006.

Koch, D., Schulz, M., Kinne, S., Mcnaughton, C., Spackman, J. R., Balkanski, Y., Bauer, S., Berntsen, T., Bond, T., Boucher, O., Chin, M., Clarke, A., De Luca, N., Dentener, F., Diehl, T., Dubovik, O., Easter, R., Fahey, D., Feichter, J., Fillmore, D., Freitag, S., Ghan, S., Ginoux, P., Gong, S., Horowitz, L., Iversen, T., Kirkevag, A., Klimont, Z., Kondo, Y., Krol, M., Liu, X., Milelr, R., Montanaro, V., Moteki, N., Myhre, G., Penner, J., Perlwitz, J., Pitari, G., Reddy, S., Sahu, L., Sakamoto, H., Schuster, G., Schwarz, J., Seland, O., Stier, P., Takegawa, N., Takemura, T., Textor, C., van Aardenne, J., and Zhao, Y.: Evaluation of black carbon estimations in global aerosol models, Atmospheric Chemistry and Physics, 9, 9001–9026, 2009.

Koffi, B., Schulz, M., Bréon, F.-M., Griesfeller, J., Winker, D., Balkanski, Y., Bauer, S., Berntsen, T., Chin, M., Collins, W. D., Dentener, F., Diehl, T., Easter, R., Ghan, S., Ginoux, P., Gong, S., Horowitz, L. W., Iversen, T., Kirkevå g, A., Koch, D., Krol, M., Myhre, G., Stier, P., and Takemura, T.: Application of the CALIOP layer product to evaluate the vertical distribution of aerosols estimated by global models: AeroCom phase I results, Journal of Geophysical Research, 117, D10 201, doi:10.1029/2011JD016858, http://doi.wiley.com/10.1029/2011JD016858, 2012.

Lequy, E., Conil, S., and Turpault, M.-P.: Impacts of Aeolian dust deposition on European forest sustainability: A review, Forest Ecology and Management, 267, 240–252, doi:10.1016/j.foreco.2011.12.005, http://linkinghub.elsevier.com/retrieve/pii/S0378112711007365, 2012.

Lin, S.-J. and Rood, R. B.: Multidimensional flux-form semi-Lagrangian transport schemes, Monthly Weather Review, 124, 2046–2070, 1996.

Lohmann, U. and Feichter, J.: Impact of sulfate aerosols on albedo and lifetime of clouds : A sensitivity study with the ECHAM4 GCM, Journal of Geophysical Research, 102, 13,685–13,700, 1997.

Lohmann, U. and Feichter, J.: Global indirect aerosol effects : a review, Atmospheric Chemistry and Physics, 5, 715–737, 2005.

Maher, B., Prospero, J., Mackie, D., Gaiero, D., Hesse, P., and Balkanski, Y.: Global connections between aeolian dust, climate and ocean biogeochemistry at the present day and at the last glacial maximum, Earth-Science Reviews, 99, 61–97, doi:10.1016/j.earscirev.2009.12.001, http://linkinghub.elsevier.com/retrieve/pii/S0012825210000024, 2010.

Mann, G. W., Carslaw, K. S., Spracklen, D. V., Ridley, D. a., Manktelow, P. T., Chipperfield, M. P., Pickering, S. J., and Johnson, C. E.: Description and evaluation of GLOMAP-mode: a modal global aerosol microphysics model for the

UKCA composition-climate model, Geoscientific Model Development, 3, 519–551, doi:10.5194/gmd-3-519-2010, http://www.geosci-model-dev.net/3/519/2010/, 2010.

McTainsh, G. and Strong, C.: The role of aeolian dust in ecosystems, Geomorphology, 89, 39–54, doi:10.1016/j.geomorph.2006.07.028, http://linkinghub.elsevier.com/retrieve/pii/S0169555X06003564, 2007.

Quaas, J., Ming, Y., Menon, S., Takemura, T., Wang, M., Penner, J. E., Gettelman, A., Lohmann, U., Bellouin, N., Boucher, O., Sayer, a. M., Thomas, G. E., McComiskey, A., Feingold, G., Hoose, C., Kristjánsson, J. E., Liu, X., Balkanski, Y., Donner, L. J., Ginoux, P. a., Stier, P., Feichter, J., Sednev, I., Bauer, S. E., Koch, D., Grainger, R. G., Kirkevåg, A., Iversen, T., Seland, O., Easter, R., Ghan, S. J., Rasch, P. J., Morrison, H., Lamarque, J.-F., Iacono, M. J., Kinne, S., and Schulz, M.: Aerosol indirect effects – general circulation model intercomparison and evaluation with satellite data, Atmospheric Chemistry and Physics Discussions, 9, 12 731–12 779, doi:10.5194/acpd-9-12731-2009, 2009.

Remer, L., Kaufman, Y., Tanre, D., Mattoo, S., Chu, D., Martins, J., Li, R.-R., Ichoku, C., Levy, R., Kleidman, R., Eck, T., Vermote, E., and Holben, B.: The MODIS Aerosol Algorithm, Products, and Validation, J. Atmospheric Sciences, 62, 947 – 973, 2005.

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, doi:10.1109/LGRS.2006.879562, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1715312, 2006.

Roeckner, E., Bäuml, G., Bonaventura, L., Brokopf, R., Esch, M., Giorgetta, M., Hagemann, S., Kirchner, I., Kornblueh, L., Manzini, E., Rhodin, A., Schlese, U., Schulzweida, U., and Tompkins, A.: The atmospheric general circulation model ECHAM5, part I: model description, Tech. Rep. 349, Max Planck Institute for Meteorology, Hamburg, 2003.

Roeckner, E., Rokopf, R. B., Sch, M. E., Iorgetta, M. G., and Agemann, S. H.: Sensitivity of Simulated Climate to Horizontal and Vertical Resolution in the ECHAM5 Atmosphere Model, Journal of Climate, 19, 3771–3791, 2006.

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, doi: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.

Schulz, M., Textor, C., Kinne, S., Balkanski, Y., Bauer, S., Berntsen, T., Berglen, T., Boucher, O., and Dentener, F.: Radiative forcing by aerosols as derived from the AeroCom present-day and pre-industrial simulations, Atmospheric Chemistry and Physics, 6, 5225–5246, 2006.

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

Shi, Y., Zhang, J., Reid, J. S., Holben, B., Hyer, E. J., and Curtis, C.: An analysis of the collection 5 MODIS over-ocean aerosol optical depth product for its implication in aerosol assimilation, Atmospheric Chemistry and Physics, 11, 557–565, doi:10.5194/acp-11-557-2011, http://www.atmos-chem-phys.net/11/557/2011/, 2011.

Smith, K. R., Jerrett, M., Anderson, H. R., Burnett, R. T., Stone, V., Derwent, R., Atkinson, R. W., Cohen, A., Shonkoff, S. B., Krewski, D., Pope, C. A., Thun, M. J., and Thurston, G.: Public health benefits of strategies to reduce greenhouse-gas emissions: health implications of short-lived greenhouse pollutants., Lancet, 374, 2091–103, doi:10.1016/S0140-6736(09)61716-5, http://www.ncbi.nlm.nih.gov/pubmed/19942276, 2009.

Stier, P., Feichter, J., Kinne, S., Kloster, S., Vignati, E., Wilson, J., Ganzeveld, L., Tegen, I., and Werner, M.: The aerosol-climate model ECHAM5-HAM, Atmospheric Chemistry and Physics, 5, 1125–1156, 2005.

Stier, P., Seinfeld, J. H., Kinne, S., and Boucher, O.: Aerosol absorption and radiative forcing, Atmospheric Chemistry and Physics, 7, 5237–5261, 2007.

Takemura, T.: Simulation of climate response to aerosol direct and indirect effects with aerosol transport-radiation model, Journal of Geophysical Research, 110, doi:10.1029/2004JD005029, http://doi.wiley.com/10.1029/2004JD005029, 2005.

Takemura, T., Okamoto, H., Maruyama, Y., Numaguti, A., Higurashi, A., and Nakajima, T.: Global three-dimensional simulation of aerosol optical thickness distribution of various origins, Journal of Geophysical Research, 105, 17 853, doi:10.1029/2000JD900265, http://doi.wiley.com/10.1029/2000JD900265, 2000.

Takemura, T., Nakajima, T., Dubovik, O., Holben, B. N., and Kinne, S.: Single-Scattering Albedo and Radiative Forcing of Various Aerosol Species with a Global Three-Dimensional Model, J. of Climate, 15, 333–352, 2002.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophysical Research, 106, 7183–7192, 2001.

Textor, C., Schulz, M., Guibert, S., Kinne, S., Balkanski, Y., Bauer, S., Berntsen, T., Berglen, T., Boucher, O., Chin, M., Dentener, F., Diehl, T., Easter, R., Feichter, H., Fillmore, D., Ghan, S., Ginoux, P., Gong, S., Grini, A., Hendricks, J., Horowitz, L., Huang, P., Isaksen, I., Iversen, I., Kloster, S., Koch, D., Kirkevå g, A., Kristjansson, J. E., Krol, M., Lauer, A., Lamarque, J. F., Liu, X., Montanaro, V., Myhre, G., Penner, J., Pitari, G., Reddy, S., Seland, O., Stier, P., Takemura, T., and Tie, X.: Analysis and quantification of the diversities of aerosol life cycles within AeroCom, Atmospheric Chemistry and Physics, 6, 1777–1813, doi:10.5194/acp-6-1777-2006, http://www.atmos-chem-phys.net/6/1777/2006/, 2006.

Textor, C., Schulz, M., Guibert, S., Kinne, S., Balkanski, Y., Bauer, S., Berntsen, T., Berglen, T., Boucher, O., Chin, M., Dentener, F., Diehl, T., Feichter, J., Fillmore, D., Ginoux, P., Gong, S., Grini, A., Hendricks, J., Horowitz, L., Huang, P., Isaksen, I. S. a., Iversen, T., Kloster, S., Koch, D., Kirkevå g, A., Kristjansson, J. E., Krol, M., Lauer, A., Lamarque, J. F., Liu, X., Montanaro, V., Myhre, G., Penner, J. E., Pitari, G., Reddy, M. S., Seland, O., Stier, P., Takemura, T., and Tie, X.:

The effect of harmonized emissions on aerosol properties in global models – an AeroCom experiment, Atmospheric Chemistry and Physics, 7, 4489–4501, doi:10.5194/acp-7-4489-2007, http://www.atmos-chem-phys.net/7/4489/2007/, 2007.

Twomey, S.: Pollution and the planetary albedo, Atmospheric Environment, 8, 1251–1256, 1974.

Vignati, E., Wilson, J., and Stier, P.: M7: An efficient size-resolved aerosol microphysics module for large-scale aerosol transport models, Journal of Geophysical Research, 109, D22 202, doi:10.1029/2003JD004485, http://doi.wiley.com/10.1029/2003JD004485, 2004.

Vink, S. and Measures, C.: The role of dust deposition in determining surface water distributions of Al and Fe in the South West Atlantic, Deep Sea Research Part II, 48, 2787–2809, doi:10.1016/S0967-0645(01)00018-2, http://linkinghub.elsevier.com/retrieve/pii/S0967064501000182, 2001.

Zhang, J. and Reid, J. S.: MODIS aerosol product analysis for data assimilation: Assessment of over-ocean level 2 aerosol optical thickness retrievals, Journal of Geophysical Research, 111, D22 207, doi:10.1029/2005JD006898, http://doi.wiley.com/10.1029/2005JD006898, 2006.

Zhang, K., O'Donnell, D., Kazil, J., Stier, P., Kinne, S., Lohmann, U., Ferrachat, S., Croft, B., Quaas, J., Wan, H., Rast, S., and Feichter, J.: The global aerosol-climate model ECHAM-HAM, version 2: sensitivity to improvements in process representations, Atmospheric Chemistry and Physics, 12, 8911–8949, doi:10.5194/acp-12-8911-2012, http://www.atmos-chem-phys.net/12/8911/2012/, 2012.

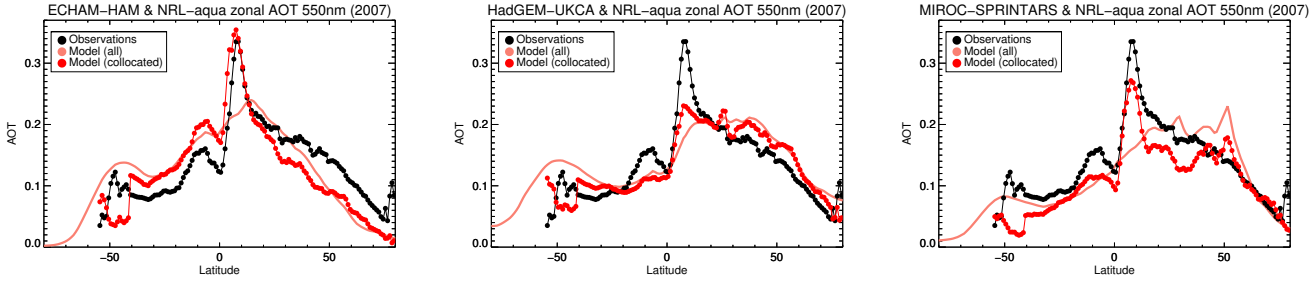


Figure 1. Yearly and zonal averages of AOT from the observations (MODIS-NRL Aqua) and the models. The latter are averaged in two different ways: using either all model data or only those sampled at the observation times.

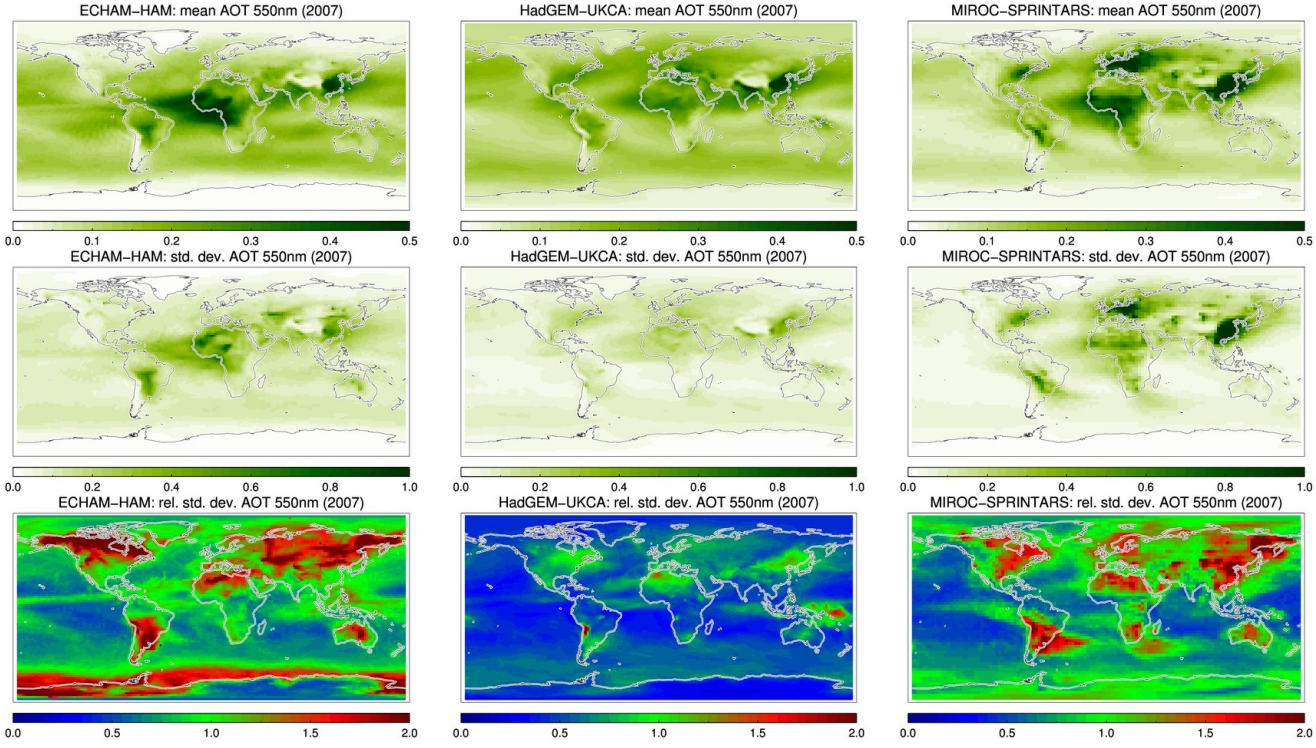


Figure 2. Global maps of yearly mean, standard deviation, and relative standard deviation in AOT.

Table 1. Models used in this study.

	ECHAM-HAM	HadGEM-UKCA	MIROC-SPRINTARS
version	ECHAM6.1-HAM2.2	HadGEM3-A-GLOMAP UKCA 8.4	MIROC3.2-SPRINTARS3.84
resolution	$1.9^o \times 1.9^o$, 31 levels (T63L31)	$1.9^o \times 1.25^o$, 85 levels (N96L85)	$2.8^o \times 2.8^o$, 20 levels (T42L20)
reanalysis	ERA-interim	ERA-interim	NCEP
natural emissions	model time-step	model time-step	model time-step
anthropogenic emissions	monthly	monthly	monthly
wildfire emissions	daily	monthly	monthly

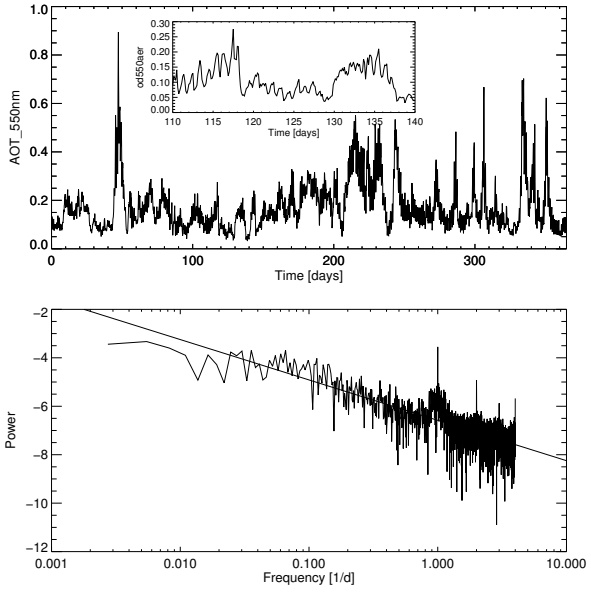


Figure 3. AOT time-series and its associated power spectrum at 75W, 0N in HadGEM-UKCA. For this power spectrum, the ratio of power at 1 day over total power is ~ 0.01. The straight line in the lower plot has a slope of -5/3.

Table 2. Median ratio of daily AOT variation δ_o/δ_m across all AERONET sites. Only sites with more than two observations for at least 24 days during 3 months were used. Since the estimated error in AERONET AOT measurements is 0.01, the error in measured δ_o can be no more than $\sqrt{2*0.01^2}$.

δ_o	ECHAM-HAM	HadGEM-UKCA	MIROC-SPRINTARS
≥ 0	2.48	2.36	2.08
$\geq \sqrt{2*0.01^2}$	3.66	3.75	2.95
$\geq 2\sqrt{2*0.01^2}$	4.19	4.95	3.35

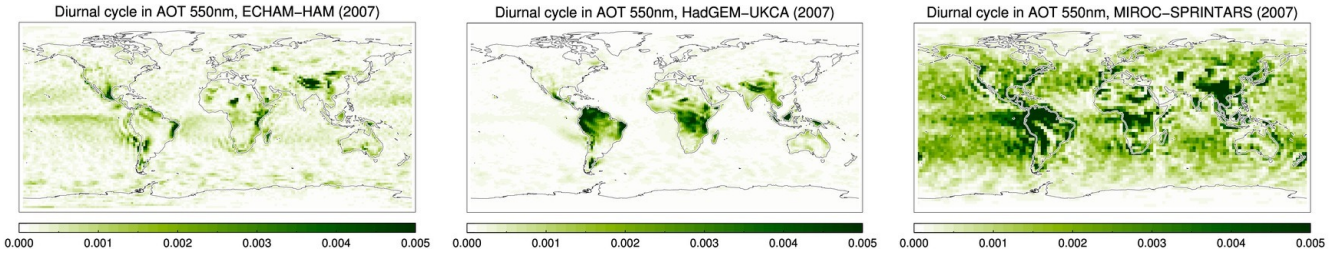


Figure 4. Global maps of the relative importance of the diurnal cycle: the ratio of power at a period of 1 day to total power for AOT time-series.

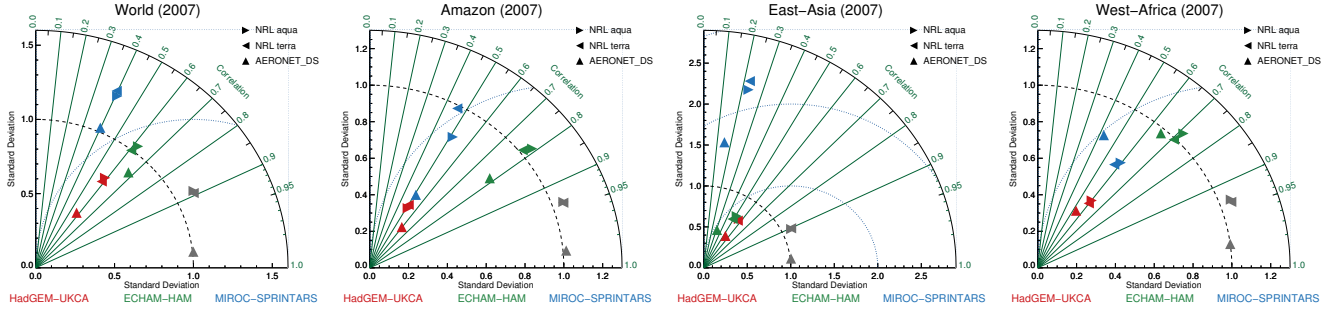


Figure 5. Taylor plots showing the standard deviation and correlation of modelled AOT against MODIS-NRL and AERONET observations. The dark grey symbols refer to observations randomly perturbed with estimated observational errors.

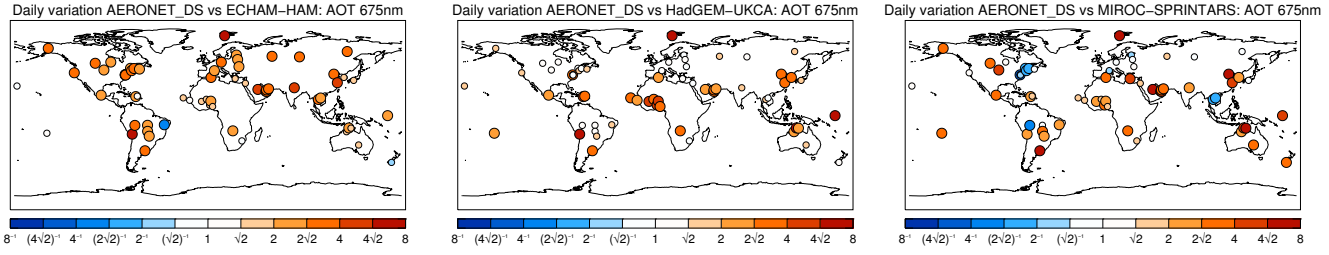


Figure 6. Median ratio of daily difference in maximum and minimum AOT δ_o/δ_m, either observed or modelled, for all AERONET sites. Smaller circles indicate sites where $0.5 \leq$ median $\delta_o/\delta_m \leq 2$. Only sites with more than two observations per day for at least 24 days during 3 or more separate months were used.

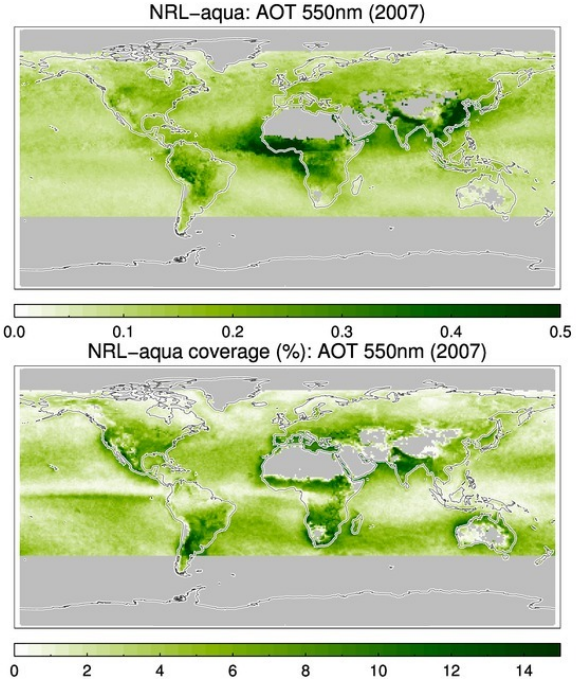


Figure 7. Top panel shows the 2007 yearly mean AOT of MODIS-NRL Aqua. Bottom panel shows the temporal coverage (%) of observations during that year. Grey colors indicate no observations available at all.

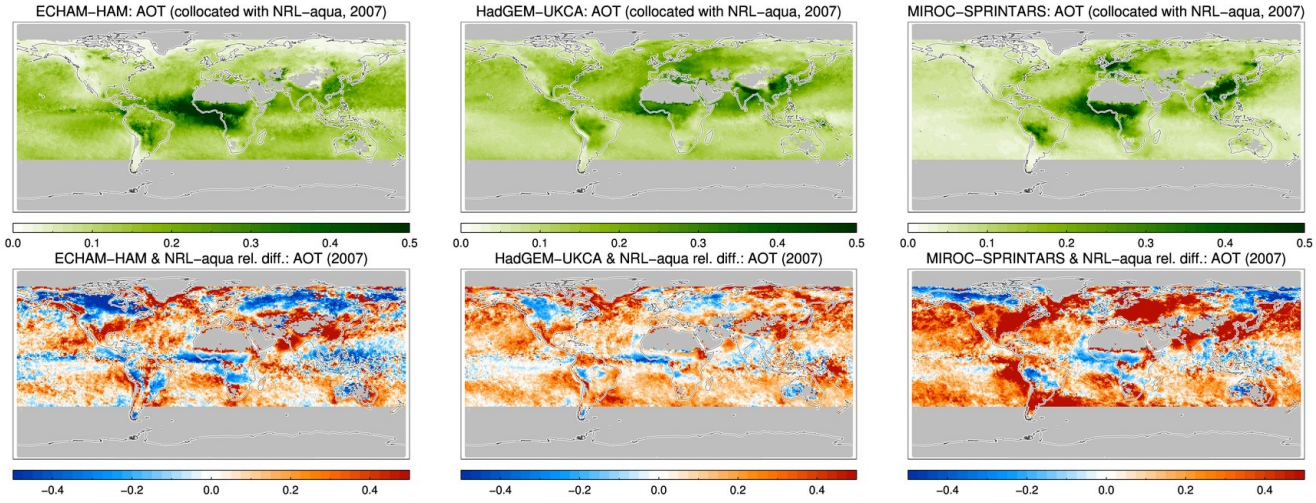


Figure 8. Top row shows global map of yearly mean of model AOT collocated with MODIS-NRL Aqua AOT data. Bottom row shows the relative temporal sampling error using MODIS-NRL Aqua AOT sampling. Grey colors indicate no observations available.

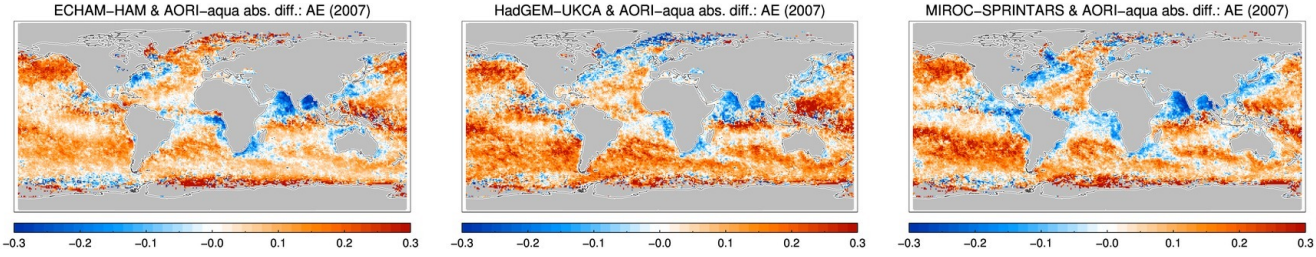


Figure 9. Global map of absolute temporal sampling error using MODIS-AORI Aqua AE sampling. Grey colors indicate no observations available.

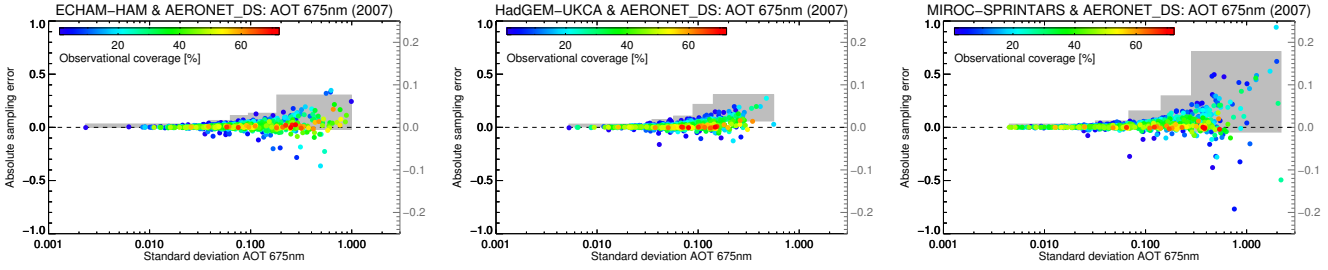


Figure 10. Absolute temporal sampling error for monthly mean AOT using AERONET sampling, as a function of time-series standard deviation and observational coverage. The grey boxes in the background (axis on the right), show the 25% and 75% quantiles of those errors for equally sized sub-samples.

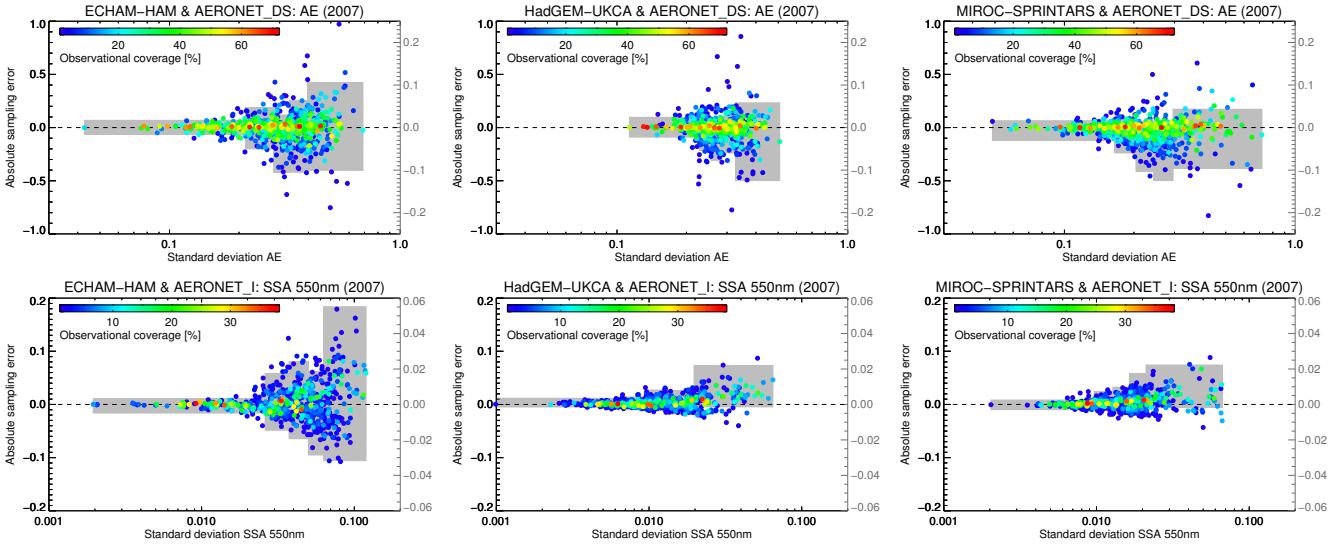


Figure 11. Absolute temporal sampling error for monthly mean AE and SSA using AERONET sampling, as a function of time-series standard deviation and observational coverage. The grey boxes in the background (axis on the right), show the 25% and 75% quantiles of those errors for equally sized sub-samples.

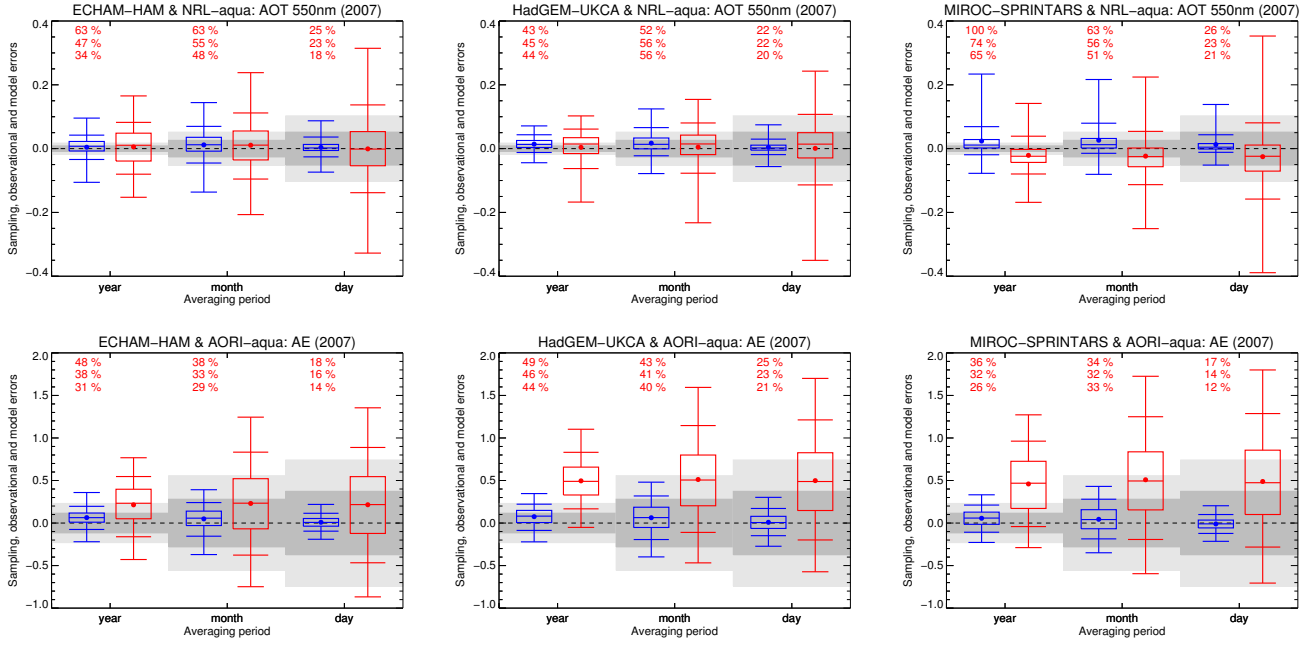


Figure 12. Box-whisker plots of temporal sampling errors, observational and model errors. For MODIS Aqua, sampling errors (blue), observational errors (grey) and model errors (red) are shown side by side as a function of averaging period (yearly, monthly, daily). The box-whisker plot shows the 2%, 9%, 25%, 50%, 75%, 91% and 98% quantiles of these errors, the dot shows the mean error. The grey shading shows the 9%, 25%, 75% and 91% quantiles of the observational errors. The red numbers above each blue box-whisker show how the different quantile ranges (from top to bottom: 2-98%, 9-91% and 25-75%) compare between the sampling error and the model errors.

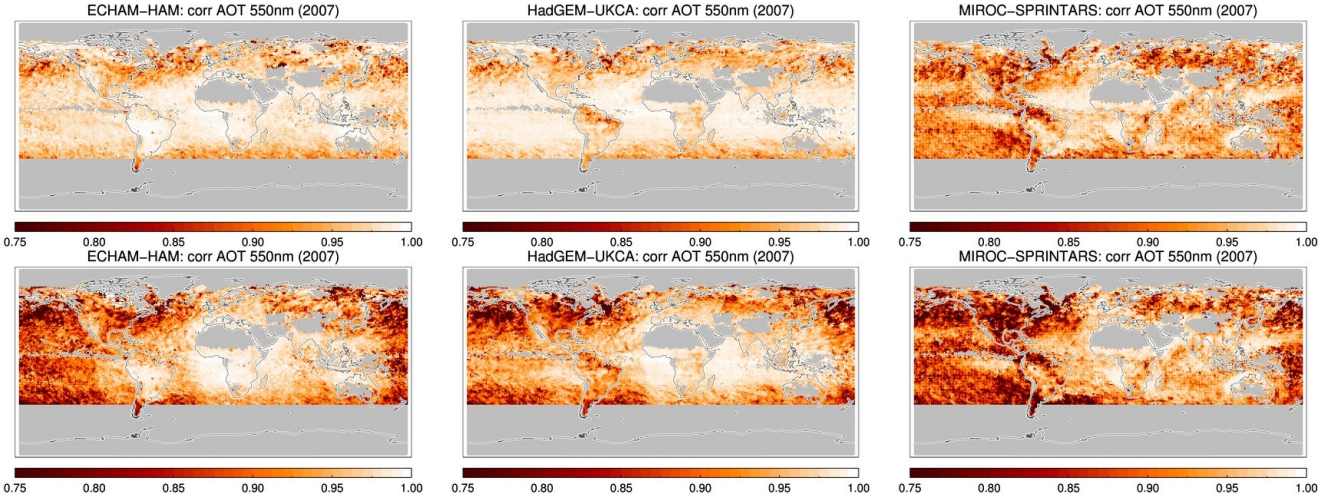


Figure 13. Correlation between model AOT at the time of MODIS-NRL Aqua observation and a 24-hour model AOT average. Top: 24-hour model average uses overpass time as the centre for the 24-hour average; bottom: 24-hour average defined according to UTC. Note how in the bottom plot, correlations drop off towards the date-line.

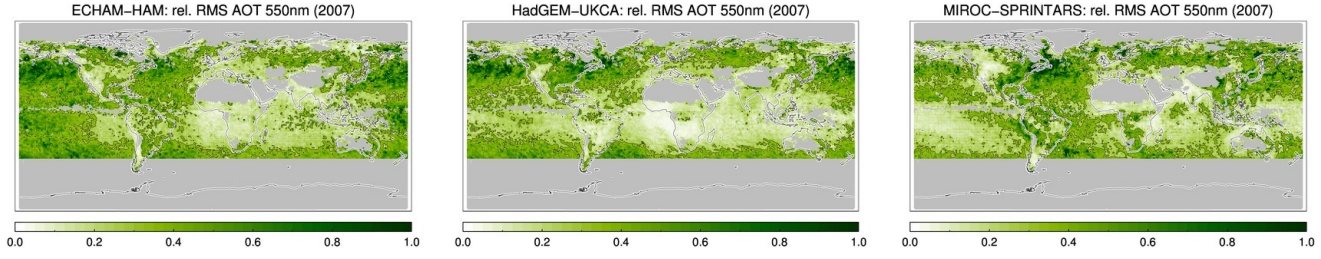


Figure 14. Ratio of the RMS (root mean squared) temporal sampling daily error to RMS model daily error, i.c. MODIS-NRL Aqua. The daily errors were defined according to UTC. The thin black contour represents a ratio of 0.3.

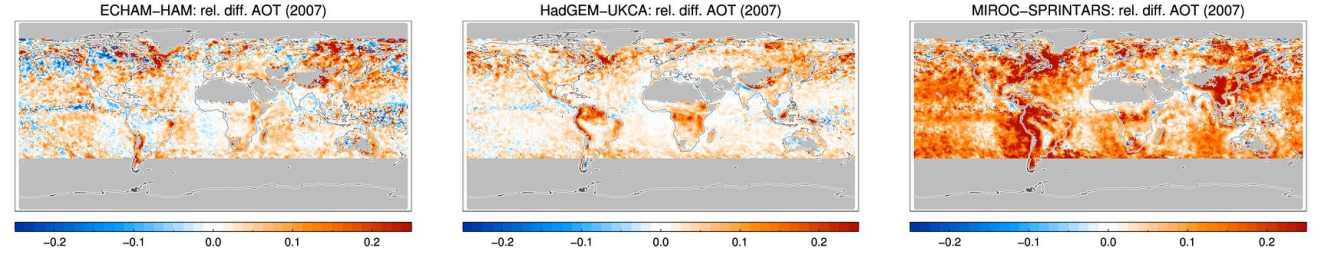


Figure 15. Relative temporal sampling error using MODIS-NRL Aqua AOT sampling. The yearly averages are constructed from daily data, excluding days for which no observations were present. This figure can be compared to the bottom row of Fig. 8 (but note that the colour bar has half its range). Grey colors indicate no observations available.

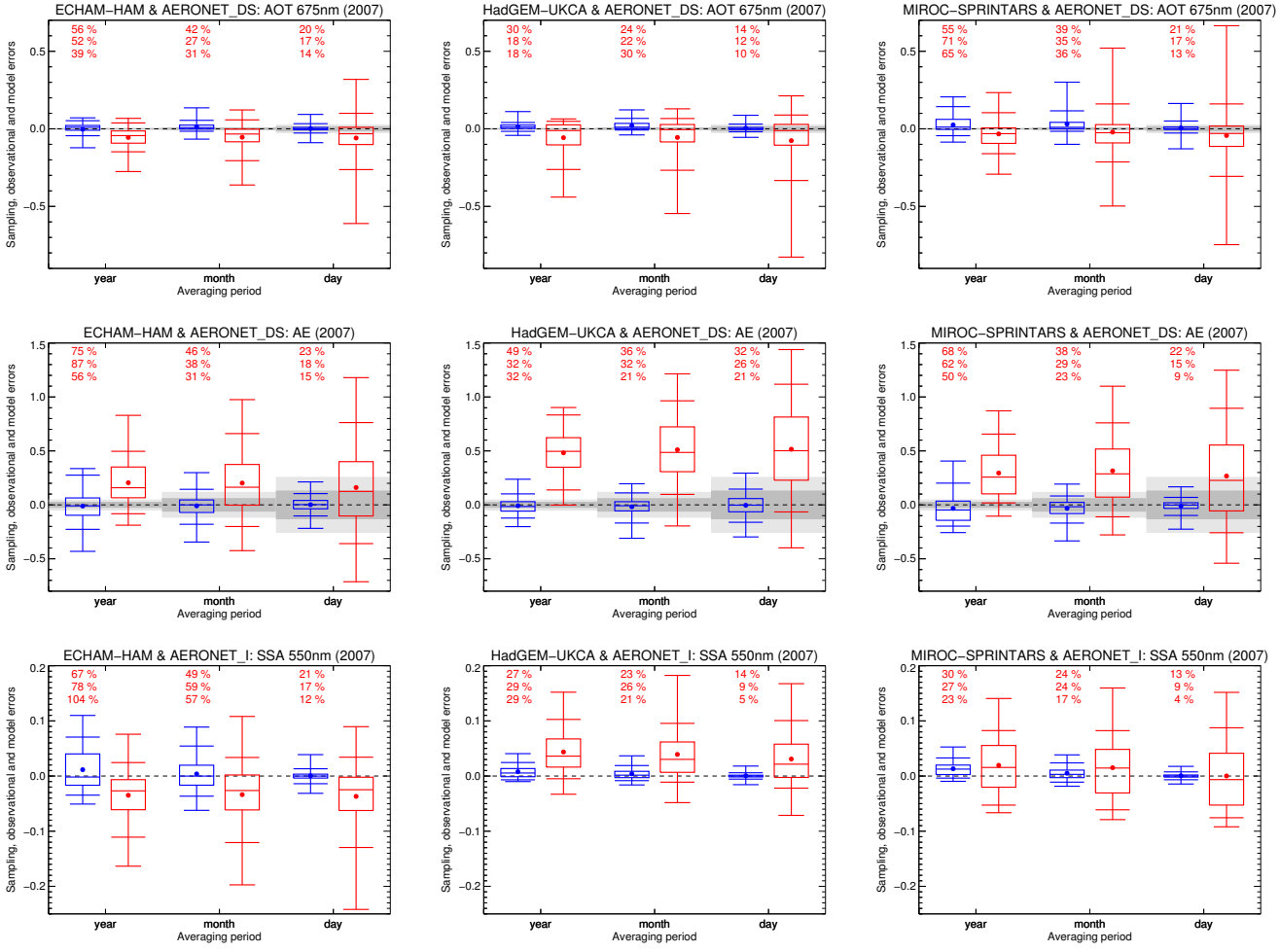


Figure 16. Box-whisker plots of temporal sampling errors, observational and model errors. For AERONET, sampling errors (blue), observational errors (grey) and model errors (red) are shown side by side as a function of averaging period (yearly, monthly, daily). The box-whisker plot shows the 2%, 9%, 25%, 50%, 75%, 91% and 98% quantiles of these errors, the dot shows the mean error. The grey shading shows the 9%, 25%, 75% and 91% quantiles of the observational errors. The red numbers above each blue box-whisker show how the different quantile ranges (from top to bottom: 2-98%, 9-91% and 25-75%) compare between the sampling error and the model errors. For SSA, no observational error estimates were available.