We thank our two reviewers for carefully reading our manuscript and supplying us with useful comments and constructive criticism.

Response to reviewer 1

Main comments

Sections 4 and 9: I may not be as clever as the average Atmos. Chem. Phys. reader, but I have difficulties understanding why errors increase with distance to the grid-point of the 210x210 box. If I understand section 4 and Figure 2 correctly, if w = 1, all observations (the 10x10 boxes) within a model 210x210 gridbox are compared to the same value. That value is the average of all observations in the 210x210 gridbox, as calculated by Equation 2. So it should not matter where the 10x10 box is within the 210x210 gridbox, since the value being compared against is always the same. Where do I go wrong here?

The reviewer's understanding of our analysis procedure is correct. If the values in the different 10x10 boxes where just random numbers independently drawn from the same distribution, the sampling errors should not depend on distance. However, the observations are not drawn from the same distributions nor are they independent. The true distribution (defined by a multi-year time-series) will depend on the location of the observation relative to sources, with respect to the atmospheric flow. Clearly this will differ for each observation within a 210x210 box. For the same reason, nearby observations will tend to be correlated; both our simulated aerosol and observed aerosol (Anderson et al JAS 2003) exhibit spatial correlations over distances of 10-100 km. Consequently, an observation in the center of a 210x210 box should be strongly correlated with that 210x210 average. But a 10x10 box in the top right corner may only be strongly correlated with the top right quadrant of the 210x210 box but not its bottom-left quadrant. A simpler and less accurate way to put it would be to say that the region for which the observation is representative and the grid-box itself overlap less and less as the distance increases. This hopefully explains the increase of errors with distance. We have added more explanation to Sect 9.

The difference in behaviour between observables is fascinating. Looking at figures 2 and 3, one does not see an essential difference between AOT and BC that could explain the different error statistics (Figure 6) and different responses to temporal sampling (Figure 5, page 7 line 15). Yet they differ. The authors offer hints at possible causes throughout the paper, especially in section 6 where they discuss the narrowness of BC plumes. I recommend adding a more selfcontained discussion in the conclusion. That discussion could also be more quantitative. In data assimilation, where they encounter very similar problems, they characterise distributions with correlation length scales. The results of the present paper suggest that AOT, for example, has a longer correlation length scale than BC concentrations, although this is not obvious from looking at Figure 2. Collins et al. (2001) use a correlation length scale of 200 km for AOT, which sounds large compared to what the authors imply here. Correlation length scale would also inform the model/observation comparison strategy, with distributions with shorter correlation length scales requiring greater caution and a

more adapted distribution of observations.

We were also surprised by the large difference in behavior. Static graphs like the ones we show in the paper do little to explain this. We have been looking at movies (available as a video supplement) of how the different fields evolve and that is far more instructive. It becomes obvious that bc and number densities in the lowest model layer 'stick' very closely to their sources (they either are deposited or elevated to higher layers). As a result, correlation length scales are very short and a single observation not very representative of the larger grid-box. In contrast, AOT is much better mixed horizontally (plus: vertical transport does not affect it) and shows longer correlation length-scales. We hesitate to explore this more fully in the current paper: it is big enough already and clearly many aspects go into this (it is a bit like asking why different models give different results: a valid question but with no easy answer). Also, we plan to return to this issue in a follow-up paper.

Other comments

-Page 4, lines 1–2: SPRINTARS can diagnose number concentrations, but that facility was not used in this study. Is that correct?

Yes, that is correct. As SPRINTARS *diagnoses* number densities (instead of calculating them prognostically) we decided there was not much to be gained from an analysis that would essentially yield the same results as say mass concentrations (e.g. pm25). Same for GOCART (WRF-Chem) over ocean.

– Page 4, lines 12–20: The authors seem to worry about the impact of hygroscopic growth on number concentrations, but I do not understand what the problem is in the context of the study. Can that point be clarified?

We discuss here how well WRF-Chem is able to simulate properties *as they might be observed*. Many particle counters first dry the aerosol, then filter by size. We only use standard WRF-Chem output which provides us with number densities in each of its three modes *at ambient conditions*. Backing out number densities at dry conditions requires information on aerosol wet-growth which was not readily available (not in the least because of the complex ammonia & nitric-acid & sulfuric-acid & water equilibrium used in WRF-Chem). We have tried to explain this more clearly in the paper.

– Page 21, caption of Figure 2: The meaning of "at 10 days, 00 hours" is unclear. I suggest "at T+10 days".

Agreed and changed.

– Page 11, section 10.1: Even if I fully understood the reason why errors increase with distance to the grid point, wouldn't that fact be an artifact of the methods used, where model data are regridded versions of observations?

In the real world, the two are independent, so distance to gridpoint might be less relevant, undermining the strategy of using only observations close to the gridpoint.

For reasons explained before, we consider this effect (increasing error with distance) to be real and directly related to the spatio-temporal structure of the aerosol fields. Obviously, our models may be over- or underestimating the spatio-temporal correlations in these fields.

– Page 11, line 30: The distance at which errors are zero in the case of a linear weighting function looks to be two thirds of the gridbox size. Is that expected mathematically, or is it a coincidence?

The errors do not become zero, merely small. Intuitively, we can expect this: the linear weighting function effectively defines a smaller grid-box (smaller than 210x210 but much larger than 10x10) for which the global model data is representative. If our observational aggregate approaches this size, errors are likely to be smallest. Mathematically, it will depend on the spatio-temporal correlations in the field itself. In practice, we do not see a big difference in this length scale amongst our six simulations.

Technical comments

– Page 3, line 20: Repeated word "from"

Corrected

– Page 4, lines 18–19: Reference is wrongly formatted.

Corrected

– Page 23, Figure 5: It would help if the blue line were thicker.

Corrected

– Page 7, line 32: Typo: "quite a bit"

Corrected

– Page 8, line 22: large -> larger

Corrected

– Page 12, line 16: Typo: "a more localised weighting function"

Corrected

Response to reviewer 2

In Page 1,
Line 5 "the field-of-view of ground-sites" is implied to be 10 km. In Page 2 Line 12 the AERONET is said to sample "no more than 5 km". In fact, most of current ground-based aerosol measurements sample significantly shorter lengths over their native integration time (between _1s and minutes). The lengths are some centimeters or meters, depending chiefly on horizontal wind speed and secondarily on instrument flow rate and sunphotometer light collector width. Ground-based measurements represent a distance comparable to 10 km only if integrated over tens of minutes or a couple of hours. Perhaps the authors have an integration time of one hour in mind for the ground-based observations, as this is the temporal resolution of the simulations. This is very different from the native integration time and should be noted.

The reviewer is correct that the 10 km resolution of our models does not do full justice to the scales at which some observations are made. This is a limitation in our analysis and was discussed in our summary. We point out that, like many atmospheric properties, aerosol shows less variation on short length-scales than on larger length scales (power spectra of aerosol distributions in space or time show a typical power-law behavior). Consequently, we suspect that variability over 10 km will not substantially alter our conclusions (although our results may somewhat underestimate sampling errors).

A few words on the expression of differences and errors would be nice. The expression(observation-model)/model (eq 4) produces numbers widely different from (modelobservation)/observation (e.g., +160% vs -62%). The latter expression encounters division by zero when the observation is zero, but is nonetheless fairly commonly used. I would recommend stating explicitly that the present study treats the model as the reference against which the "observation" is evaluated, and not the other way around.

For the reason given by the reviewer, we chose to use model data (210x210 box average) as the reference. This was explained in Sect. 3 but we now repeat it in the Summary and Introduction as well. We've also discussed this in more detail in Sect. 3.

Page 3. Line 20. Remove one of the two "from".

Corrected

Page 3. Line 28. Replace "main island Japan" with "the largest island of Japan". Japan has three more main islands.

While the reviewer is correct, the island we refer to is called Honshu, which means 'main island'. It would be identified as the main island by most Japanese. In the interest of non-Japanese readers we have changed the text to 'largest

island'.

Page 4. Line 10. Move the first parenthesis to immediately before 2013.

Corrected

Page 4. Line 19. Move the first parenthesis to immediately before Seinfeld.

Corrected

Page 7. Line 20. I do not see the exception in Figure 6.

The bars to compare are the two orange bars in the left plot. The left bar is WRF-Chem, the right bar is EMEP. They are quite similar, certainly compared to the other bars in these plots. We have removed this line to prevent confusion.

Page 7. Line 29. "wet and dry and wet deposition" should read "wet and dry deposition" or "dry and wet deposition".

Corrected

Page 8. Line 17. Remove "e.g."

Corrected

Page 8. Line 26. I would think short life-time works to increase spatial heterogeneity, not decrease.

That is a good point. What we see in this simulation is low spatial but high temporal variation over Ocean. Presumably that is due to spatial correlations in rapidly varying windspeeds. In the current paper, only spatial sampling is considered and hence sampling errors over Ocean are small. In a paper that we are currently working on, also temporal sampling is considered which substantially increases errors over Ocean. We have corrected the text and removed reference to short life-times.

Page 10. Line 14. "Sofar" should read "So far". Also Line 27.

Corrected.

Page 11. Line 25. Remove the hyphen from "More-over".

Corrected

Page 13. Line 21. Replace no with not.

Corrected.

Will a perfect model agree with perfect observations? The impact of spatial sampling.

N.A.J. Schutgens[1], E. Gryspeerdt[2], N. Weigum[1], S. Tsyro[3], D. Goto[4], M. Schulz[3], and P. Stier[1]

[1]Department of Physics, University of Oxford, Parks road, OX1 3PU, England
[2]Institute for Meteorology, University of Leipzig, Stephanstr. 3, 04103 Leipzig, Germany
[3]Norwegian Meteorological Institute, O313 Oslo, Norway
[4]National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba, 305-8568, Japan

Correspondence to: Nick Schutgens (schutgens@physics.ox.ac.uk)

Abstract. The spatial resolution of global climate models with interactive aerosol and the observations used to evaluate them is very different. Current models use grid-spacings of ~ 200 km, while satellite observations of aerosol use so-called pixels of ~ 10 km. Ground site or air-borne observations concern even smaller spatial scales. We study the errors incurred due to different resolutions by aggregating high-resolution simulations (10 km grid-spacing) over either the large areas of global model grid-boxes ("perfect" model data) or small areas corresponding to the pixels of satellite measurements or the field-of-view of ground-sites ("perfect" observations). Our analysis suggests that instantaneous RMS differences ~~between these perfect observations and~~ of perfect observations from perfect global models can easily amount to 30–160%, for a range of observables like AOT (Aerosol Optical Thickness), extinction, black carbon mass concentrations, $PM_{2.5}$, number densities and CCN (Cloud Condensation Nuclei). These differences, due entirely to different spatial sampling of models and observations, are often larger than measurement errors in real observations. Temporal averaging over a month of data reduces these differences more strongly for some observables (e.g. a three-fold reduction i.c. AOT), than for others (e.g. a two-fold reduction for surface black carbon concentrations), but significant RMS differences remain (10-75%). Note that this study ignores the issue of temporal sampling of real observations, which is likely to affect our present monthly error estimates. We examine several other strategies (e.g. spatial aggregation of observations, interpolation of model data) for reducing these differences and show their effectiveness. Finally, we examine consequences for the use of flight campaign data in global model evaluation and show that significant biases may be introduced depending on the flight strategy used.

1 Introduction

Airborne aerosols are a fascinating component of the Earth's atmosphere. They come in a bewildering variety of shapes, sizes and compositions. More importantly, they can affect the radiative budget and energy and hydrological balances of the atmosphere (Angstrom, 1962; Twomey, 1974; Albrecht, 1989; Hansen et al., 1997; Lohmann and Feichter, 2005, 1997). Dust aerosols may transport nutrients for the biosphere over long distances (Swap et al., 1992; Vink and Measures, 2001; McTainsh and Strong, 2007; Maher et al., 2010; Lequy et al., 2012) and air pollution aerosol 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). Aerosols have also been suggested as disease vectors (Ballester et al., 2013). For a recent review of some of these aspects, see Fuzzi et al. (2015).

Models provide powerful tools to explore the role of aerosols, but require evaluation against observations in order to quantify their skill and detect possible model errors. AEROCOM is an international community of scientists (http://aerocom.met.no) involved in evaluating global aerosol models (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) but model evaluations are also routinely performed by individual research groups around the world. It is therefore surprising that evaluation strategies themselves have received relatively little scrutiny.

Due to constraints on computational resources, global aerosol-climate models are currently run at spatial resolutions of ~ 200 km. This of course limits their ability to resolve fine-scale structure (Benkovitz and Schwartz, 1997;

Weigum et al., 2012) which will affect the comparison of global model data with observations: models and observations represent averages over different spatial areas. Satellite remote sensing observations are made for nominal pixels of 10 km as for MODIS (MODerate resolution Imaging Spectroradiometer) or 17 km as for MISR (Multi-angle Imaging SpectroRadiometer) or 3 km as for SEVIRI (Spinning Enhanced Visible and InfraRed Imager). Ground stations from AERONET can be estimated to sample no more than 5 km horizontally away from the site. In-situ measurements see even less of the atmosphere surrounding them. Yet, observed aerosol fields are known to exhibit variations over relatively short distances of 10 to 100 km (Anderson et al., 2003; Kovacs, 2006; Santese et al., 2007; Shinozuka and Redemann, 2011; Schutgens et al., 2013). Note that the spatial resolution of global models also impacts global model data due to the non-linear nature of many physical and chemical processes (Qian et al., 2010; Gustafson et al., 2011; Stroud et al., 2011; Weigum et al., 2015), but that is not the topic of this paper.

Recently, the disparity of spatial scales between global models and observations has attracted some attention. Using satellite retrieved solar surface radiation estimates to assess spatial representativeness, Hakuba et al. (2014b, a) estimated differences of 1–2% resp. 2–3% in 5-year seasonal means between either $1^o \times 1^o$ or $3^o \times 3^o$ areas and point measurements. Cavanaugh and Shen (2015); Director and Bornn (2015) showed that the standard deviation, skewness and kurtosis of climate data (e.g. temperature) can be be significantly different between point values and gridded values (in their analysis means were identical by construction).

We use high-resolution model simulations (with a 10 km grid-spacing) to simulate both perfect global model data and perfect observations. These data are considered perfect in the sense that they are both derived from the same high-resolution simulation that we treat as the truth. In fact, the only difference between the global model data and observations is the area over which the high-resolution simulation is averaged (see Sect. 3). No measurement errors are added to the observations. The high-resolution simulations allow us to build up statistics of the difference between observations and model data, under a large variety of scenarios. In particular, we consider different observables like AOT, $PM_{2.5}$, number densities and CCN for different regions on the globe. We also evaluate a variety of averaging and interpolation strategies designed to bring model data and observations closer together. These high-resolution model simulations provide us with a toy model of what happens when global model data are evaluated with observations, ignoring both model and observation errors.

Since we simulate global model data as an average over the high-resolution data, a very relevant question is: 'what average is appropriate?'. This question is closely tied to the question of what the grid-point value of a global model represents and will be addressed later.

Section 2 introduces the three different models and 6 different regions for which we have high-resolution simulations. We also explain how the simulated fields were turned into observables. Section 3 describes in more detail how both global model data and observations are generated from the high-resolution simulations. In particular, Sect. 3.1 discusses various interpretations that may be given to a global model's grid-point value. Section 4 then introduces the concept of spatial sampling as a source of error through some examples. More substantive statistics can be found in Sect. 5, 6, 7, 8 and 9. An evaluation of several strategies to reduce spatial sampling differences is given in Sect. 10. A preliminary analysis of the consequences of spatial sampling for the use of flight campaign data can be found in Sect. 11. The paper concludes with a summary (Sect. 12)

Note that Sect. 3.2 contains some general guidelines to interpreting many of the figures and statistics that appear in this paper.

2 The regional models

Three different regional models were used to create high-resolution simulated fields (10 km, 1 hour) of common observables: AOT, extinction, $PM_{2.5}$, black carbon mass concentration, number densities and CCN. Fig. 1 shows the simulation regions, and Table 1 summarises the most important information on these simulations.

WRF-Chem (Grell et al., 2005; Fast et al., 2006) was run for three regions using the MADE/SORGAM aerosol module (Ackermann et al., 1998; Schell et al., 2001), and one region using the GOCART bulk aerosol scheme. The meteorology was nudged to NCEP-FNL operational analysis data. The West-Europe and Oklahoma runs used emission scenarios (TNO MEGAPOLI-2005 or US National Emissions Inventory NEI-2005) with imposed 24-hour cycles for the anthropogenic emissions. These regions are characterised by fairly localised spatially-fixed sources. The Congo experiment used daily biomass burning emissions derived from MODIS fire counts and is characterised by highly localised sources that differ in location from day to day. The MADE/SORGAM module assumes the aerosol to exist in three modes (Aitken, accumulation and coarse) of varying species mixtures (sulfate, nitrate, organic and black carbon, sea salt and dust). MADE/SORGAM explicitly treats nitrates and SOA (secondary organic aerosol).

An expanded version of EMEP/MSC-W (Simpson et al., 2012) that includes calculations of aerosol bulk optical properties (based on work by Hess et al. (1998) and Chin et al. (2002)) was run at a $0.1^\circ \times 0.1^\circ$ grid, using ECMWF-IFS meteorology for 2008 and TNO-INERIS emissions for 2009 for Europe. Emissions of black carbon were derived from the emissions of primary $PM_{2.5}$, using EMEP standard split-factors (per country and sector). Monthly, day-of-week and hourly temporal profiles were applied to the annual emis-

sions. The EMEP chemical scheme includes approximately 160 reactions. The aerosols are represented as bulk-mass distributed between a fine fraction (including sulfate, nitrate, ammonium, organic and black carbon sea salt and dust) and a coarse fraction (nitrate, sea-salt and dust). Ammonium nitrate is calculated with the equilibrium model MARS, and the formation of coarse nitrate from nitric acid depends on the relative humidity. SOA is calculated using the VBS approach. For all details see Simpson et al. (2012) and references therein.

NICAM-SPRINTARS (see Goto et al. (2015) and references therein) was run in global mode with a stretched grid that had a resolution of 11 km over a part of Honshu (the largest island of Japan). Its meteorology was nudged to NCEP-FNL reanalysis data. SPRINTARS uses a mass bulk scheme with individual modes for sulfate, organic carbon, black carbon and bins for sea-salt and dust. Two different organic/black carbon mixtures are also represented by individual modes. Anthropogenic emissions of black carbon and the SO4 precursor gas SO2 had a prescribed diurnal cycle. SOA were treated in the simple manner of scaling aerosol emissions. Nitrate aerosol were ignored in this SPRINTARS simulation.

Both EMEP and SPRINTARS do not calculate number densities as prognostic variables (SPRINTARS can diagnose them from assumed size distributions) and consequently did not provide those fields for our analysis. Both EMEP and SPRINTARS data were regridded from their original model grids to regular grids with 10 km spacings.

2.1 Observable parameters

~~The simulated fields examined in this paper are , for obvious reasons, all observables.~~ In this sub-section we discuss how well our models are able to simulate aerosol properties (see Table 2) as they would be observed. All of the models provided AOT, extinction and (dry) $PM_{2.5}$, although WRF-Chem calculates AOT and extinction for 600 nm and EMEP and NICAM-SPRINTARS for 550 nm. ~~WRF-Chem MADE also provided number densities and CCN at various super-saturations S.~~

Real black carbon measurements by SP2 (single particle soot photometer) require a minimum black carbon content per particle. ~~Due to~~ In models with mass bulk schemes, particles either contain only black carbon or none at all. Modal aerosol schemes also cannot properly simulate SP2 measurements, due to the instantaneous redistribution of black carbon mass over many particles of mixed species ~~, modal aerosol schemes cannot properly simulate SP2 measurements, see also~~ which leads to very low concentrations per particle (Kipling et al., 2013). We decided to ignore this minimum black carbon content and used the total black carbon concentration as provided by the ~~model~~models.

~~Number densities (N3~~Real number density measurements dry out the particles before selecting only those above a certain diameter. Hence, N10 and N50 ~~, i.e. number densities for particles with wet~~ refer to number densities of particles with dry diameters in excess of ~~3,~~10 or 50 nm~~) are constrained by a lower size threshold. In actual measurements, this lower size threshold applies to particles in relatively dry air but~~ . WRF-Chem provides only modal ~~information for~~ number densities at ambient humidities. ~~However, based on dry $PM_{2.5}$ and its associated water content~~ Based on auxiliary model data, we ~~estimate~~ estimated that 'taking out' the water has at most a 10% effect on N10 or N50 values. ~~Furthermore, from $PM_{2.5}$ data we also conclude that taking account of dry conditions may actually~~ We also concluded that this may increase the spatial sampling errors we are studying.

~~Finally, since WRF-Chem~~ Furthermore, the model calculates the equilibrium of the ammonia & nitric-acid & sulfuric-acid & water system (Seinfeld and Pandis, 2006), ~~'drying out' the modelled particles amounts to~~ and 'drying-out' particles involves much more than simply removing the water ~~: (~~it would lead to a shift in the equilibrium). Currently WRF-Chem provides no mechanism to simulate this aspect of observed number densities. So we decided for a practical approach and use ambient number densities to calculate N10 and N50.

3 Simulating observational and global model data

This section briefly describes the main methodology used in this paper. Using the high resolution simulated fields, we have generated both perfect observations and perfect global model data. The high resolution field v has a regular rectilinear horizontal grid (10×10 km), and a regular temporal spacing (1 hour). Only the vertical spacing is non regular and differs among the models. The field v can be thought of as 3 or 4-dimensional data cubes v_{xyt} or v_{xyzt} where $x = 1 \ldots n_x$ and $y = 1 \ldots n_y$ are indices to the horizontal coordinates, $z = 1 \ldots n_z$ is an index to the vertical coordinate and $t = 1 \ldots n_t$ is an index to the time coordinate. In the following, the z coordinate is ignored for brevity's sake. A single perfect observation O_{xyt} at time t and location x, y is simulated by:

$$O_{xyt} = v_{xyt}. \tag{1}$$

A perfect global model grid point's value M_{xyt} can be simulated by averaging v_{xyt} over a global model grid-box area $(2\Delta x + 1) \times (2\Delta y + 1)$ in the high-resolution field:

$$M_{xyt} = \sum_{i=-\Delta x}^{\Delta x} \sum_{j=-\Delta y}^{\Delta y} w_{ij} v_{x+i;y+j;t}, \tag{2}$$

where Δx and Δy represent the longitudinal and latitudinal half-sizes of a grid-box, as measured in the coordinate in-

dices. Here w is a normalised weighting function (to be defined later). Note that perfect model data can only be calculated on an inner domain of the high-resolution ~~run~~ region of $1+\Delta x \leq x \leq n_x - \Delta x; 1+\Delta y \leq y \leq n_y - \Delta y$.

In the case that the location of the observation and the grid-point coincide, an instantaneous spatial sampling error can now be defined as:

$$\epsilon_{xyt} = O_{xyt} - M_{xyt} \tag{3}$$

where we use the perfect model value as a reference, since it is the model value that we want to evaluate in actual comparisons of observational and model data. It is straightforward to define a relative sampling error for time-averaged data by

$$\varepsilon_{xyt} = \left(\sum_{k=t-\Delta t}^{k=t+\Delta t} O_{xyt} - M_{xyt} \right) / \left(\sum_{k=t-\Delta t}^{k=t+\Delta t} M_{xyt} \right), \tag{4}$$

where $2\Delta t + 1$ is an arbitrary averaging interval. Using the global model value (instead of the observation) as reference prevents denominators from becoming zero.

A subset of the data cube of our regional simulations is used to build up error statistics. In addition to the limitation imposed by the Eq. 2 (already discussed), the outer 50 km of the simulated region was excluded from our analysis to reduce boundary effects. Similarly, the first two days of each simulation were used as a spin-up and excluded from analysis. At various points in our analysis, we have studied the sensitivity of our results to these choices but found no significant impact.

3.1 Interpretation of the grid-point value

We generate the global model grid-point value M_{xyt} as a weighted average of the high-resolution simulation over a large area, see Eq. 2. The weighting function w represents our interpretation of the global model's grid-point value. The question is what are realistic w like for actual global models?

A numerical grid with spacing ~~Δ~~ L can represent standing or travelling waves with a wavelength of in theory ~~2Δ~~ $2L$ and in practice ~~$4\Delta - 6\Delta$~~ $4L - 6L$. This suggests that the grid-point value of a low resolution model is at best some average of a high resolution simulation over the grid-box ~~$\Delta \times \Delta$~~ $L \times L$. Moreover, at horizontal resolutions of ~ 200 km, there is no evidence that actual global models have converged numerically (Pope and Stratton, 2002; Roeckner et al., 2006; Williamson, 2008). As the resolution of global models is increased, various aspects of the models are tweaked to obtain best agreement with either observations or reanalysis datasets (see Pope and Stratton for a very clear description). Diffusion is adapted to prevent numerical instabilities and the gravity-wave drag coefficients are modified according to the resolution of the orography. ~~Most famously~~Best known, various parameters related to sub-grid cloud processes are tuned to obtain radiative balance at the top-of-atmosphere. Our point here is that the strategy for tweaking the global model to best reflect an observational or reanalysis dataset effectively determines w, although this is never explicitly discussed. In addition, models are tuned for only a few parameters for which abundant observations or reliable reanalysis data are available (e.g. pressure, temperature). There is no reason to assume that other parameters require the same weighting function, as these models are non-linear.

Hence we argue that w is fundamentally unknown (and may actually vary with time and location). To conduct our analysis, we therefore assumed three different weighting functions and performed sensitivity studies (to be described later). The weighting function most used in this paper is a constant value throughout the grid-box. This corresponds to the mental model that many scientists have of the physics processes that occur in a grid-box. The other two weighting functions favour the area near the grid-point more than the outer edges of the grid-box. One weighting function uses a linear profile (highest at the grid-point, zero at the edge) and another uses a Dirac-δ (centred at the grid-point). The latter we consider a rather unlikely choice of w but it does correspond to the case where the model has numerically converged.

3.2 Some conventions used in this paper

This paper contains many figures and statistics of spatial sampling error distributions. Instead of repeating the same information, some aspects are explained here. Error distributions are always given for either instantaneous ('hourly') or monthly data over a single region, see Table 1. These error distributions are quantified through Root-Mean-Square (RMS) values or quantiles. They represent typical errors per region (over no more than a month), which should not be mistaken for the typical error in any one longitude/latitude location. We use the so-called parametric 7-number summary of the 2, 9, 25, 75, 91 and 98% quantiles q of the errors because for a normal distribution these quantiles are equally spaced. Any skewness or extended wings in a distribution will be readily visible. In particular, we often refer to the inter-quantile ranges $\Delta q_{50} = q_{75} - q_{25}$, $\Delta q_{82} = q_{91} - q_9$ and $\Delta q_{96} = q_{98} - q_2$. In e.g. Fig. 5 different shades of grey are used to denote these interquantile ranges: light grey for Δq_{96}, medium grey for the Δq_{82} and dark grey for Δq_{50}. The solid blue line represents the median error. In Fig. 6, box-whisker plots show the error distributions. Different widths of the bars are used to denote different inter-quantile ranges: narrow for Δq_{96}, medium for Δq_{82} and wide for Δq_{50}. The black rectangle represents the median error and the black circle the mean error. In a few figures, additional error distributions are shown using colored lines: the Δq_{50}, Δq_{82} and Δq_{96} ranges will be indicated by resp. solid, dashed and dotted lines.

The standard measure of uncertainty, the standard deviation, is ~~of course~~ half the $q_{84.1} - q_{15.9}$ inter-quantile range. For a Gaussian distribution, Δq_{50} is 1.35 times the standard

deviation, and Δq_{82} is 2.68 times the standard deviation. For a Gaussian distribution with zero mean, the standard deviation and the RMS value will of-course agree.

4 Examples of spatial sampling errors

In Fig. 2, we show instantaneous simulated AOT and surface black carbon concentration after 10 days in the WRF-Chem W-Europe run. By comparing the field in a small 10×10 km box to the average of a large 210×210 km box surrounding it (approximate size of present-day global model grid box), we assess spatial sampling errors. The centre of the large box we refer to as grid-point (of the global model). By moving these two boxes together throughout the region, we can build up statistics of spatial sampling errors (also shown in Fig. 2). These errors can reach $\sim 100\%$ and form coherent patterns several global model grid-boxes large. Time series of the global model and observed values at a single location are shown in Fig. 3. In the case of AOT, we see that the perfect observation can both over- and under-estimate the perfect model value with variations on a time-scale of a day or so. The black carbon time-series, on the other hand, shows systematic underestimation by the perfect observation over long periods for most of the month (note that events of overestimation also occur but on smaller time-scales). Although these time-series vary a lot throughout the region, this example is nevertheless typical.

Since these spatial sampling errors are substantial, it makes sense to try and reduce them by temporally averaging the data. In Fig. 4, we show monthly averaged simulated AOT and surface black carbon concentration from the same run. The spatial sampling errors in monthly averaged observations are also shown in Fig. 4. They are smaller than the errors for instantaneous fields but still quite substantial (up to $\sim 20\%$ for AOT and $\sim 65\%$ for black carbon). Note also that the error patterns have become larger and more coherent. As a matter of fact, notice how closely the patterns in sampling errors for black carbon agree with its emission sources. Except that sampling errors are negative (and quite large) where concentrations are quite low: when defining areas downstream from sources where the aerosol is supposedly well-mixed spatially it is important to consider the grid-box size of the model which is evaluated as much as the length-scales involved in the actual aerosol processes.

The effectiveness of temporal averaging is shown in Fig. 5, where the spatial sampling errors are shown as a function of averaging period. Time-averaging does decrease spatial sampling errors but not as fast as one would expect if instantaneous sampling errors ~~were~~ behaved like independent Gaussian noise. This is understandable because the persistence of emission sources and flow patterns in the atmosphere create temporal correlations in the fields of a few hours to a few days. Note that AOT is more strongly (beneficially) affected by time-averaging than surface black carbon concentrations.

5 Agreement among models

Before studying these spatial sampling errors in more detail, we consider how (dis)similar they are among different models. The Europe region simulated by EMEP encompasses the W-Europe region simulated by WRF-Chem MADE and so these two models allow ~~for ready intercomparison~~ ready intercomparison for May 2008, see Fig. 6. We see that both instantaneous and monthly errors as predicted by WRF-Chem and EMEP are of similar magnitude although WRF-Chem generally produces larger errors~~(note the exception of instantaneous errors for extinction near 2 km AGL)~~. Error magnitudes for different observables behave similarly among WRF-Chem and EMEP: monthly errors for AOT and surface black carbon are the smallest resp. largest errors. EMEP monthly error maps (see Fig. 7) also look similar to WRF-Chem results (Fig. 4), especially for black carbon surface concentrations.

It would be interesting to understand the reason for the differences. From global model studies in the context of AEROCOM (e.g. Myhre et al. (2013); Randles et al. (2013); Stier et al. (2013)), we know that such attribution is difficult and here we limit ourselves to pointing out some obvious differences between WRF-Chem and EMEP. First, there are differences in emission inventories and sea-salt emission parametrisations. Second, although both models were nudged to reanalysis data, transport will be different due to different dynamical cores and vertical resolution (WRF-Chem uses twice the vertical resolution as EMEP). For similar reasons wet and dry deposition are different. Both models also use a very different aerosol scheme (mass bulk or two moment scheme). All of this will affect aerosol life-times, which in turn will affect the spatio-temporal variability of aerosol.

It should also be pointed out that EMEP shows quite a bit of month-to-month variation: e.g. January 2008 errors for AOT and March 2008 errors for surface black carbon concentration are markedly bigger than those estimated for May~~(not shown)~~.

The most important point here is that both models suggest spatial sampling errors of similar magnitude with similar spatial patterns.

6 Different observables and different regions

Figure 8 shows relative spatial sampling errors (either instantaneous or monthly) for all observables and the three WRF-Chem MADE regions (see also Tab. 1 and Fig. 1). Instantaneous RMS errors are large: from 20 % up to 160% depending on observable and region (the RMS errors are calculated over a single region for the full month, see Table 1). There are clear and (mostly) systematic differences among the three regions in that W-Europe shows the largest errors and Congo the smallest. This may be related to the overall wind-flow:

Congo shows the most laminar flow (and hence most coherent aerosol plumes), while W-Europe shows a very turbulent flow (we do not wish to discount other effects like the spatio-temporal distribution of sources but a full explanation is outside this paper's scope). Two observables (black carbon concentrations near 2 km AGL for all three regions and surface CCN at $S = 0.02\%$ in W-Europe) show errors down to -100%. In the case of black carbon, this is due to narrow black carbon plumes travelling through an otherwise pristine air layer: the observation often sees the pristine air but the model always includes contributions from the plume. In the case of CCN, the background CCN at $S = 0.02\%$ is very low, especially close to sources where many small particles are emitted. But once in a while a plume of larger particles travels over giving rise to much larger CCN at low supersaturation $S = 0.02\%$.

The monthly errors can be reduced quite a bit compared to the instantaneous errors. For many observables, RMS errors are 5–15%, although for observables like surface black carbon concentrations and N10 it can be resp. 30–50% and 30–80%, with individual errors reaching over 100%. Congo represents quite a different situation from the other two regions: the reduction due to averaging is much less, and in the case of surface N10 there is actually a slight increase in errors. An important difference between W-Europe & Oklahoma on one hand and Congo on the other is that the first have mostly fixed aerosol sources with a prescribed diurnal cycle. The latter has emission sources (fires) in different locations from day to day. ~~The explanation for the larger N10 errors is similar to that of the large instantaneous errors for black carbon plumes, except now the temporal extent should be taken into account as well.~~

Figure 9 shows relative spatial sampling errors for the other 3 regions, all simulated by models with mass-bulk schemes for aerosol. In general, spatial sampling errors appear to be a bit smaller than in Fig. 8, but note the exception of extinction near 2km AGL. ~~Most monthly~~ Monthly sampling errors over ocean are very low, ~~presumably due to the short life-time and diffuse source regions of~~ due to spatial correlations in the near-surface wind-speeds that cause sea-salt aerosol emission. But large errors are found for extinction over ocean near 2 km AGL, that seem partly due to isolated plumes of sea-salt but mostly due to a broken cloud field that rains out sea-salt locally. Both instantaneous and monthly errors over Japan become larger if only observations over the land area are considered. The Japan region includes parts of the Japan sea and the North Pacific ocean that account for more than 80% of the simulated area. Also, the Japan simulation, like the Congo simulation, shows rather laminar flow from meso-scale to synoptic scale. Finally, simple statistics like in Fig. 9 cannot convey that over an extended region like Europe there are areas with systematically small or large sampling errors due to source locations and orography (see also Fig. 4 and 7).

In the case of actual observations, there may be quite a bit of intermittency in their temporal sampling suggesting that the spatial sampling decreases we have shown here for monthly averages represent a best case scenario.

7 Vertical distribution of sampling errors

The vertical distribution of spatial sampling errors can be very different depending on observable and region. Figures 10 and 11 show the instantaneous and monthly relative spatial sampling error profiles for extinction, N10 and black carbon concentrations.

We see that although errors are typically largest at and near the surface, this does not preclude large errors higher up in the atmosphere. The instantaneous errors for black carbon concentrations actually show largest errors from 2 to 7 km AGL. This is due to black carbon plumes in a relatively pristine background, which also explains why the error distribution is so clear skewed to negative values (observation sees the pristine background while the model also includes plumes). Black carbon's only source is surface emission, but both extinction and N10 also have sources throughout the troposphere (nucleation, condensation and in-cloud production of sulfate) which likely explains the difference between these observables. ~~Note that by design, WRF-Chem black carbon concentrations cannot go below a minimum value of $2 \cdot 10^{-16}$ μg/m^3; this should result in an underestimation of the sampling errors.~~

For the monthly errors, both extinction and black carbon concentrations show secondary maxima in sampling errors well above the surface, while N10 errors drop steadily with altitude.

We have analysed the sampling errors at their original model levels, which for these simulations occur at fairly constant altitude above ground. Note that the errors estimated in this subsection do not take into account that a global model's grid-box may have a vertical extent larger than that of our regional simulations. Taking this into account would only increase the estimated errors. The profiles of spatial sampling errors for the bulk mass simulations are rather constant and therefore not discussed here.

8 Impact of grid-box size and shape

8.1 Impact of latitude

Although our high resolution simulations were made at different latitudes on Earth, so far we have assumed that the global model grid-box size is equal to the grid-box size of a T63 grid at the equator (210 by 210 km). At higher latitudes, the longitudinal extent of the grid-box shrinks (at least for rectangular grids), which may reduce spatial sampling errors. This is explored in Fig. 12. As we can see, smaller longitudinal extent leads to smaller errors although the effect is

rather mild. When the longitudinal extent is halved, errors in monthly-averaged fields decrease between 10 and 30% of the original errors, with $\sim 20\%$ a very typical value. Also, larger errors are usually less affected than smaller errors although the differences are not very big. ~~The figure for black carbon is typical, while AOT is rather the exception to this.~~ Spatial sampling errors in instantaneous fields behave very similar (not shown), although fields that show very large errors (like surface BC or surface CCN at $S = 0.02\%$) tend to show less improvement ($\sim 10\%$) when the grid-box longitudinal extent is halved.

Note that the longitudinal extent only has an impact on spatial sampling errors because there are spatial and temporal correlations in the aerosol fields. If these fields were independent random noise, decreasing longitudinal extent would barely have an impact on sampling errors.

8.2 Impact of grid-box size

The impact of model resolution is also easily explored, see Fig. 13. ~~Aggregation~~ Monthly sampling errors decrease by 10 to 50% from T63 (210 by 210 km) to T106 (125 by 125 km, a third of the T63 grid-box area), with 40% a rather typical value. Surface observations are less affected with decreases of $\sim 30\%$, especially N10 whose spatial sampling errors in all three simulations only decreased by $\sim 20\%$ when the grid-box size was halved. For instantaneous values (not shown), the typical reduction in sampling error is smaller, $\sim 30\%$, especially for surface fields: $\sim 20\%$.

As with the longitudinal extent, gridbox-size only has an impact because of the spatial and temporal correlations in the aerosol fields. A field of independent random noise exhibits sampling errors quite independently of gridbox-size (unless the box, and the number of values therein, becomes very small).

9 Observations offset from the grid point

So far we have considered observations at the exact grid-point of a global model's grid-box which is a useful starting point but also quite unrealistic. For a sample of randomly distributed observations in a 210 by 210 km grid-box, only 2% will be within 10 km of the grid-point and 50% will be more than 84 km away from it. By considering observations located throughout the grid-box, and not just its centre, it is possible to show how monthly sampling errors increase with distance of the observation to the grid-point, see Fig. 14. As a matter of fact, 50% of possible AOT observations have errors at least twice as large as found for ~~observation coinciding with~~ an observation at the grid-point. Observations in the very corners of the grid-box exhibit errors three times as large.The increase of sampling errors with distance to the grid-point for surface black carbon concentrations is not as large but still significant.

That sampling errors increase with distance may be surprising but can be explained. The evolution of aerosol across a global model grid-box may differ quite a bit due to differences in sources, flow and deposition (especially wet). Nevertheless, as is well known from observations, aerosol exhibits correlations over several 10s of km (Anderson et al., 2003; Kovacs, 2006; Santese et al., 2007; Shinozuka and Redemann, 2011; Schutgens et al., 2013) and our high-resolution simulations are no different. Hence, an observation at the centre of a grid-box will correlate strongly with a large part of that grid-box while an observation in the upper-right corner will only correlate strongly with (part of) the upper-right quadrant of that grid-box but less so with the lower-left quadrant. It is important to realise that aerosol in individual 10×10 km^2 boxes *cannot* be considered as independent and identically distributed (i.i.d.) random variables as is sometimes assumed in climate studies (Cavanaugh and Shen, 2015; Director and Bornn, 2015). If aerosol behaved like i.i.d. random variables, sampling errors would *not* increase with distance.

Figure 15 shows box-whisker plots of monthly sampling errors for several observables, either at the grid-point, or at a distance of 70 or 100 km, for the W-Europe region. Similar results can be shown for Oklahoma and Congo, where the relative increase with distance is often (but not always) larger. For all three regions and all observables, the increase for Δq_{82} at 70 km is between $1.2 - 2.3\times$ and the increase at 100 km is between $1.4 - 3.4\times$. Instantaneous spatial sampling errors increase less fast with distance but still significantly: typical increases for Δq_{82} at 70 km is 1.3 for AOT and 1.2 for surface black carbon concentration (i.e. monthly averaging is more beneficial for an observation at the grid-point than one at 70 km distance).

As discussed before (Sect. 3.1), the meaning of a global model's grid-point value is not obvious. So far we have assumed that the grid-point value is the unweighted average of the high-resolution field over the global model's grid-box (i.e. a constant weighting function w). Here, we explore how the sampling errors depend on different weighting functions. Fig. 16 shows how a constant, linear or Dirac-δ weighting function affects sampling errors as a function of distance to the grid-point. For the Dirac-δ weighting function, sampling errors are equal to zero at a distance of zero: the global model's value is equal to the observation (since both are perfect). But as distance increases, so will the spatial sampling errors. Actually, for distances larger than ~ 30 km, the three very different weighting functions give rather similar sampling errors (but notice that more localised weighting functions yield larger errors as expected). Since for randomly distributed observations, only $\sim 6\%$ would be closer than 30 km to the grid-point, we feel it is justified to conclude that the shape of the weighting function has only a small impact on statistics of spatial sampling errors. The spatio-temporal variation of the field is far more important.

10 Strategies for reducing sampling errors

The typical sampling errors when the observation is at the model grid-point are lower than those for an observation offset from the grid-point. It seems unlikely that we can devise strategies to reduce "centre-of-grid-box" errors, other than temporal averaging (see Sect. 6) or further averaging global model data (and their associated observations) over multiple grid-boxes. But the sampling errors for observations offset from a grid-point might be reduced by proper screening, interpolation within the model grid, or considering multiple observations at the same time.

10.1 Observations close to the model grid-point

As Fig. 14 shows, the smallest spatial sampling errors occur for observations close to the model grid-point. As a matter of fact, within a distance of 30 km, there is hardly any change in the errors (note: this figure uses the constant weighting function). To keep sampling errors as small as possible, one might only select observations that are within 30 km of a model grid-point. For a T63 grid-box at the equator ($210 \times 210 = 44100$ km^2), that implies only 6% of randomly distributed observations would be usable, a substantial reduction of potential observational data. For an upper distance of 50 km, this increases to 18% of observations, still representing a significant loss of observational data.

One benefit of selecting only observations close to the grid-point is that here the impact of the weighting function is most pronounced (see also Fig. 16). So within 30 km of the grid-point, spatial sampling errors may actually be very small if the weighting function is highly localised. Since it is impossible to know the actual weighting function, it may be difficult to assess whether it is localised or not.

10.2 Aggregating observations over the model grid-box

It has been suggested (e.g. Sayer et al. (2010)) that aggregating observations over a model grid-box is the best strategy for comparing models with observations. Obviously, such a strategy is only possible for satellite data that provide contiguous wide swath observations (e.g. MODIS, MISR, POLDER, SEVIRI). Moreover, it can be expected that the success of this strategy depends on the weighting function that is applicable. Figure 17 shows relative spatial sampling errors in case of observations that are spatially aggregated before comparison to the model (it is assumed the aggregation is space-filling). Here the model grid-point and the centre of the aggregated observations coincide. As a result, sampling errors go to zero for the constant weighting function as the observational aggregation approaches the extent of the grid-box. For the linear weighting function, we see that errors initially become smaller as the aggregation increases and then grow again as the observational aggregation approaches the extent of the grid-box. Still, sampling errors are halved when aggregating observations over the full grid-box so there is clearly a benefit. The extreme weighting function of the Dirac-δ obviously leads to large errors.

For actual satellite measurements it will be difficult to observe the complete grid-box, due to e.g. cloud cover, sun glint or high surface albedo. Sayer et al. (2010) show that in the case of AATSR observations (nominal 10×10 km pixel) and the GEOS-Chem model ($5^\circ \times 4^\circ$ grid-box) it is extremely unlikely that more than 50% of a model grid-box would be covered by observations, that is: space-filling aggregations over global model grid-boxes are very unlikely.

10.3 Multiple observations in a model grid-box

Instead of a space-filling aggregation, one could average multiple observations in the same grid-box before comparison to the grid-point value and hopefully reduce sampling errors. The idea here is that if the observations are sufficiently far apart and represent fairly independent samplings of the field within the grid-box, their average should be distributed closer to the (weighted) grid-box average than an individual observation. This is similar to the previous sub-section, except far fewer observations are needed and no space-filling aggregation is required. This strategy may be employed for surface sites as well as for satellite data.

Figure 18 show errors in case of 4 independently distributed observations throughout the grid-box. Clearly, averaging multiple observations helps to reduce spatial sampling errors, even when the Dirac-δ weighting function is assumed! But note that this improvement is less in case of more localised weighting functions. For the constant weighting function, we also see that smallest errors now occur not at a distance of 0 km, but at a distance of 50 to 70 km (for the linear weighting function this minimum shifts closer to the grid-point). This is quite understandable: close to the grid-point multiple observations are clustered together. Hence they will not be very different. As distance increases, the randomly distributed observations sample more of the grid-box. Obviously, using more observations than 4 will give better results (not shown).

Note that Fig. 18 does *not* suggest that *any* set of 4 observations reduces sampling errors: if those observations are very close together, averaging them will hardly improve on the error.

10.4 Interpolating model data among grid-points

By interpolating the model data to the location of an observation, it may be possible to reduce spatial sampling errors for observations located away from the model grid-point. The idea is to construct virtual model data for a virtual grid-box centred on the observation. This interpolation can be performed in different ways; here we consider linear interpolation and distance-weighted averaging. Figure 19 shows that linear interpolation i.c. of a constant weighting function

clearly has a beneficial effect on spatial sampling errors, especially for observations far from the global model's grid-point. Notice that from about 80 km distance, errors become constant and no longer increase with distance (they are always larger than the errors for an observation *at* the grid-point). Obviously, the impact depends on weighting function and interpolation method, as shown in Fig. 20. Figure 20 shows that interpolation is most beneficial for observations farthest from the grid-point and can actually lead to larger errors close to the grid-point (especially for distant-weighted averaging). Interestingly, the more localised the weighting function, the more beneficial the interpolation (presumably because the global model data are now identical to observations at the grid-point). Finally, this figure shows that linear interpolation performs better than distance-weighted average. This holds for all observables and all regions we considered.

Much the same conclusions can be stated for instantaneous values, except that the beneficial impact of interpolation is less pronounced.

11 Flight campaigns

Unlike satellite or ground-site observations, measurements taken during a flight campaign cannot be properly averaged over time (at least on time-scales from days to months and longer). To simulate the (nearly) instantaneous measurements during horizontal legs of flight campaigns, we use narrow tracks: 10 km wide and 210 km long, centred on the grid-point and running in either East-West or North-South direction. Profiles of spatial sampling errors for such flight campaign data can be seen in Fig. 21. Compared to instantaneous point observations (also shown), the flight campaign observations are less affected by spatial sampling issues because they sample a larger part of the grid-box. Even so, significant instantaneous RMS errors exist, varying between 10-41% for extinction, 10-46% for N10 and 21-100% for black carbon concentrations at different altitudes and for different regions (these errors are for a best case scenario: a grid-box long flight path centred on the grid-point). For Congo, spatial sampling errors can be quite different depending on whether the flight path runs North-South or East-West Around 6 km AGL prevailing wind flows are East-West, resulting in similarly orientated plumes. If the flight track observations are within and along such a plume, spatial sampling errors will be large and positively biased. If the flight track observations are across such a plume, errors will be smaller and (over a large domain) unbiased.

The Congo results highlight a particular issue with flight campaign data: if the flight tracks ~~deliberately have~~ have deliberately been chosen to follow observed aerosol plumes, perfect observations will overestimate perfect model values by significant amounts.

~~Near~~ Almost vertical legs of flight campaigns should experience errors like those discussed for point observations, Sec. 7~~, but notice~~. Notice that we do not consider the vertical extent of a global model's grid-box in our analysis.

12 Conclusions

The spatial resolutions of current global aerosol models and the observations used to evaluate them are very different. Model grid-point values are representative of areas of $\sim 200 \times 200$ km^2 but individual observations seldom see more than $\sim 10 \times 10$ km^2 of the atmosphere. This difference in 'field-of-view' should affect the evaluation of models with observations but has received little attention in the literature. We believe our paper is the first systematic and qualitative study of the differences between a perfect model and perfect observations due to spatial sampling.

Using high-resolution simulations for 6 different regions by 2 different regional models and 1 global model, we show that spatial sampling errors can be substantial across a range of observables (AOT, extinction, $PM_{2.5}$, black carbon concentrations, number concentrations and CCN). These spatial sampling errors fluctuate in time and space, depending on emission sources, grid locations, weather and aerosol processes. Ultimately, they constitute a noise that will be present in any model evaluation and that can not be eliminated entirely unless model grid sizes become smaller than observational fields-of-view.

Assuming observations that do not coincide with the global model's grid-point but are offset by 80 km (54% of randomly located observations in a 210×210 km grid-box will be further away), the following statistics are offered. For instantaneous data, RMS spatial sampling errors (defined as observation minus global model value) are larger than 30%, typically between 40 and 80% and may go up to 160% (depending on observable and region). These errors are typically positively skewed and highly non-Gaussian. For monthly data, RMS sampling errors are larger than 10%, typically between 10 and 40% and may go up to 75% (depending on observable and region).

This noise can however be reduced: we have explored the impact of spatial or temporal averaging of data as well as selection of observations based on distance to a grid-point or interpolation of model data to the location of an observation. Our study suggests that while increased model resolution will of course be beneficial, resolutions will need to be 4 times higher (50×50 km^2 grid-box area) before spatial sampling errors become significantly smaller. In the mean time, we recommend that both model data and observations are spatio-temporally averaged to ensure best agreement. Here the model data must first be spatially interpolated to and temporally collocated with the observation. Optimal averaging procedures will depend on the spatio-temporal sampling of the observations, the characteristics of the observable and

the requirements of the scientific community, so we offer no single prescription although the results in this paper provide some guidelines. Optimal strategies for evaluating models with observations need to receive more attention from researchers.

Our results suggest that caution is needed when using in-situ measurements in global model evaluation. These measurements consistently led to larger spatial sampling errors than remote sensing measurements like AOT. For instance, monthly surface black carbon concentrations & number densities for our simulations have RMS spatial sampling errors of at least 30% and and up to 80%. Best case scenarios for flight campaign data still allowed spatial sampling errors of 100% and typically the observation would underestimate the model.

Regarding the large sampling errors in case of black carbon, other species (e.g. sulfate, sea-salt) were not explicitly analysed in this paper but show different results (not shown). Sulfate errors tend to be rather small, probably due to the multitude of sources and relatively long-life times. Sea-salt, on the other hand, shows large and systematic monthly sampling errors along coast lines (unsurprisingly). Given the size of our global model's grid-box, these errors extend quite far into land or over sea. The important point here is that sampling errors for species mass concentrations can be very different dependent on species and hence have a big impact on the evaluation of a model's particle speciation.

It is likely that the spatial sampling errors estimated in this paper are under-estimates. First, Qian et al. (2010) showed that model spatial variability over 75 km increased significantly (by 60 to 100%) when model resolution changed from 15 to 3 km. Our current high-resolution simulations have resolutions of 10 km. Second, our high-resolution simulations do not resolve fine-structure below 10 km while many in-situ measurements actually have fields-of-view on the order of millimetres to centimetres (e.g. particle inlets). Third, our models are more limited in the spatio-temporal variation of their emission sources than reality due to assumed and constant diurnal patterns in anthropogenic emissions. Finally, even high-resolution models will have to take a broad view of aerosol and describe average properties (e.g. mass and/or number densities) instead of modelling individual aerosols in all their variety.

On the other hand, it is possible that in areas far away from sources (e.g. the free troposphere over the remote ocean) aerosol has mixed sufficiently to strongly reduce ~~these~~ spatial sampling errors (e.g. HIPPO measurements over the Pacific, see also Weigum et al. (2012)). Our simulations do not really allow us to explore this scenario.

In the interest of comparing likes to likes, this paper does not consider ~~the fact~~ that real observations may have very intermittent temporal sampling. Nor does it consider the impact that precipitation may have on spatio-temporal variability of aerosol (Gryspeerdt et al., 2015, for example). These issues are the subject of further investigation.

Acknowledgements. This work was supported by the Natural Environmental Research Council grant nr NE/J024252/1 (Global Aerosol Synthesis And Science Project).

E. Gryspeerdt acknowledges funding from the European Research Council under the EU Seventh Framework Programme FP7-306284 ('QUAERERE'). D. Goto was supported by the Global Environment Research Fund S-12 of the Ministry of the Environment (MOE)/Japan, the Grant-in-Aid for Young Scientist B (Grant Number 26740010) of the Ministry of Education, Culture, Sports and Science and Technology (MEXT)/Japan, and KAKENHI/Innovative Areas (Grant Number 24110002) of MEXT/Japan. P. Stier acknowledges funding from the European Research Council under the EU Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement FP7-280025.

Michael Schulz and Svetlana Tsyro acknowledge funding from the Norwegian Research Council under the KLIMAFORSK project 'AeroCom-P3'. Their work was supported by EMEP under UNECE.

The figures in this paper were prepared using David W. Fanning's coyote library for IDL.

References

Ackermann, I. J., Hass, H., Memmesheimer, M., Ebel, A., Binkowski, F. S., and Shankar, U. M. A.: MODAL AEROSOL DYNAMICS MODEL FOR EUROPE : DEVELOPMENT AND FIRST APPLICATIONS, Atmospheric Environment, 32, 2981–2999, 1998.

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.

Ballester, J., Burns, J. C., Cayan, D., Nakamura, Y., Uehara, R., and Rodó, X.: Kawasaki disease and ENSO-driven wind circulation, Geophysical Research Letters, 40, 2284–2289, doi:10.1002/grl.50388, http://doi.wiley.com/10.1002/grl.50388, 2013.

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.

Benkovitz, C. M. and Schwartz, E.: Evaluation of modeled sulfate and SO2 over North America and Europe for four seasonal months in 1986-1987, J. Geophysical Research, 102, 25 305–25 338, 1997.

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.

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

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

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.

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

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.

Fast, J. D., Gustafson, W. I., Easter, R. C., Zaveri, R. A., Barnard, J. C., Chapman, E. G., Grell, G. A., and Peckham, S. E.: Evolution of ozone, particulates, and aerosol direct radiative forcing in the vicinity of Houston using a fully coupled meteorology-chemistry-aerosol model, Journal of Geophysical Research, 111, D21 305, doi:10.1029/2005JD006721, http://doi.wiley.com/10.1029/2005JD006721, 2006.

Fuzzi, S., Baltensperger, U., Carslaw, K., Decesari, S., Denier van der Gon, H., Facchini, M. C., Fowler, D., Koren, I., Langford, B., Lohmann, U., Nemitz, E., Pandis, S., Riipinen, I., Rudich, Y., Schaap, M., Slowik, J. G., Spracklen, D. V., Vignati, E., Wild, M., Williams, M., and Gilardoni, S.: Particulate matter, air quality and climate: lessons learned and future needs, Atmospheric Chemistry and Physics, 15, 8217–8299, doi:10.5194/acp-15-8217-2015, http://www.atmos-chem-phys.net/15/8217/2015/, 2015.

Goto, D., Dai, T., Satoh, M., Tomita, H., Uchida, J., Misawa, S., Inoue, T., Tsuruta, H., Ueda, K., Ng, C. F. S., Takami, A., Sugimoto, N., Shimizu, A., Ohara, T., and Nakajima, T.: Application of a global nonhydrostatic model with a stretched-grid system to regional aerosol simulations around Japan, Geoscientific Model Development, 8, 235–259, doi:10.5194/gmd-8-235-2015, http://www.geosci-model-dev.net/8/235/2015/, 2015.

Grell, G. A., Peckham, S. E., Schmitz, R., McKeen, S. A., Frost, G., Skamarock, W. C., and Eder, B.: Fully coupled "online" chemistry within the WRF model, Atmospheric Environment, 39, 6957–6975, doi:10.1016/j.atmosenv.2005.04.027, http://linkinghub.elsevier.com/retrieve/pii/S1352231005003560, 2005.

Gryspeerdt, E., Stier, P., White, B. A., and Kipling, Z.: Wet scavenging limits the detection of aerosol effects on precipitation, Atmospheric Chemistry and Physics, 15, 7557–7570, doi:10.5194/acp-15-7557-2015, http://www.atmos-chem-phys.net/15/7557/2015/, 2015.

Gustafson, W. I., Qian, Y., and Fast, J. D.: Downscaling aerosols and the impact of neglected subgrid processes on direct aerosol radiative forcing for a representative global climate model grid spacing, Journal of Geophysical Research: Atmospheres, 116, 1–28, doi:10.1029/2010JD015480, 2011.

Hakuba, M. Z., Folini, D., Sanchez-lorenzo, A., and Wild, M.: Spatial representativeness of ground-based solar radiation measurements - extension to the full Meteosat disk, Journal of Geophysical Research: Atmospheres, 16, 50 673, doi:10.1002/jgrd.50673., 2014a.

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

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

Hess, M., Koepke, P., and Schult, I.: Optical Properties of Aerosols and Clouds: The Software Package OPAC, Bulletin of the American Meteorological Society, 79, 831–844, doi:10.1175/1520-0477(1998)079<0831:OPOAAC>2.0.CO;2, 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.

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.

Kipling, Z., Stier, P., Schwarz, J. P., Perring, a. E., Spackman, J. R., Mann, G. W., Johnson, C. E., and Telford, P. J.: Constraints on aerosol processes in climate models from vertically-resolved aircraft observations of black carbon, Atmospheric Chemistry and Physics, 13, 5969–5986, doi:10.5194/acp-13-5969-2013, 2013.

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.: Evalua-

tion 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., Kirkevåg, 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.

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

Lequy, É., 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.

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.

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.

Myhre, G., Samset, B. H., Schulz, M., Balkanski, Y., Bauer, S., Berntsen, T. K., Bian, H., Bellouin, N., Chin, M., Diehl, T., Easter, R. C., Feichter, J., Ghan, S. J., Hauglustaine, D., Iversen, T., Kinne, S., Kirkevåg, A., Lamarque, J.-F., Lin, G., Liu, X., Lund, M. T., Luo, G., Ma, X., van Noije, T., Penner, J. E., Rasch, P. J., Ruiz, A., Seland, Ø., Skeie, R. B., Stier, P., Takemura, T., Tsigaridis, K., Wang, P., Wang, Z., Xu, L., Yu, H., Yu, F., Yoon, J.-H., Zhang, K., Zhang, H., and Zhou, C.: Radiative forcing of the direct aerosol effect from AeroCom Phase II simulations, Atmospheric Chemistry and Physics, 13, 1853–1877, doi:10.5194/acp-13-1853-2013, http://www.atmos-chem-phys.net/13/1853/2013/, 2013.

Pope, V. D. and Stratton, R. a.: The processes governing horizontal resolution sensitivity in a climate model, Climate Dynamics, 19, 211–236, doi:10.1007/s00382-001-0222-8, 2002.

Qian, Y., Gustafson, W. I., and Fast, J. D.: An investigation of the sub-grid variability of trace gases and aerosols for global climate modeling, Atmospheric Chemistry and Physics, 10, 6917–6946, doi:10.5194/acp-10-6917-2010, 2010.

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, Ø., 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, 8697–8717, doi:10.5194/acpd-9-12731-2009, 2009.

Randles, C. a., Kinne, S., Myhre, G., Schulz, M., Stier, P., Fischer, J., Doppler, L., Highwood, E., Ryder, C., Harris, B., Huttunen, J., Ma, Y., Pinker, R. T., Mayer, B., Neubauer, D., Hitzenberger, R., Oreopoulos, L., Lee, D., Pitari, G., Di Genova, G., Quaas, J., Rose, F. G., Kato, S., Rumbold, S. T., Vardavas, I., Hatzianastassiou, N., Matsoukas, C., Yu, H., Zhang, F., Zhang, H., and Lu, P.: Intercomparison of shortwave radiative transfer schemes in global aerosol modeling: results from the AeroCom Radiative Transfer Experiment, Atmospheric Chemistry and Physics, 13, 2347–2379, doi:10.5194/acp-13-2347-2013, http://www.atmos-chem-phys.net/13/2347/2013/, 2013.

Roeckner, E., Brokopf, R., Esch, M., Giorgetta, M., Hagemann, S., Kornblueh, L., Manzini, E., Schlese, U., and Schulzweida, U.: Sensitivity of Simulated Climate to Horizontal and Vertical Resolution in the ECHAM5 Atmosphere Model, Journal of Climate, 19, 3771–3791, 2006.

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

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

Schell, B., Ackermann, I. J., Hass, H., Binkowski, F. S., and Ebel, A.: Modeling the formation of secondary organic aerosol within a comprehensive air quality model system, J. Geophys. Res.-Atmos., 106, 28 275–28 293, doi:10.1029/2001JD000384, 2001.

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. 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.

Seinfeld, J. H. and Pandis, S. N.: Atmospheric chemistry and physics: from air pollution to climate change, John Wiley & Sons, Hoboken, New Jersey, 2nd edn., 2006.

Shinozuka, Y. and Redemann, J.: Horizontal variability of aerosol optical depth observed during the ARCTAS airborne exper-

iment, Atmospheric Chemistry and Physics, 11, 8489–8495, doi:10.5194/acp-11-8489-2011, 2011.

Simpson, D., Benedictow, A., Berge, H., Bergström, R., Emberson, L. D., Fagerli, H., Flechard, C. R., Hayman, G. D., Gauss, M., Jonson, J. E., Jenkin, M. E., Nyúri, A., Richter, C., Semeena, V. S., Tsyro, S., Tuovinen, J. P., Valdebenito, A., and Wind, P.: The EMEP MSC-W chemical transport model – Technical description, Atmospheric Chemistry and Physics, 12, 7825–7865, doi:10.5194/acp-12-7825-2012, 2012.

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., Schutgens, N. A. J., Bellouin, N., Bian, H., Boucher, O., Chin, M., Ghan, S., Huneeus, N., Kinne, S., Lin, G., Ma, X., Myhre, G., Penner, J. E., Randles, C. a., Samset, B., Schulz, M., Takemura, T., Yu, F., Yu, H., and Zhou, C.: Host model uncertainties in aerosol radiative forcing estimates: results from the AeroCom Prescribed intercomparison study, Atmospheric Chemistry and Physics, 13, 3245–3270, doi:10.5194/acp-13-3245-2013, http://www.atmos-chem-phys.net/13/3245/2013/, 2013.

Stroud, C. a., Makar, P. a., Moran, M. D., Gong, W., Gong, S., Zhang, J., Hayden, K., Mihele, C., Brook, J. R., Abbatt, J. P. D., and Slowik, J. G.: Impact of model grid spacing on regional- and urban- scale air quality predictions of organic aerosol, Atmospheric Chemistry and Physics, 11, 3107–3118, doi:10.5194/acp-11-3107-2011, 2011.

Swap, R., Garstang, M., Greco, S., Talbot, R., and Kallberg, P.: Saharan dust in the Amazon Basin, Tellus, 44B, 133–149, doi:10.1034/j.1600-0889.1992.t01-1-00005.x, 1992.

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., Kirkevåg, A., Kristjansson, J. E., Krol, M., Lauer, A., Lamarque, J. F., Liu, X., Montanaro, V., Myhre, G., Penner, J., Pitari, G., Reddy, S., Seland, Ø., 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., Kirkevåg, 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, Ø., 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.

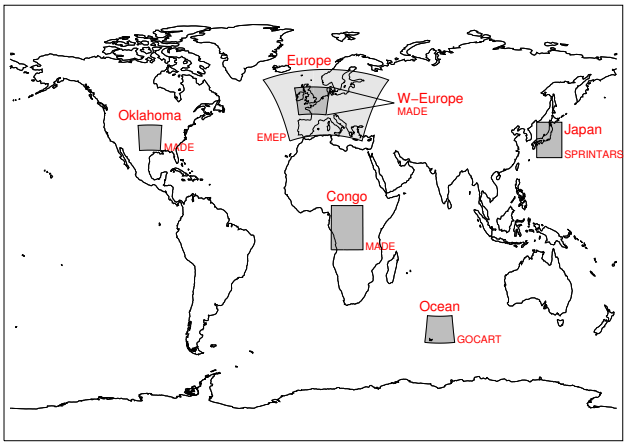

Figure 1. Three models were used in this study to simulate a variety of aerosol fields. The regional names used to identify these simulations are given in large font, while the models are denoted in small font. MADE and GOCART refer to the WRF-Chem version used.

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.

Weigum, N., ~~Stier, P., and~~ Schutgens, ~~N~~N.A.J., and Stier, P.: ~~Modelling of stratocumulus cloud layers in a large eddy simulation model with explicit microphysics~~Effect of aerosol sub-grid variability on aerosol optical depth and cloud condensation nuclei: Implications for global aerosol modelling, submitted to Atmospheric Chemistry and Physics~~, in preparation, 2015.~~.

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

Williamson, D. L.: Convergence of aqua-planet simulations with increasing resolution in the Community Atmospheric Model, Version 3, Tellus A, 60, 848–862, doi:10.1111/j.1600-0870.2008.00339.x, http://tellusa.net/index.php/tellusa/article/view/15499, 2008.

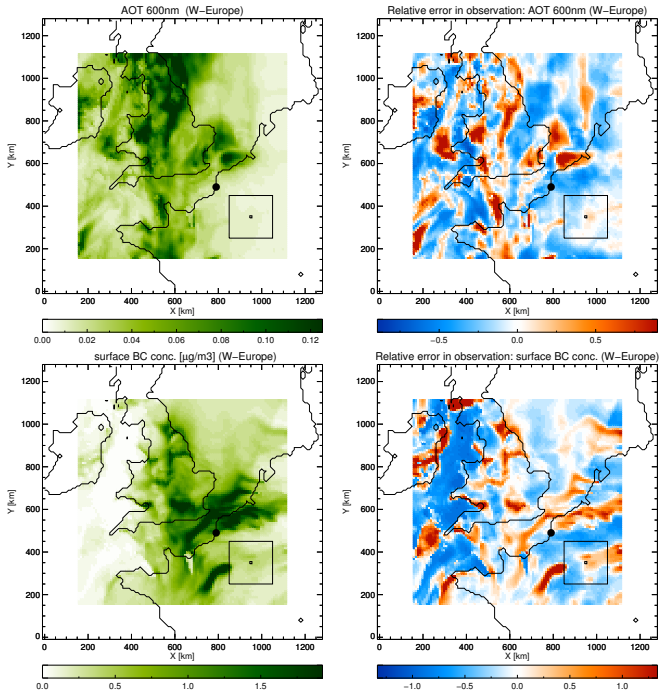

Figure 2. Snapshots of the simulated field and the relative spatial sampling error in the observation of AOT and surface black carbon concentration, ~~as simulated~~ over W-Europe ~~at~~ exactly 10 days ~~, 00 hours~~ into the simulation by WRF-Chem MADE. Also shown are two square boxes (10×10 and 210×210 km) and a single location (fat dot), south of Calais, France. Note that the high-resolution simulations encompass the whole region shown, while our analysis is only made for the colored domain.

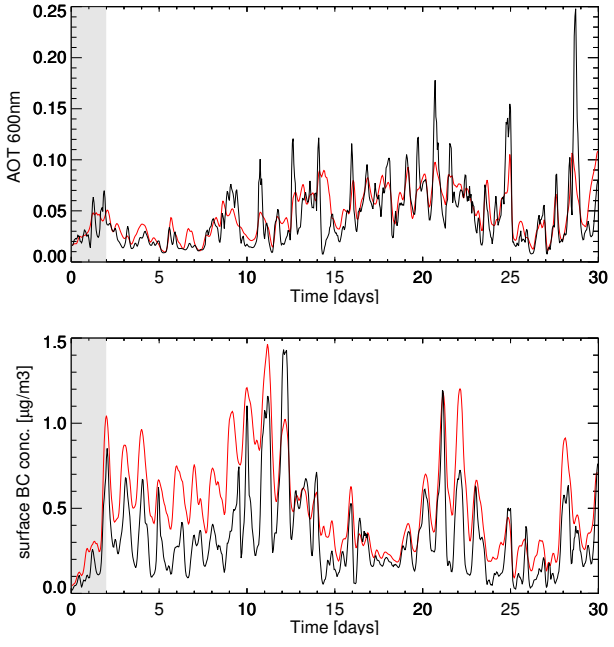

Figure 3. Timeseries of global model (red) and observed (black) AOT and surface black carbon concentration as simulated at a location south of Calais (France) by WRF-Chem MADE, see also Fig. 2. The grey ~~bar~~ area to the left shows the model's spin-up period.

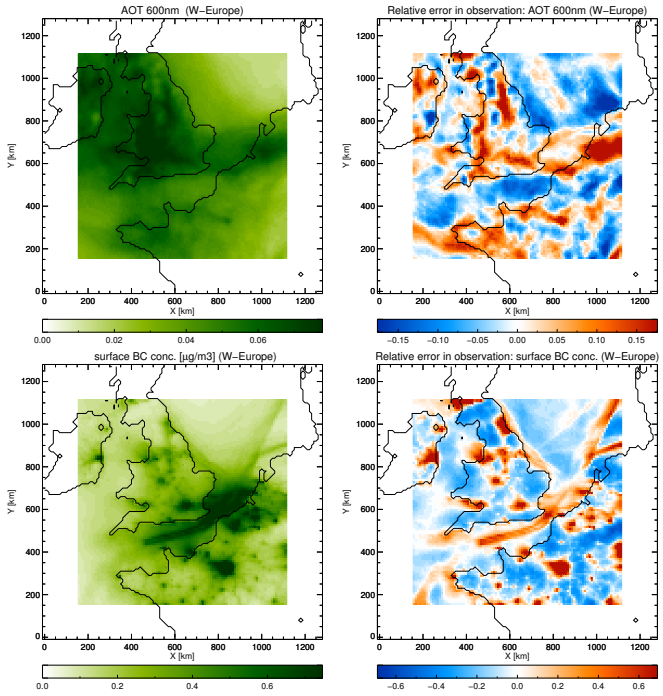

Figure 4. Monthly average of the simulated field and the relative spatial sampling error in the observation of AOT and surface black carbon concentration, as simulated over W-Europe by WRF-Chem MADE. Note that the high-resolution simulations encompass the whole region shown, while our analysis is only made for the colored domain.

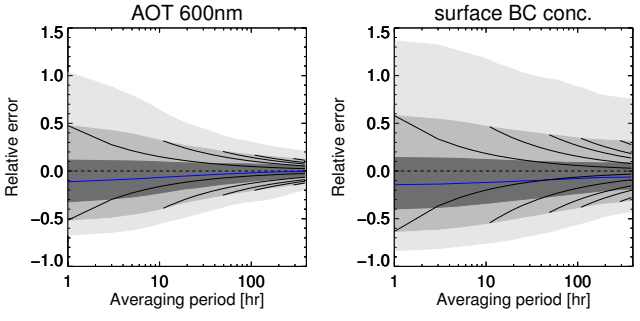

Figure 5. Relative spatial sampling error as a function of averaging period. The thin black lines are prognosis of the 9 and 91% quantiles *in case* these errors behaved like independent Gaussian errors (i.e. $1/\sqrt{n}$, with n the number of observations). Results from WRF-Chem MADE over W-Europe. Further explanation in Sec. 3.2.

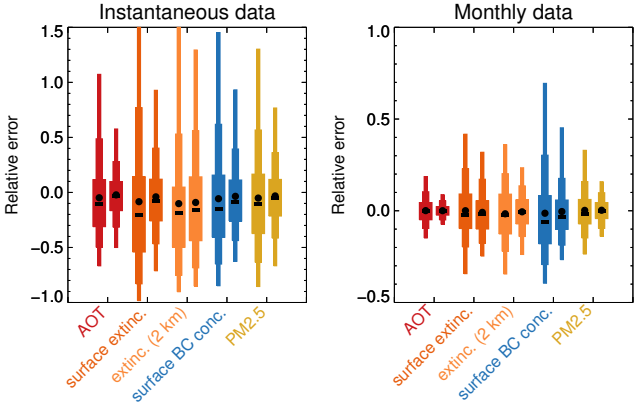

Figure 6. Relative spatial sampling errors (for either instantaneous or monthly data, note the different vertical axes) over the W-Europe region as calculated by WRF-Chem MADE (left bar) and EMEP (right bar) in May 2008. Further explanation in Sec. 3.2.

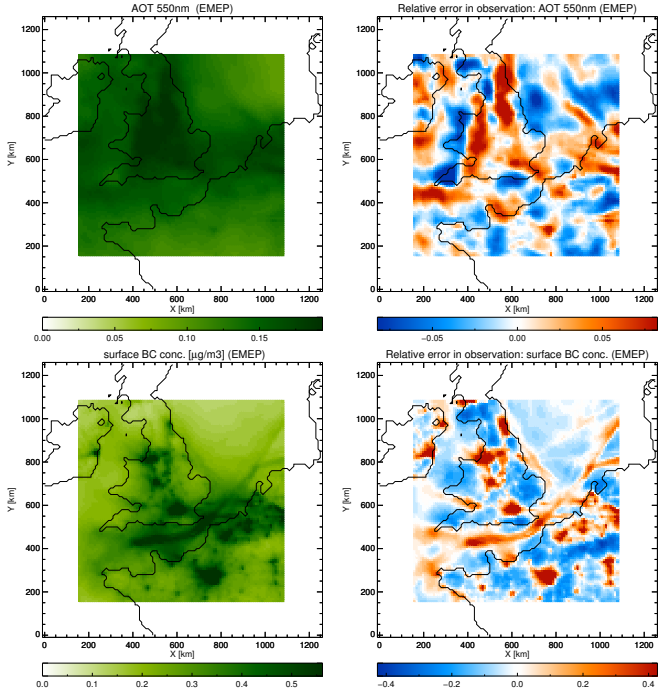

Figure 7. Monthly average of the simulated field and the relative spatial sampling error in the observation of AOT and surface black carbon concentration, as simulated over W-Europe by EMEP. This can be compared to results for WRF-Chem MADE as shown in Fig. 4 but note that the colour bars have different ranges to bring out spatial patterns better.

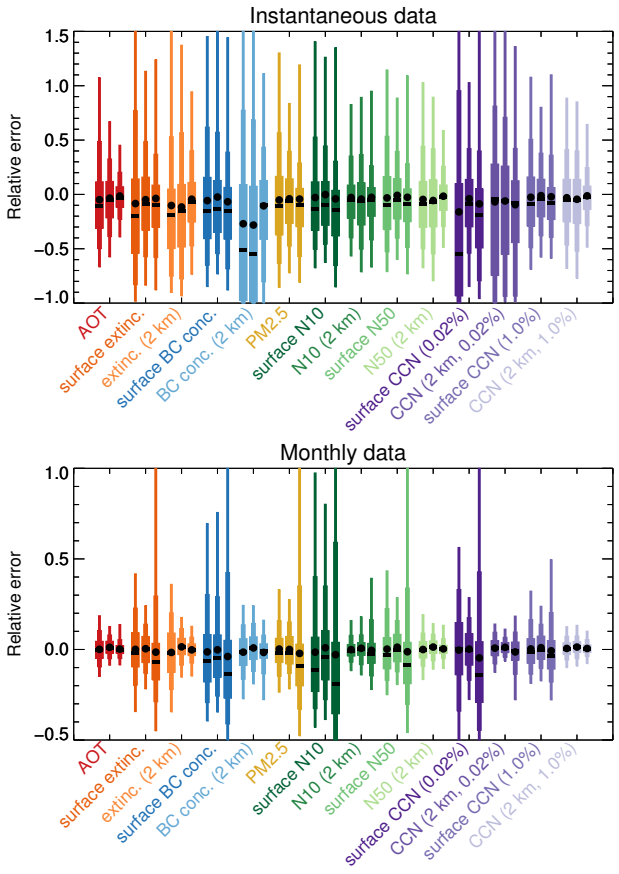

Figure 8. Relative spatial sampling errors (for either instantaneous or monthly data, note the different vertical axes) for all WRF-Chem MADE regions (left bar: W-Europe; centre bar: Oklahoma; right bar: Congo). Further explanation in Sec. 3.2.

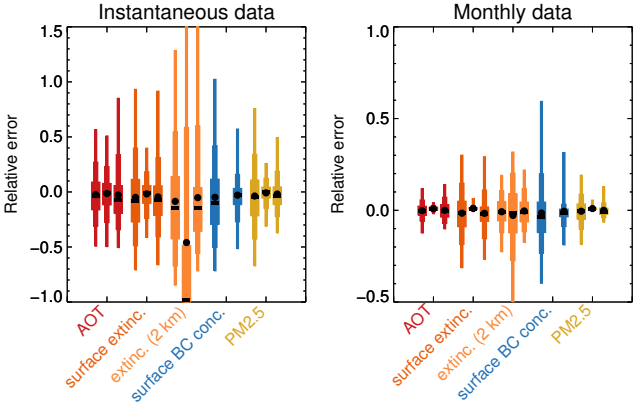

Figure 9. Relative spatial sampling errors (for either instantaneous or monthly data, note the different vertical axes) for three regions simulated with mass-bulk schemes (left bar: Europe; middle bar: Ocean; right bar: Japan). Black carbon concentrations over Ocean are zero and so are related spatial sampling errors. Further explanation in Sec. 3.2.

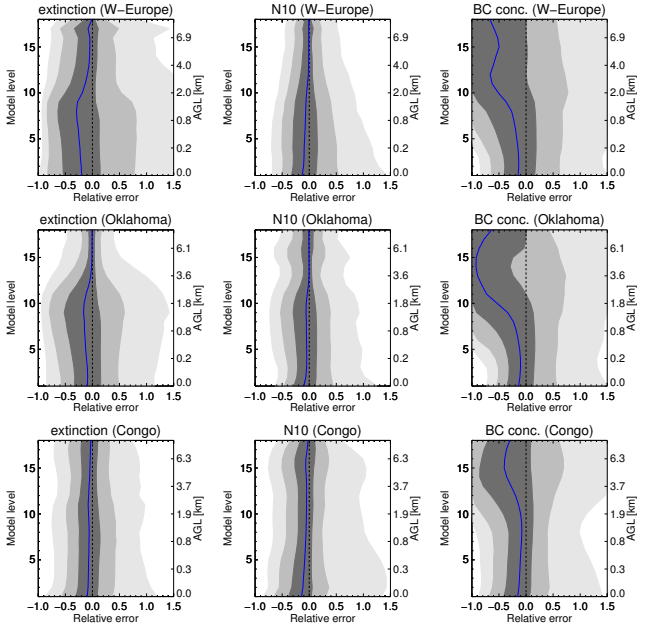

Figure 10. Relative spatial sampling error (instantaneous data) as a function of model level (left vertical axis) and altitude above ground level (AGL, right vertical axis) for extinction, N10 and black carbon concentrations. Results for the WRF-Chem MADE simulations. Further explanation in Sec. 3.2.

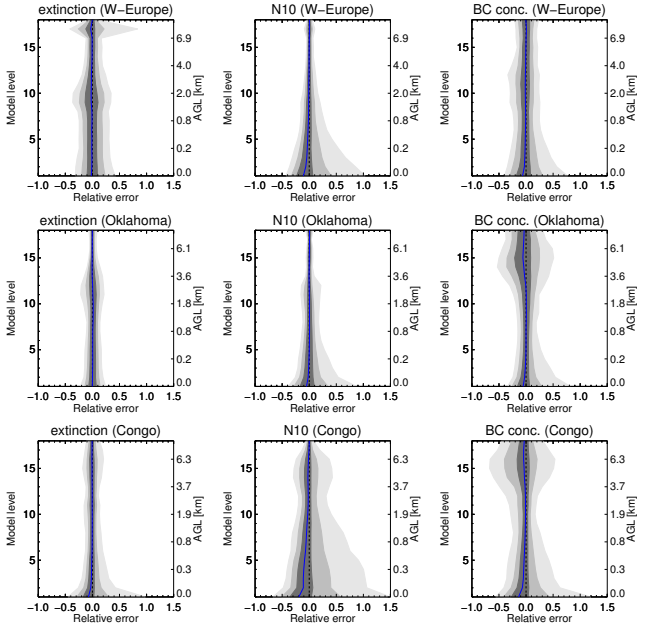

Figure 11. Relative spatial sampling error (monthly data) as a function of model level (left vertical axis) and altitude above ground level (AGL, right vertical axis) for extinction, N10 and black carbon concentrations. Results for the WRF-Chem MADE simulations. Further explanation in Sec. 3.2.

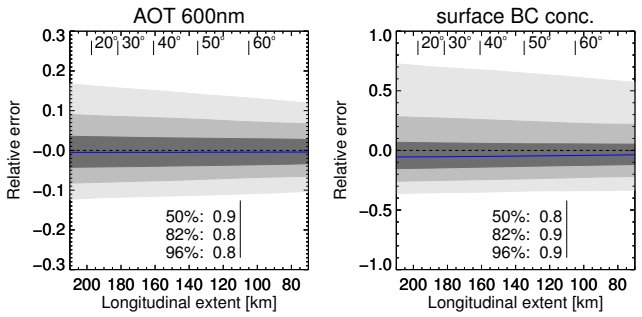

Figure 12. Relative spatial sampling errors (monthly data) as a function of longitudinal extent of the grid-box (due to latitude). Near the top horizontal axis, latitudes are given. Near the bottom horizontal axis, the ratios of $\Delta q_{25}, \Delta q_{82}$ and Δq_{96} at two different longitudinal extents (110 over 210 km) are given. Results from WRF-Chem MADE over W-Europe. Further explanation in Sec. 3.2.

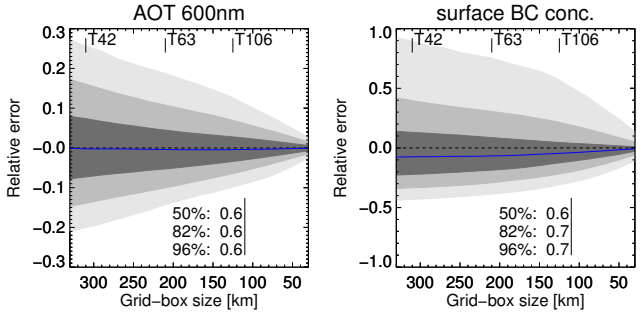

Figure 13. Relative spatial sampling errors (monthly data) as a function of grid-box size. Near the top horizontal axis, standard spectral grid sizes are shown. Near the bottom horizontal axis, the ratios of $\Delta q_{25}, \Delta q_{82}$ and Δq_{96} at two different grid-box sizes (110 and 210 km) are given. Results from WRF-Chem MADE over W-Europe. Further explanation in Sec. 3.2.

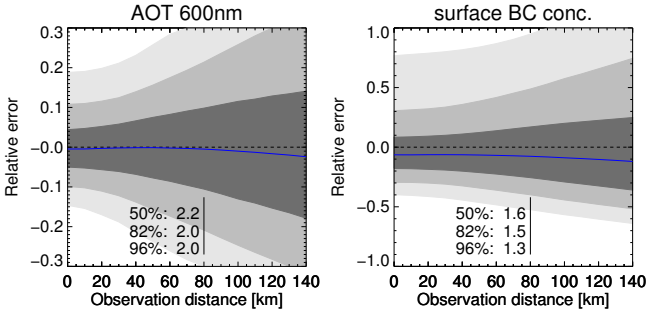

Figure 14. Relative spatial sampling error (monthly data) as a function of distance of the observation to the grid-point. Near the bottom horizontal axis, the ratios of $\Delta q_{25}, \Delta q_{82}$ and Δq_{96} at a distance of 80 and 0 km are given. Results from WRF-Chem MADE over W-Europe. Further explanation in Sec. 3.2.

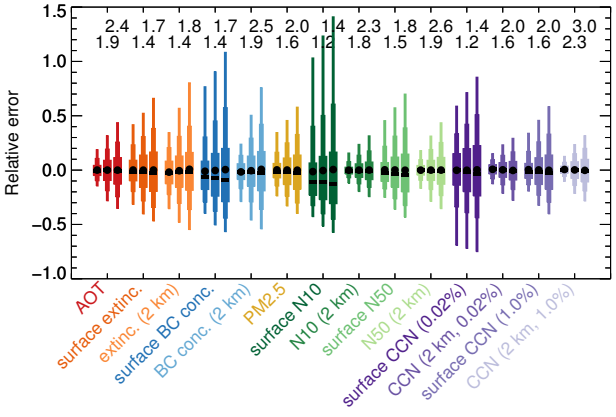

Figure 15. Relative spatial sampling error (monthly data) as a function of distance of the observation to the grid-point. The numbers near the top horizontal axis show the increase of Δq_{82} at resp. 70 and 100 km relative to 0 km. Results from WRF-Chem MADE over W-Europe. Further explanation in Sec. 3.2.

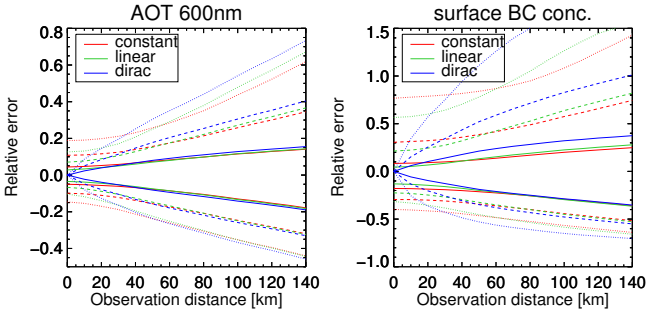

Figure 16. Relative spatial sampling error (monthly data) as a function of distance of the observation to the grid-point, for three different weighting functions. Results from WRF-Chem MADE over W-Europe. The usual inter-quantile ranges Δq_{50} (solid), Δq_{82} (dashed) and Δq_{96} (dotted) are shown.

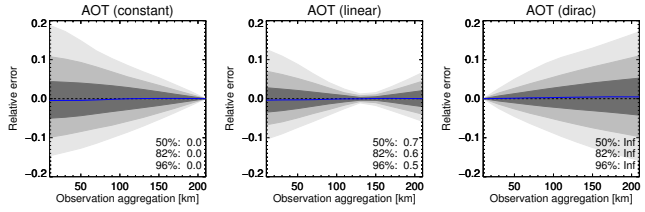

Figure 17. Relative spatial sampling error (monthly data) as a function of aggregation extent of the AOT observations, using three different weighting functions. The centre of the aggregated observations is assumed to coincide with the model's grid-points. In the lower right corner, the ratios of $\Delta q_{25}, \Delta q_{82}$ and Δq_{96} at two different aggregation extents (210 to 0 km) are given. Results from WRF-Chem MADE over W-Europe. Further explanation in Sec. 3.2.

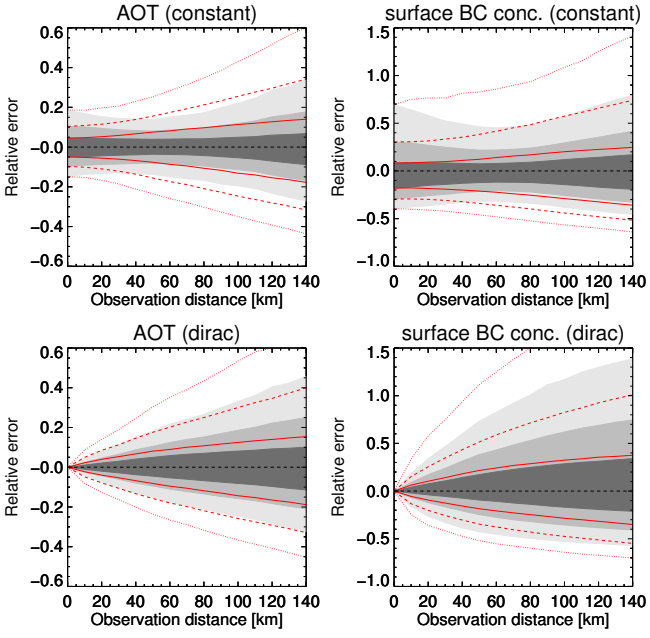

Figure 18. Relative spatial sampling error (monthly data) for 4 randomly distributed sites as a function of distance to the grid-point, assuming two different weighting functions. The red lines indicate the errors for a single site (see also Fig. 14). Results from WRF-Chem MADE over W-Europe. Further explanation in Sec. 3.2.

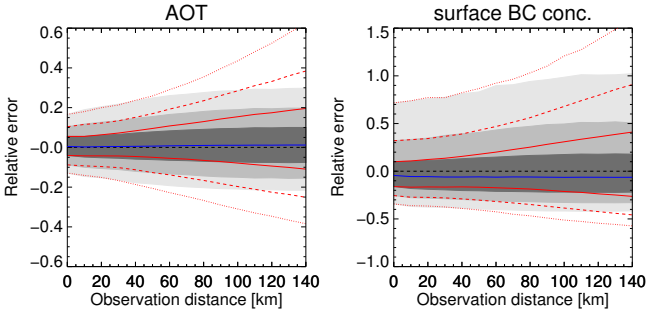

Figure 19. Relative spatial sampling error (monthly data) i.c. of linear interpolation of model values to the observation, as a function of distance to the grid-point. The red lines indicate the errors without interpolation (see also Fig. 14). Results from WRF-Chem MADE over W-Europe. Further explanation in Sec. 3.2.

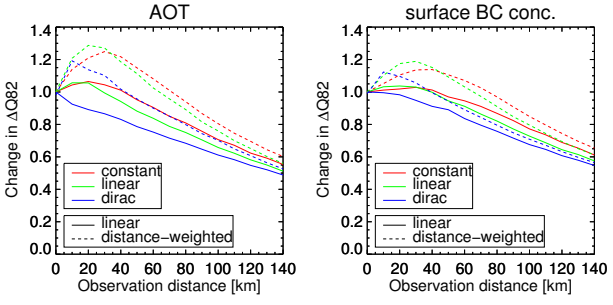

Figure 20. Change (relative to Fig. 14) in Δq_{82} (for monthly relative sampling errors) due to interpolation, as a function of distance to the grid-point. All three weighting functions and two interpolation methods are considered. Similar graphs for Δq_{50} and Δq_{96} can be shown.

Table 1. Simulations analysed in this study

region	size [km^2]	period	model	scheme	comments
W-Europe	1280×1280	May 2008	WRF-Chem	MADE	2-moments modal
Oklahoma	1190×1190	March 2007	WRF-Chem	MADE	2-moments modal
Congo	2090×2090	March 2007	WRF-Chem	MADE	2-moments modal
Ocean	1270×1270	March 2007	WRF-Chem	GOCART	mass bulk
Europe	4000×3100	January - June 2008	EMEP		mass bulk
Japan	1500×1250	August 2007	NICAM	SPRINTARS	mass bulk

Table 2. Simulated observables

	AOT	extinction	$PM_{2.5}$	BC conc.	N10, N50	CCN
WRF-Chem MADE	✓	✓	✓	✓	✓	✓
WRF-Chem GOCART	✓	✓	✓			
EMEP	✓	✓	✓	✓		
NICAM-SPRINTARS	✓	✓	✓	✓		

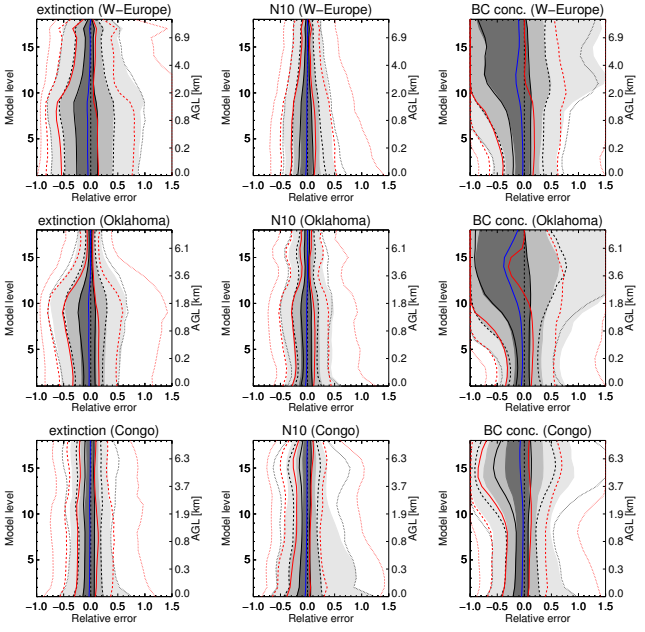

Figure 21. Relative spatial sampling error (for measurements during horizontal legs of a flight campaign) as a function of model level (left vertical axis) and altitude above ground level (AGL, right vertical axis) for extinction, N10 and black carbon concentrations. The grey shaded error ranges are for North-South flights. Similar error ranges for East-West flights are shown in black lines. The results of Fig. 10 are also shown in red lines. The usual inter-quantile ranges Δq_{50} (solid), Δq_{82} (dashed) and Δq_{96} (dotted) are shown. Further explanation in Sec. 3.2.