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