We thank the reviewers for their careful reviews and helpful comments. The manuscript has been revised accordingly and our point-by-point responses are provided below. (Reviewer's comments are in italic and the responses in standard font).

Reviewer #2

The manuscript presents MAM4, an extended version of the three-mode MAM3 aerosol scheme incorporating an additional externally-mixed mode for primary carbonaceous aerosol. This allows it to represent the delayed transition to the hydrophilic state via ageing processes, but without the full complexity of the seven-mode MAM7 scheme. Improvements to the aerosol distribution are well demonstrated, and also compared with those from increasing model resolution.
The manuscript is well presented, based on sound methodology and shows clear results. Subject to the comments below, I would certainly recommend it for publication in GMD.

1 General comments:

Since MAM4 is positioned between MAM3 and MAM7 in terms of its complexity, it is a pity that it is only compared against the former. An additional run with MAM7 in set 1 would demonstrate how much of the improvement in MAM7 is achieved by MAM4 with its lower computational cost.

Reply: We conducted an experiment with MAM7. Since MAM4's primary carbon mode is the same as that in MAM7, the difference in BC and POM simulations between MAM4 and MAM7 is negligible. Thus we don't include the results of MAM7 in the comparison. We have made this clear in the model experiment section 3 and conclusion section 5.

For the set 2 experiments, the model is constrained by the YOTC analysis. Does this cover the time period of all the observational campaigns used, in order that these evaluations can be performed on the correct year? If not, how do the authors deal with interannual meteorological variability when comparing simulations for one year with observations for another?

Reply: We agree with the reviewer that the interannual meteorological variability is not accounted for in this evaluation strategy. Unfortunately the YOTC analysis is only available between May 2008 and May 2010, while some field campaigns used in this study were conducted in other years (e.g., NASA ARCTAS and NOAA ARCPAC campaigns took place in April 2008). However, our purpose is to present the new improvement in MAM4 with two sets of sensitivity experiments using the same evaluation strategy as in Liu et al. (2012) for MAM3 and MAM7 to highlight the importance of the BC/POM aging process and model resolution. We have acknowledged the potential effect of interannual meteorological variability on our model-observation comparisons in the revised manuscript.

Similarly, could the authors please clarify the temporal nature of the biomass-burning emissions in particular? These are highly variable, and there is likely to be a large *difference between using climatological or hindcast-style "correct year" emissions when comparing to specific campaigns.*

Reply: We agree with the reviewer that the time-dependent nature of biomass-burning emissions can affect our model-observation comparisons for specific campaigns. Our experiments are designed to be consistent with Liu et al. (2012), to highlight the importance of the BC/POM aging process and model resolution. Our set 1 experiments were conducted using the present-day (i.e., year 2000) climate forcing conditions, including aerosol emissions, instead of AMIP-type configurations. Thus the interannual variability for biomass-burning emissions is not accounted for in this study. We have acknowledged the effect of interannual variability of biomass-burning emissions on our model comparisons to specific campaigns in the revised manuscript.

We checked the interannual variability of biomass-burning emissions from GFED version 3.1 (van der Werf et al., ACP, 2010) for the time period of year 2004 to 2011. The annual fire BC plus OC emissions in boreal regions (including N. America and Russia) vary by more than a factor of 2 from 2007 with the lowest emissions to 2008 with the strongest emissions (Figure R1). However, the interannual variability of biomass burning emissions still can not account for the large difference between model simulations and observations (by more than a factor of 10 in the Arctic).

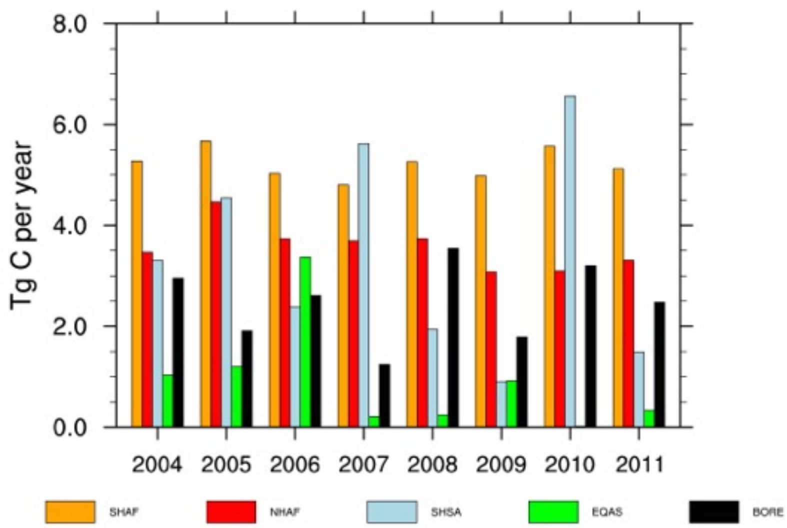

Figure R1. Annual BC plus OC emissions in different regions from 2004 to 2011 from the GFED emission inventory. SHAF: southern hemisphere Africa; NHAF: northern hemisphere Africa; SHSA: southern hemisphere southern America; EQAS: equatorial Asia; BORE: boreal region (in both N. America and Russia).

In the budget analysis presented in Section 4.2, is it possible to quantify the interannual variability in these budgets and thus to determine whether or not the differences are statistically significant?

Reply: It is a great idea and we have added the standard deviation (from the 10-year average) in the budget analysis. We have added the standard deviations in Tables 2-3 for the set 1 experiments in the revised manuscript.

When comparing the model to the aircraft campaigns in Figures 9–12, how is the model sampled in space and time to match the observations? (Is it interpolated to the flight time and location, or are e.g. monthly means over some region used? This will affect the sampling error to be expected in the comparison.)

Reply: The model results are sampled in the same way as Liu et al. (2012): Simulated profiles are averaged over the points on the map and the indicated month of each campaign. This sentence has been added to the figure captions of Figures 7-12.

2 Detailed comments

Page 8344, lines 3–5. If the models compare better with ground-based sun photometer measurements than with satellite retrievals, doesn't this suggest that the poorer comparison with satellite might be due in part to either retrieval errors or collocation/sampling issues between models and observations, rather than deficiencies in the models?

Reply: This is indeed a possibility that certainly requires further investigation. We have added the reviewer's comments in the manuscript and call for further studies.

Page 8344, lines 6–9. Hydrophobic and water-insoluble are not quite the same thing. Insoluble materials may nevertheless be hydrophilic ("wettable") and thus act as CCN via adsorption, although this is rarely treated in models. Also, "and are not able to nucleate cloud droplets" is unnecessary – this is what "cannot serve as CCN" means.

Reply: Following the reviewer's comment, we have removed "water insoluble" and "and are not able to nucleate cloud droplets" in the revision.

Page 8345, line 20. Remove hyphen in "high latitudes" and insert "the" before "northern hemisphere".

Reply: Done.

Page 8346, line 21. Please explain how homogeneous nucleation might be affected by aerosol particles. Unlike for heterogeneous nucleation, this is not clear.

Reply: homogeneous nucleation is the spontaneous freezing of aerosol solution droplets such as sulfate aerosol. Change in the sulfate aerosol number can affect the number and size of ice crystals from homogeneous nucleation due to the competition of water vapor among the ice crystals.

We have added "of aerosol solution droplets (e.g., sulfate aerosol)" after "homogeneous ice nucleation" in the revision.

Page 8348, line 23. Change "standard-alone" to "stand-alone" and explain what is meant by this. I presume an uncoupled atmosphere-only simulation with prescribed SST and sea ice?

Reply: We made the correction, and added "atmosphere-only simulation with prescribed SST and sea ice". Yes, this is what the "stand-alone" means.

Page 8349, line 23. By "specified dynamics" do you mean what is often referred to as "nudging" (i.e. Newtonian relaxation of model fields to the (re)analysis ones)? It's probably worth either using the term "nudging", or explaining how the technique differs.

Reply: Yes, they are the same thing. We acknowledge that there are different terminologies and have added "also known as the nudging technique" in the revised manuscript.

Page 8350, line 9. Delete "the" before "comparison".

Reply: Done.

Page 8350, line 12, and captions to Figures 2–5. "Latitude and longitudinal" is grammatically inconsistent. I'd suggest "zonal and meridional" or "latitudinal and longitudinal" instead.

Reply: Done. Changed to "latitudinal and longitudinal".

Page 8351, line 6. Although primary carbon particles may not act as CCN, they are still subject to impaction scavenging by cloud droplets, ice crystals and falling precipitation, leading to wet deposition. Please clarify if this process is included in the model or not.

Reply: Yes, the impaction scavenging by precipitating hydrometeors is included in the model. We have added the sentence "(while still subject to impaction scavenging by precipitating hydrometeors)" in the revised manuscript. The primary carbon particles are in the dry diameter range of 0.04-0.2 μm calculated from our model (Liu et al., 2012). Particles in this size range, even if hydroscopic, are less efficiently scavenged by the precipitating hydrometeors than by the in-cloud nucleation scavenging.

Page 8351, lines 28–29. Why is this of importance for aerosol–cloud interactions in particular?

Reply: Ma et al. (2015) found that increasing resolution leads to enhanced but less frequent droplet nucleation (Ma et al., 2015) due to (1) stronger subgrid vertical velocity, (2) reduced collocation of aerosols and clouds, and (3) higher aerosol concentration. We have revised the sentence for clarification. Now it reads as "The resolution sensitivity of POM/BC, as well as other aerosol particles, can contribute to the resolution sensitivity of aerosol-cloud interactions such as the enhanced but less frequent droplet nucleation due to stronger subgrid vertical velocity, reduced collocation of aerosols and clouds, and higher aerosol concentration (Ma et al., 2015)."

Page 8352, lines 6–11. Please justify, quantify and show evidence for, the difference between SD and free-running simulations being due to a convectively less stable atmosphere.

Reply: This is from our previous study but since it is not necessary to this study, we removed the statement in the revised manuscript: "In particular, we find that the free-running CAM5 produces a convectively less stable atmosphere, which enhances the vertical transport of aerosols to higher altitudes where the aerosol lifetimes are longer."

Page 8352, lines 17–18. This gives the impression that dry deposition overtakes wet deposition as the dominant process; please clarify that wet deposition remains very much the dominant sink even at 8 mono-layers, although dry deposition becomes a larger secondary sink.

Reply: Following the reviewer's comment, we changed the sentence to "the contribution of wet deposition to the total BC sink somewhat decreases and the contribution of dry deposition increases" in the revised manuscript.

Page 8353, line 23. Inserting "absolute" before "wet removal sinks" would make it clear how these statements are consistent.

Reply: Done.

Page 8354, lines 7–8. A citation of Schwarz et al. (2013; 10.1002/2013GL057775) is probably in order for the HIPPO1–5 SP2 observations.

Reply: Done. Citation added in the revised manuscript.

Page 8355, lines 3–5. A similar attribution of excess BC in the upper troposphere to the relationship between convective transport and scavenging has been done for other models, notably HadGEM3–UKCA (Kipling et al., 2013; 10.5194/acp-13-5969-2013).

Reply: The citation has been added in the revision to "Excessive BC aloft was also found in the HadGEM3–UKCA model and was attributed to the coupling of convective transport and convective scavenging (Kipling et al., 2013)."

Pages 8356–8357, and/or Figures 9 and 11. Please add references for the aircraft campaigns used, where available.

Reply: References are added for the aircraft campaigns in the revised manuscript.

Page 8358, line 7. Should be either "which however weakens" or "which, however, weakens".

Reply: Done.

Page 8359, lines 23–25. Citation or evidence for these deficiencies in the emission inventory?

Reply: Following the reviewer's comment, we have added citations for the deficiencies of aerosol emission inventory in and around the Arctic (e.g., Stohl et al., 2013) and in East Asia (e.g., Cohan and Wang, 2014).

Page 8360, lines 14–16. Probably worth using the AR5 terminology explicitly, i.e. either "ERFaci" or "ERFari+aci", depending which this is.

Reply: Done. We changed to use the AR5 terminology $ERF_{ari+aci}$.

Table 1 caption. Try "number of ageing mono-layers set to 8", and ". . . with MAM3 is run for the comparison".

Reply: Done.

Figures 2, 4, 6, 7–12. Font sizes are very small, although this may only be due to the reduction required to fit in discussion page layout. Please check that labels etc. will be easily legible in the final paper.

Reply: Thanks for the suggestion. We will check the readability of labels etc. in these figures in the final paper.

Figure 3. Using the same colour scale for a difference plot as for the absolute burdens is confusing. Also, relative difference ((x1-x0)/x0) on a logarithmic scale seems strange. Either relative difference on a linear scale, or ratio (x1/x0) on a logarithmic scale would be easier to interpret.

Reply: We are using the same color scale for the relative differences between different MAM4 experiments and MAM3, so as to compare the differences more clearly. We prefer plotting the relative difference with logarithmic scale to capture the large spread of the relative difference for different MAM4 experiments. We think it is easy to understand.

Figures 7–12. The standard-deviation shading looks very odd, often being highly asymmetric and reaching down to zero. It's also not described in the caption of Figure 7. This is probably the result of using arithmetic standard deviations on a logarithmic scale – on such a scale, and for a quantity which roughly follows a log-normal distribution, the geometric mean and standard deviation would be more appropriate (or alternatively median and interquantile range). Including both means and medians on the plot doesn't *seem to add much, and makes them quite cluttered. I'd suggest sticking with one, and removing the other, unless some important conclusion relies on the distinction.*

Reply: Following the reviewer's comment, we removed dark dashed curves for the medians in Figure 7-12. We also added the description of means and standard deviations in the caption of Figure 7.