Review of "An evaluation of global organic aerosol schemes using airborne Observations"

Two organic aerosol schemes in GEOS-chem are evaluated against observations. A comprehensive set of aircraft measurements of aerosol composition, mainly using the ToF-AMS instrument, is assembled and classified into regimes in which the model is evaluated. A simple scheme with non-volatile primary organic aerosol and with fixed yields for secondary organic aerosols is compared with a complex scheme, in which semivolatile primary and secondary organics are produced in a volatility basis set framework and aqueous reactive uptake for isoprene is included. The performance of the two schemes is similar. High-resolution (or higher-resolution, at least) nested simulations and some tweaks to the simple scheme are used to gain further insights into the behavior of the mechanisms. The paper builds on previous work by many of the same authors, published in Heald et al 2011.

The paper is a valuable addition to the literature. It is well-written, insightful, and will attract attention from the community. I recommend it for publication in ACP. I have the following minor suggestions for improvements.

Specific comments:

Introduction: I think it's worth remarking (with appropriate references) that the "simple scheme" is fairly similar to what is used in most climate and earth system models submitted to CMIP6 (NorESM, HadGEM3/UKESM1, I think also GFDL AM4), while a couple of climate models (CESM and E3SM as far as I know) have configurations that are more similar in at least some important respects to the complex scheme (such as including the volatility basis set for semivolatile SOA).

L103 "we perform a series of simulations from 2008 to 2017 using two distinct model scheme": It would be helpful to include a table summarizing the simulations that were performed (one simple, one complex, then some simulations in which the simple scheme was modified, the simple SOA with complex POA, etc), and a more detailed explanation of which periods were simulated – just the times of the flights, or the whole year. Also, at or around L223: "The observations were averaged over the model grid-boxes and timestep." It would be great to be a bit more explicit about how the comparison was done - for example, was the model output also diagnosed and written out to a file on every timestep, or every few hours?

L112 "A standard bulk aerosol scheme": which one? Also please put into context the subsequent sentence "GEOS-chem also simulates sulfate aerosol…" – is this somehow a separate issue from the bulk aerosol scheme?

Figure 4: It's rather unfortunate that the differences between the two schemes are greatest in areas where you have no measurements: central Africa and inland in China. This is pointed out in the conclusions, but perhaps Africa should be added to the list on line 612 – you could also point out that there do exist some datasets that might already help with resolving the large discrepancies there (DACCIWA, ORACLES, for example)?

"The explicit aqueous uptake mechanism for the isoprene-derived SOA products also results in substantially larger global isoprene SOA burdens (0.31 Tg) when compared to the 'pure-VBS' treatment of isoprene-derived SOA that simulates an annually averaged ISOA burden of 0.12 Tg" - -so was there an additional simulation performed with the aqueous uptake turned off? Can you be a bit more detailed about what differences this causes, maybe adding

another row to Table 2 where isoprene SOA is split into aqueous and non-aqueous contributions?

L295. It is stated in the abstract that the model skill is superior to previous model evaluations, and in this section at line 295 the model is compared to an ensemble from Tsigaridis et al 2014. However, the reasons why the model differs from the ensemble probably vary from model to model. For the GEOS-chem model, the authors already include some comments about how the current model differs from that in Heald et al 2011 at lines 427 and 438. Can the authors identify whether it is changes to the emissions inventories since the 2011 paper, or changes to the OA schemes, that are responsible for the differences? Also, perhaps it is worth saying why some of the campaigns from Heald et al 2011 were used in the current study, but not others (presumably this was just to avoid running the simulations for unfeasibly many years)?

L343: The regime analysis is interesting and very useful in the following interpretation, but needs some further explanation, or possibly further tuning of the classifications, because some features of Figure 5 are rather surprising. In Figure 5, many regions that must be at least relatively pristine compared to the eastern United States are categorized as anthropogenic (large portions of the North Atlantic and North Pacific ATOM flights, much of the Canadian Arctic) and perhaps this can explain the sentence "Median concentrations over anthropogenic regions are markedly lower than those over other sources"?

Then, it looks like most of the eastern USA is classified as "remote". Is "remote" being plotted on top of "anthropogenic" for example, so the high volume of data would cause misleading results where only the last plotted regime shows up? Or is the North Atlantic characterized as anthropogenic because the aerosol mass concentration can be quite high due to a lot of dust (and the North Pacific, potentially, due to volcanic sulfur). Could other aerosol types have been included in the 'aggregate OA mass concentration' threshold of 0.2mugsm-3?

The classification would presumably be quite different if the figure was remade using regime types from the complex scheme rather than the simple scheme. Perhaps this would be interesting to try, just to reproduce Figure 5 (no need to repeat the whole analysis!) It would overcome the shortcoming of the simple scheme already mentioned in the text that it tends to count (for example) anthropogenically influenced biogenic SOA as anthropogenic SOA.

Another way to make this figure 5 less surprising might be to introduce separate categories for 'remote anthropogenic' and 'polluted anthropogenic' based on another mass threshold.

Figures 7 and 8 show that in the remote/marine region, the two schemes also disagree radically on whether the aerosol is primary or secondary, above 2km altitude. This is well discussed in the first paragraph of the conclusion but also seems worthy of greater emphasis and discussion in the paper around line 505. It is remarkable how well both schemes reproduce observations despite this (at least in Figure 7 and in summer, and even the lack of variability noted at line 378 does not seem to be a large effect in Figure S2). It makes sense that semivolatile POA gets to high altitude more effectively than non-volatile POA, so it stands to reason that the complex scheme is doing well. So then it seems surprising that overall the model with complex POA and simple SOA, from fig S7, seems to underestimate OA in the remote region (negative NMB)- if I understand how the NMB is defined, it should overestimate it. Similarly, the reverse arrangement, with simple POA and complex SOA, should underestimate, but the NMB is positive. What does the altitude profile look like?

There are several ways to calculate NMB – please can the authors include an equation somewhere in the text to define it?

In Figure 8, ATOM-2-W shows both models substantially overestimate SOA at high altitude, while ATOM-1-W is fine. This is explained in the text as a seasonal effect, L455. Does the overestimate square with the near-perfect agreement in Figure 7 remote/marine?

I realise this is outside the scope of the current study, but do the authors intend to make use of a fuller range of capabilities of the ToF-AMS in tracking signatures of different aerosol sources – for example signatures of biomass burning (f60), SOA (f44) etc, relevant fragment ratios, etc, etc in future work? Or even PMF factors? I'm not an instrumentation expert, but my understanding is that the ToF-AMS can provide much richer information than simply OA, sulfate, and total mass concentrations, and this could be used in future model evaluations to great effect. It is also one reason why the observation dataset is substantially improved relative to Heald et al 2011. I think this merits a comment in the conclusions alongside the comments about the importance of additional observational constraints from new campaigns, since the expense of new observations would be much easier to justify once the existing datasets have been fully exploited.

Technical comments

Figure 6, 10, and S2: the colors are confusing compared to the AMS conventions, please use red for sulfate and green for complex OA.

In Table 3, the standard deviation is often greater than the mean and median, yet negative concentrations aren't possible. This is clearly a matter of opinion and a pretty minor point, but maybe presenting the interquartile range (or better still, the upper and lower quartiles separately) would be more instructive? Or a figure like figure S4, but just to represent variability in observations?

L524 I know it should be obvious but it may be worth saying "biogenic SOA yields for the simple scheme" as presumably this wasn't done for the complex scheme as the dependence is already present.