Response to Anonymous Referee #1 for "An evaluation of global organic aerosol schemes using airborne observations"

We thank the referee for their consideration of our manuscript. Below are our responses (in red) to each of the comments **(in black)**, including the proposed changes to our revised manuscript. Given extensive additions to the SI as a result of review, we do not quote text changes below but provide a track changes version of the main text and the SI.

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 semi-volatile SOA).

This is a good point; we have added a statement in the introduction with an appropriate reference from the 2014 Tsigaridis model intercomparison that can point readers to various ESMs that use similar OA schemes.

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.

We have added a table summarizing the various model simulations to the supplementary information in order to provide greater detail.

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?

We have reworded the statement in Section 3 to be more explicit about the process and have also included more information on the plane-flight diagnostic in the supplementary materials

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?

We have added an appropriate reference to the aerosol scheme used in this study and have reworded the paragraph to limit confusion.

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)?

We have added Africa to the list and have mentioned some relevant campaigns.

"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 'pureVBS' 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?

The isoprene SOA in the complex scheme does not include any non-aqueous contribution and is an explicit scheme that does not include any reversible partitioning. The 'pure VBS' simulation models isoprene SOA using the VBS (no explicit mechanism) as in Pye et al., (2010) and was used in the paper in order to provide context for the complex scheme results. We have updated our model description to explicitly clarify this in order to limit any confusion and have also added a separate category to the model descriptions in the SI.

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)?

While it is likely that the model improvement is largely due to the combination of changes in OA schemes and emissions, it is difficult to attribute model improvement between the different versions given the large number of changes to model code and inputs without running an extensive series of sensitivity experiments with the older code. This is thus regrettably beyond the scope of the work.

We selected a list of representative campaigns to conduct this study and, given the number of simulations required per campaign and our limited computation resources, we decided to focus on campaigns with AMS observations, within the last decade, and that were publicly-available when we started this project in 2017. Text to this effect has been added to Section 3.

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"?

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.

We acknowledge the reviewer's concern and have modified the paragraph to explicitly state that our classification of anthropogenic OA (and indeed all other categories) includes both 'fresh' and 'aged' regions which, particularly in the case of anthropogenic SOA, explains the lower median. The regime analysis is imperfect but is intended to provide broad classification and we have found that adding additional categories over the ones currently in the study can be overwhelming while adding little to the underlying classification that is based on a relative weighting as opposed to an absolute one. With regards to the North Pacific, we track much of that pollution to east Asian emissions from China and the surrounding countries, while the North Atlantic is influenced by both European and African emissions.

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 0.2mugsm3?

As the reviewer points out, the reason there appear to be a number of remote points over the eastern US is due to the density of observations, resulting in observations in the upper troposphere that are below 0.2 ug/m3 (and thus, 'remote') being plotted over points lower in the troposphere. We have added a few lines to the paragraph in order to explicitly clarify this. The regime analysis has been conducted exclusively with OA concentrations and is thus not influenced by other aerosol types. The reason portions of the North Atlantic are classified as anthropogenic is because they do not meet the threshold for remote concentrations (0.2 μg sm^{-3}) and are composed of a minimum of 70% anthropogenic OA (from various continental sources)

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.

In an attempt to address the reviewer's comment, we experimented with source regimes using the complex scheme, however, given the complexities involved in partitioning the POA it is difficult to separate the contribution from anthropogenic and fire sources without large structural changes to the code along with the addition of a number of new tracers (and repeating a suite of simulations).

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?

We agree that it is surprising that the simple and complex scheme broadly track each other given the large differences in OA contributions from the different sources. We have added a short discussion on this topic. We have also included an additional section on why we chose R^2 and NMB as representative metrics in the main text and have included the equations for how we calculate both metrics. The simple SOA is non-volatile and is thus less sensitive to the altitude profile than the complex SOA. This results in a larger complex SOA loading at higher altitudes and leads to a larger NMB.

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?

The overestimated ATom points are classified as Anthropogenic, not remote. This is why the overestimate is not present in the remote comparison.

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.

We wholeheartedly agree with the reviewer! Unfortunately, such tracers, PMF factors, and P-ToF size distributions are typically not provided in publically-accessible datasets. We have added a statement in the conclusions addressing this point and stating that the standardized reporting of chemical signatures from AMS data could enable further model evaluation.

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.

We apologize for the confusion; given our use of two OA schemes, we could not use standard "AMS green" for organics, so decided to differentiate these throughout with red and blue. As a result, we cannot use red for sulfate. However, we agree that using green could be confusing and have changed the sulfate plots to purple so as to be clearly differentiated from the AMS color scheme.

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?

In order to prevent a skew in our analysis, we have not excluded negative observations from our dataset. We chose to use the mean and median as metrics to allow for ease of interpretation but have added a figure to the SI to represent the variability in the 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.

We have made the change to clarify the statement.