"Regional and global climate response to anthropogenic SO2 emissions from China in three climate models" by M. Kasoar et al.

Author response to anonymous referee #3

The authors wish to express their sincere gratitude to the anonymous referee for their invaluable comments and positive appraisal of our study. They have provided thorough and thought-provoking points and we very much appreciate the time taken to do so.

Below we detail our responses to each major and minor comment in turn. We hope that these responses will satisfactorily address all the points raised. The referee's original comment is included in italics, with our response and change to the manuscript in normal font.

Comment 1:

"*While the perturbation applied is well specified, the model output and diagnostics retrieved seems to vary a lot. I realize it's hard to do anything about this once the simulations are done, but for later studies I would encourage the authors to use a wider output protocol. E.g. clear-sky vs all-sky fluxes should be possible to diagnose for all these climate models, and for sulphate perturbations their difference can be very instructive due to differences in treatment of the indirect effect.*"

We absolutely agree. This shortcoming was mentioned also by the first reviewer, and the lessons from this study are indeed being learnt in the discussion of potential future collaborations, which have developed following presentations of the results in this study. As described in the responses to Referee #1, we have in fact also taken the step of extending the simulations with CESM1 for a short period to diagnose the previously missing clear-sky SW flux (which we expect has lower variability than temperature, and so probably doesn't need the same 150-year averaging period), and so the discussion in Sections 4.2 and 4.3 has been updated with this new data.

Comment 2:

"*Page 8, line 23++: For a sudy such as this one, a good diagnostic of TOA RF is very useful. It can be extracted from relatively short and inexpensive fSST runs, as was done here for HadGEM3. I would encourage the authors to add this also for the two other models, and to take the results into their intercomparison discussions.*"

We do agree with the reviewer that including additional simulations would be helpful to get a more precise measure of the radiative responses. However, we have opted already to use the available time to extend the coupled simulations with CESM in order to diagnose clear-sky fluxes as requested by the first reviewer, which we decided was a more critical deficiency. Although more thorough RF diagnostics would be nice for consistency, we do not anticipate they would qualitatively change any of our findings, and we strongly believe that our analysis with the presently available diagnostics already robustly supports the points we make in the conclusions. Given the number of single model studies that have appeared recently in the literature and that have not always considered structural uncertainties, we believe these conclusions are already of sufficient importance and urgency to merit publishing this paper now, rather than incur the further delay and additional costs of additional simulations.

Comment 3:

"*Page 11, line 30-31: The GISS-E2 model has had some problems with its nitrate implementation, and e.g. pulled these results from AeroCom Phase II. Is this issue resolved for the simulations presented here? (I assume so, but still ask since nitrate here seems to be one of the drivers of intermodel differences.)*"

The GISS-E2 configuration used here is the AR5 version, meaning that it does still suffer from the issue of too high a nitrate burden, and probably an overly strong nitrate response as a result. This was, in fact, one of the first things we considered as a possible cause of the discrepancy between GISS and the other models. However, as we discuss in the paper, although there is some partial compensation by increases in nitrate, it turns out to still be a fairly minor factor in the inter-model differences in this study.

Comment 4:

"*Page 13, line 1-10: This section is very interesting, but briefly presented. I would suggest expanding it somewhat, perhaps adding some comparison plots? This would make the study even more useful for future model work.*"

P13, L1-10 discusses the comparison against AERONET, for which there is already a comparison plot in the supplement and we are not sure that there is much scope to expand on it. However, we think the reviewer may have meant Page 14, where we discuss the fractional change in AOD, which turns out to be much larger in HadGEM3 than in GISS-E2 or CESM1. In this case then yes, we do agree that this was rather interesting and could merit some more detail. We have therefore expanded the discussion and added two extra Supplementary Figures here showing firstly how the sulfate fraction of total AOD varies considerably between HadGEM3-GA4 and GISS-E2, and then also comparing the non-sulfate AOD to show that this is in fact similar in these two models, and so the discrepancy in the fraction of total AOD removed is primarily due to disagreeing on the sulfate optical depth only.

Changes made:

1) At the end of the third paragraph of Section 4.1.1, added:

   "This is illustrated further for the two extreme cases, HadGEM3-GA4 and GISS-E2, in Supplementary Fig. S3, which shows that the fraction of climatological AOD made up by sulfate is consistently higher across the east Asian region in HadGEM3-GA4 than in GISS-E2. However, the total non-sulfate AOD is fairly similar across the region in these two models (Supplementary Fig. S4), indicating that the stark difference in the fractional contribution of sulfate comes primarily from HadGEM3-GA4 simulating much greater sulfate AOD alone. Given that regionally GISS-E2 appeared to underestimate total AOD, this would then suggest that either the higher sulfate AOD in HadGEM3-GA4 is more realistic, or else both models underestimate the non-sulfate AOD."

2) Added new Supplementary Figure (S3), showing fraction of total AOD made up by sulfate in GISS-E2 and HadGEM3-GA4.

3) Added new Supplementary Figure (S4) showing total non-sulfate AOD (i.e. total AOD minus sulfate AOD) in GISS-E2 and HadGEM3-GA4.

4) Renumbered other Supplementary Figures accordingly.

Comment 5:

"*Page 15, line 12-30: This section discusses wet deposition results vs observations, and link good performance to a realistic SO4 distribution. However, isn't this also very dependent on the representation of precipitation? The China/Asia region has a lot of variability both in actual and modeled precipitation, and until it's shown that these compare to a reasonable degree I would be cautious about the above interpretation of wet deposition.*"

A valid point. We have added a caveat to this part of our discussion by noting that precipitation will influence the amount of local wet deposition, and so it is difficult to draw definite conclusions from this comparison. (Although, because wet deposition is the primary sink of sulfate aerosol, to some extent regionally it must balance the source of aerosol regardless of the precipitation, and so a large underestimate in the amount of wet deposition could be indicative of too low production of sulfate aerosol). At any rate, we do not rely on this single measure to determine which model is more accurate, but note that it appears consistent with the other observations that we compare with in suggesting that GISS-E2 likely simulates too little sulfate in the region.

Changes made:

1) Added sentence to end of wet deposition paragraph:

   " This overall picture seems consistent with that of the other observational measures looked at here, although it should be noted that wet deposition rates are dependent not just on the column sulfate burden but also on the amount and distribution of precipitation however, and so biases in wet deposition could also be due to incorrect precipitation distribution rather than sulfate."

Comment 6:

"*Page 17, line 17-19: It's hard to assess if e.g. "a 3-fold larger clear-sky SW change" is significant without some indication of the internal variability. Since the results in this paper are mostly from 150-year integrations, I would encourage the authors to add more information on the year-to-year variability (i.e. just the standard deviation of the result across the integration, or similar) throughout the manuscript.*"

We agree with the reviewer that some desirable detail on the significance of the results was either omitted or hard to find, which we have tried to rectify. In our SW and surface temperature plots for GISS-E2 and HadGEM3 we did already include a measure of significance by stippling the plots and stated in the text which temperature responses were significant, but we have now extended that by including ± 2σ uncertainty values in the Table of global and regional responses for all variables that there were sufficient data to calculate it for. This includes the clear-sky SW changes in HadGEM3 and GISS, for which the discrepancy is seen to be extremely significant (around 23 standard deviations). Extending this to all variables in the Table is complicated by the fact that the very long control simulation used to assess variability in GISS-E2 only output basic climate diagnostics and not more detailed aerosol-related diagnostics, and we have no equivalent long or ensemble control simulation at all for CESM. The SW changes and final temperature response are ultimately what we are interested in most though, so we do not think that this is too restrictive (and one can generally use the value given for HadGEM3 to get at least an order-of-magnitude estimate of the likely uncertainty where a value isn't available for the other models). We deliberately avoided estimating the significance of other variables from the year-to-year variability in these simulations though, because we do not think this necessarily leads to an accurate measure of the long-term 150-year variability which is the relevant quantity here, and on which we base our uncertainty analysis.

Changes made:

1) Added ± 2σ uncertainty values to the Ch0-Con differences in Table 2 (formerly Table 1), for all variables for which long/multiple control runs data were available (all of HadGEM3 + temperature and radiative fluxes for GISS). Added statement to Table 2 caption:

   "For models and variables where data was available, error ranges are quoted for the Ch0-Con values and indicate ± 2 standard deviations, evaluated in HadGEM3-GA4 from an ensemble of six 150-year control runs with perturbed initial conditions, and in GISS-E2 from twelve 150-year segments of a long pre-industrial control run. Values quoted without error ranges indicate that uncertainty was not evaluated."

Comment 7:

"*Table 1: The numbers listed here seem to have an unrealistically high precision (e.g. -0.034810…) Please give a reasonable number of significant digits, and also include some indication of the internal variability in each model (see previous comment).*"

We agree that the precision that the numbers are quoted to is implausibly high – this is an oversight that appears to have crept in from an old version of the table, and the numbers should have been truncated to fewer significant figures in the submitted version. This has now been corrected. See our response to the previous comment for discussion of internal variability and estimating significance – we have added error values into Table 1 for the variables and models for which these we had these figures.

Changes made:

1) Values in Table 2 truncated so that Ch0-Con values are at most 3 significant figures. Values for individual simulations have been truncated to at most the same number of decimal places as the Ch0-Con anomalies for that variable.

2) Added significance estimates to Table 2 as detailed in response to Comment 6

Minor comment 1:

"*Abstract (p2): "...and reinforces that caution must be applied when interpreting the results of single-model studies." I believe the results of this paper show that we should be cautions also in interpreting multi-model studies. They are usually just ensembles of opportunity, with little or no observational* *constraint beyond what is already taken into the model parametrizations. Hence their average values are not necessarily closer to reality, but instead just indicative of the present model diversity.*"

We have modified both the abstract and conclusion so as to not limit our statement to single-model studies.

Changes made:

1) In the abstract, changed 'single-model studies' to 'modelling studies'

2) Changed the corresponding line in the second last paragraph in the conclusion ("…and imply that care must be taken not to over-interpret the results of studies performed with single models") to:

   "…and imply that care must be taken not to over-interpret studies of aerosol-climate interaction if the robustness of results across diverse models cannot be demonstrated"

Minor comment 2:

"*Page 3, line 31-32: The Phase II AeroCom study (Myhre et al. 2013, ACP) which you cite later probably belongs in this company.*"

We agree, and have added a reference to this paper in that section as well.

Changes made:

1) Added 'Myhre et al., 2013' to bracketed list of HTAP and AeroCom references.

Minor comment 3:

"*Section 2.1: The description of HadGEM3-GA4 is very long compared to the two other models. Could the descriptions be clarified and made more uniform? Perhaps through a table of the most relevant model parameters/physical processes included?*"

Agreed – this is something that has been mentioned by another reviewer as well, though the other reviewer favoured more detail for CESM1 and GISS-E2 rather than less for HadGEM3-GA4. We have slightly cut down superfluous details in the HadGEM3 description while adding several additional details to the other model descriptions and slightly re-ordering them to make the descriptions more uniformly structured. As recommended, we have also added a new table which includes key references and features of the three models for easy reference.

Changes made:

1) Numerous changes to model descriptions which are detailed in responses to Referee #1 Minor Comment 1 and Referee #2 Minor Comments 2, 3, 4, and 5, which harmonise the model descriptions.

2) In the first paragraph of Section 2, added:

   "The models are briefly described below, and the key references and features are also summarised in Table 1."

3) Added a new table (Table 1; previous Table 1 is now Table 2) with key model details. Updated all previous instances of "Table 1" in the text to "Table 2", and updated the caption of the existing table to Table 2. Added caption to new table:

   "Table 1: Key references and features of the three models and their aerosol schemes used in this study"

Minor comment 4:

"*Page 18, line 1-2: The SO4 forcing is not very sensitive to the vertical distribution, compared e.g. to absorbing species. See e.g. Samset and Myhre, GRL 2011, doi:10.1029/2011GL049697.*"

This is very true. We have removed that speculation, and found a different (partial) explanation:

Changes made:

1) Removed "For instance, the forcing per unit AOD will be influenced by the vertical distribution of the aerosol (Myhre et al., 2013a), which could vary between models in different parts of the world."

2) Replaced with:

   "The sulfate efficiencies in Myhre et al. (2013) are calculated relative to all-sky direct radiative effect, and so local differences in vertical profiles and cloud screening may therefore change the relationship – however they also evaluated clear-sky forcing normalised by AOD for all aerosol species combined, and again found HadGEM2 to be higher than GISS ModelE."

3) Additionally, at end of this section, added text indicated in the mark-up below:

   "However, the study also found that, globally, the atmospheric component of HadGEM2 had a slightly larger~~very similar~~ forcing efficiency to CAM5.1 both for sulfate (all-sky) and all aerosols (clear-sky), unlike the somewhat smaller regional efficiency found here for HadGEM3-GA4 compared with CESM1. Given that our regional values from GISS-E2 and HadGEM3-GA4 also seem to conflict qualitatively with the global values from the AeroCom study, this would suggest that either the global comparison is not relevant on regional scales, or else the radiative efficiency is very sensitive to changes in model configuration and version.