*Reviewer comments are in **black**, author replies are in blue, and revised text is in red.*

On behalf of all authors, I would like to thank Referee #2 for their carefully thought out comments and suggestions, which I believe have substantially improved the paper. Responses to each point follow below.

Laura Revell, University of Canterbury, 12 September 2018.

Interactive comment on "Tropospheric ozone in CCMI models and Gaussian emulation to understand biases in the SOCOLv3 chemistry-climate model" by Laura E. Revell et al.

Anonymous Referee #2

Received and published: 25 August 2018

This manuscript quantifies tropospheric ozone biases in two versions of the SOCOL chemistry-climate model, as well as the CCMI models. The SOCOL bias is further investigated using an emulator. I find the methodology novel, and the Discussions and Conclusions is particularly well reasoned and should be of considerable interest to the chemistry-climate modeling community. I do believe the paper could be greatly improved if some choices and details of the methodology are better explained (and perhaps if the paper is slightly restructured) as I explain in my two major criticisms below.

General comments

1) A stronger rationalization of the input parameter choices for the emulator is needed in Section 2.4. An important reason for testing the ozone precursors [variables (1- 3)] is that they are a primary candidate for the cause of the systematic high bias in tropospheric ozone among model intercomparisons that use harmonized emissions, such as CCMI and ACCMIP. An important reason for testing (3) should be that SOCOL is very simplistic in its representation of NMVOC chemistry compared to other CCMI models (as an aside: why not vary the yield of CO from NMVOC oxidation separately to the magnitude of NMVOC emissions?). It also seems that variables (4-9) are chosen to reflect developments between SOCOLv3.0 and v3.1...is that correct? If so, I am not sure why, besides (8), they are investigated at all since the authors have already performed a sensitivity test in which they find that inclusion of heterogeneous hydrolysis of N2O5 is the main development that reduces the model's ozone bias between the two versions (P9L31).

This section has been rewritten, taking the reviewers' feedback into account. To briefly answer the questions above:

a) SOCOL's NMVOC chemistry scheme is indeed very simplistic, and CO emissions far exceed NMVOC emissions (isoprene and formaldehyde). As described in the methodology, CO is prescribed as an "additional" NMVOC in SOCOL to account for missing NMVOCs. It is for this reason that we decided to treat CO and NMVOCs together. However, as we have now noted in the manuscript (see below), we do not recommend such an approach for CCMs with more complex NMVOC schemes.

b) Yes, variables 4-9 were chosen to reflect developments between SOCOLv3.0 and 3.1 – hopefully this is now clear in the revised text (below). Even though we performed some individual sensitivity tests initially, the advantage of including them in the sensitivity analysis is that joint interactions between these variables can be identified.

Revised text in Section 2.4 is as follows:

"Although many factors influence the tropospheric ozone budget, we restricted our analysis to 9 model forcings/parametrizations (see Table 1 for details of the scalings applied). These are listed below, followed by a section rationalizing the inclusion of each variable. We reiterate that this list above does not constitute a comprehensive list of variables controlling tropospheric ozone, however by illustrating the methodology used, we aim to demonstrate its utility."

…[List follows here]…

"Variables (1-3) were selected due to their importance as tropospheric ozone precursors. CO and NMVOC emissions were varied simultaneously (3) because the only NMVOCs included explicitly in SOCOL are isoprene and formaldehyde; other NMVOCs are represented via additional CO using a 'lumped' approach (Section 2.2). For models with a more complex representation of NMVOCs, we recommend testing CO and NMVOC emissions separately when constructing a GP emulator.

The remaining variables were included to investigate the sensitivity of tropospheric ozone to the model improvements implemented in SOCOLv3.1. SOCOLv3.0 and its predecessors prescribed methane on the lowermost six model levels. This was changed to only the surface level in SOCOLv3.1, and variable (5) was included in our analysis to investigate the sensitivity of tropospheric ozone to this implementation. By doing so, we aim to test the exchange of emissions between the boundary layer and free troposphere. The lowermost level in SOCOL covers approximately 100 m, and the 6 lowermost levels combined cover approximately 2.5 km. To explore whether other ozone precursors are sensitive to the number of levels they are prescribed on, variable (4) was included, even though it is prescribed only as a surface emissions flux in most, if not all, CCMs.

Because ozone production and destruction reactions are mostly photochemical, i.e. they occur in the presence of sunlight, we selected variable (6) to test the sensitivity of the current CMF parametrization, and examine impacts of the updated LUTs on tropospheric ozone in SOCOLv3.1. HNO_3 washout is the main sink for NO_x, and therefore affects the ozone budget. Future SOCOL versions will include an online wet deposition scheme, and so variable (7) was

selected to probe the sensitivity of tropospheric ozone to the rate of HNO_3 loss. Heterogeneous N_2O_5 hydrolysis is similarly important as it leads to HNO_3 formation, however it was not included in SOCOLv3.0. Therefore variable (8) was included in our analysis to quantify its relevance for tropospheric ozone abundances. Finally, variable (9) was chosen to test the sensitivity of tropospheric ozone to the newly-implemented dry deposition parametrization (Section 2.3)."

2) It seems that the most detailed portion of the paper is focused on quantifying and understanding SOCOL's ozone biases, in part with the emulator, rather than an exploration of biases in the CCMI models (which could be a paper by itself!). With this in mind, the authors might consider first discussing SOCOL biases and then placing the results of the single model study within the wider context of the CCMI models e.g. combining Section 3.1 with the first paragraph of the Discussions and Conclusions. However, I leave this up to the authors.

We have taken this suggestion on board, and shuffled material around; the methods subsection "CCM simulations to compare with observations" has been moved to the start of the methods section, and the results subsection "Tropospheric ozone in the CCMI models" has been moved to the end of the results section. This allows a more-or-less seamless transition from: a) describing the GP emulator methodology to showing the emulator results; and b) showing the CCMI comparison to discussing the results in the Discussion and conclusions.

Secondly, and more importantly, please elaborate upon the basics of the emulation technique. Although I appreciate that the authors are probably trying to avoid jargon, as a non-statistician, I find the beginning of Section 2.4 a little confusing.

Here we refer also to Referee 1's comments and our response to those. Referee 1 has previous experience with Gaussian Process emulation, and provided many constructive comments aimed at improving the description of this technique. We hope that the revised manuscript is now clearer to read for statisticians and non-statisticians alike.

Finally, the emulator experiments are a novel contribution to this field, which should be emphasized in the Introduction and Conclusions to increase the significance of the paper. Perhaps the authors could also speak to the broader goals such as extending the emulation methodology to explore tropospheric ozone variability due to meteorological parameters (e.g. convective parameters) not investigated here, or variability in other metrics such as ozone extremes etc...

We have emphasized the novel contribution of this study in the introduction and conclusions as suggested. E.g., from the Introduction:

"This is the first time the technique has been applied to global tropospheric ozone. Our GP emulator experiments have been designed to focus on recent developments regarding

SOCOL's tropospheric chemistry scheme, however the methodology has the potential to be expanded to also include meteorological parameters."

And the end of the Discussion and conclusions section:

"Given the results of our multi-model intercomparison as well as previous multi-model studies, our results highlight the need for careful validation of emissions inventories used by global models. However, the way in which emissions are handled by the models also appears to result in biased ozone abundances, and further work is needed to address the challenges of simulating sub-grid processes of importance to tropospheric ozone, in SOCOLv3 as well as in other CCMs. GP emulation may prove a useful tool for such studies, and we have demonstrated its usefulness for understanding tropospheric ozone biases. GP emulation is a powerful tool, and should be considered for use by those wanting to perform detailed sensitivity analyses at low computational cost."

Specific comments

3) P2L21: these fractions were deduced using data over individual sites in the Southern Hemisphere and are not necessarily representative of the whole troposphere.

Noted: "Greenslade et al. 2017 calculate the mean fraction of total tropospheric ozone attributable to STE at three sites between 38—69° S as 1-3%, and show that during individual STE events, over 10% of tropospheric ozone may be directly transported from the stratosphere."

4) P2L23: specify that this is the "global tropospheric lifetime" since the ozone lifetime can vary considerably by region.

Changed as suggested.

5) P2L27: please cite Young et al. (2018) alongside Young et al. (2013) and Parrish et al. (2014).

Done.

6) P3L5: please cite Stevenson et al. (2006) for ACCENT and Young et al. (2013) for ACCMIP.

Done.

7) P3L26 (and P6L21): Do you mean non-additive instead of non-linear?

Indeed, it turns out that non-linear is not the correct term – the other reviewer advised referring to them as "interacting" contributions, which we have now done.

8) P4L3: For clarity, specify that SOCOL is a chemistry-climate model.

Done.

9) P4: Provide some information about the stratospheric boundary conditions.

This information has been added to the section "CCM simulations to compare with observations." (Added text is in bold):

"Greenhouse gas concentrations (CH_4, CO_2 and N_2O) follow observations until 2005, then Representative Concentration Pathway (RCP) 8.5 (Riahi et al., 2011). Ozone precursor emissions (including NO_x, CO and NMVOCs) follow a historical emissions inventory until 2000 (Lamarque et al., 2010), then RCP 6.0 (Masui et al., 2011). Sea surface temperatures and sea ice concentrations were prescribed following HadISST observations (Rayner et al., 2003). **Concentrations of ozone-depleting substances followed the World Meteorological Organization's A1 scenario (WMO2011), and stratospheric aerosol surface area densities and optical parameters were prescribed from the SAGE-4λ data set (Arfeuille et al. 2013, Luo 2013)."**

10) P4L16: A look-up table is an offline, not online, photolysis scheme (in agreement with the last sentence of the paragraph).

This has been corrected.

11) P5L14: This is inconsistent with P4L29, which states that methane is prescribed as a "surface mixing ratio", which implies the lowermost model level.

That sentence has now been removed from P4L29, and the discussion about how methane is prescribed is left until the section "Upgraded model version SOCOLv3.1".

12) P5L16: Naively, I would not expect methane-induced ozone production to be reduced upon prescribing methane on one level versus multiple levels since it is well mixed in the troposphere.

We in the SOCOL group were also initially surprised at the result, however the reduction in tropospheric ozone is not huge (10% at maximum). Our reasoning for the result is outlined in the next few lines of the ACPD manuscript.

13) P6 paragraph 1 and Section 3.1: I wonder how much of the inter-model differences in the tropospheric ozone burden arise from inter-model differences in tropopause height. Could this be quantified by imposing the same tropopause height across all the models and noting the difference in ozone burden?

Shown below is annual-mean tropospheric column ozone in 2005, where tropospheric ozone columns were calculated between the surface and 250 hPa, rather than the WMO-defined tropopause. As would be expected by imposing the tropopause at 250 hPa, the global-mean tropospheric ozone abundance is smaller compared with Figure 2 from our ACPD manuscript. It can also be seen that the same differences in terms of the spatial distribution and tropospheric ozone abundances exist, regardless of where the tropopause is defined. As noted in the manuscript, we opted to select the WMO-defined tropopause to enable a "like-

with-like" comparison with the OMI/MLS satellite product. Therefore the figure shown in the manuscript remains unchanged.
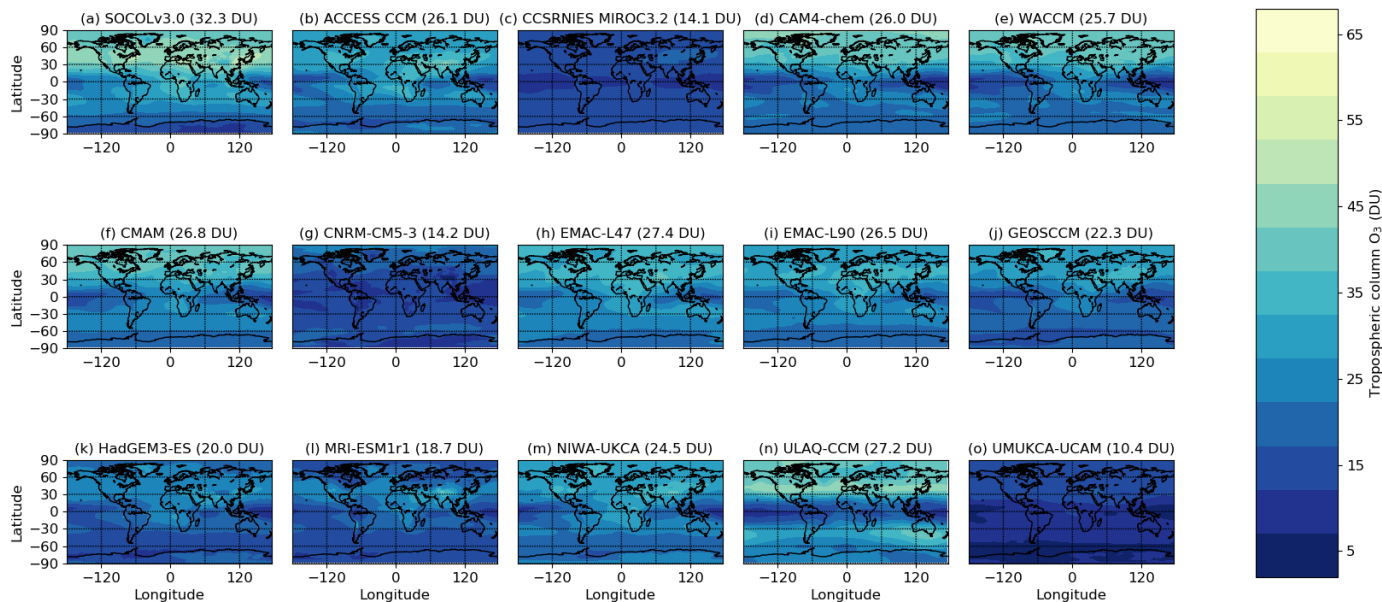
14) P6L20: Please see General Comment #2. This sentence is packed with information and is confusing to a non-statistician.

This section now reads:

"Variance-based global sensitivity analysis allows the individual contribution of a single parameter to the overall uncertainty to be quantified. Because the large number of model simulations required would make one-at-a-time testing computationally too expensive, a type of statistical model called a GP emulator can be used as a surrogate for the input-output relation of a complex model (Le Gratiet et al., 2017), such as a CCM. For "training" data on which the GP emulator is built, we know that the true value of the emulated output should be the same as the input, so the emulator should return the output with no uncertainty. For inputs that the emulator is not trained at, the outputs should have a probability distribution specified by a mean function and covariance function (O'Hagan, 2006).

Here, we use tropospheric ozone columns from SOCOLv3.1 to train the emulator. Interacting contributions to the overall uncertainty in tropospheric column ozone can be identified by comparing the main effect variance (the reduction in the ozone variance when a particular model forcing is fixed, e.g. NO_x emissions), with the total effect variance (the remaining variance in the tropospheric column ozone when everything except a particular model forcing is fixed). Various software packages are available for GP emulation. We used the Gaussian Emulation Machine for Sensitivity Analysis (GEM-SA), available at http://tonyohagan.co.uk/academic/GEM/index.html, to build an emulator for tropospheric column ozone."

15) P6 points 1 and 3: Which type of emissions? Anthropogenic/biomass burning/natural?

We have now noted these in the manuscript – NO_x: natural and anthropogenic. CO: natural and anthropogenic. NMVOCs: anthropogenic, biomass burning and biogenic.

16) P6 point 4: I am unclear as to why this is tested. Emissions are included as surface fluxes (i.e. lowest model level) in both SOCOL versions, and to my knowledge, across most models.

This variable was included following the realization that tropospheric ozone in SOCOL is slightly sensitive to the number of levels methane is prescribed on. We were curious as to whether ozone would be similarly sensitive to the number of levels NO_x, CO and NMVOCs are prescribed on. The following text has been added to clarify this:

"…variable (5) was included in our analysis to investigate the sensitivity of tropospheric ozone to this implementation. By doing so, we aim to test the exchange of emissions between the boundary layer and free troposphere. The lowermost level in SOCOL covers approximately 100 m, and the 6 lowermost levels combined cover approximately 2.5 km. To explore whether other ozone precursors are sensitive to the number of levels they are prescribed on, variable (4) was included, even though it is prescribed only as a surface emissions flux in most, if not all, CCMs."

17) P7 point 5: I would have thought a priori that the number of levels that methane is prescribed on would not matter for tropospheric ozone amounts, and this is confirmed later in the paper.

Yes – also addressed in point (12) above.

18) P7L24: I am not sure why you would test ranges that are not feasible. E.g. the maximum range for methane (4xCH4) is much larger than even RCP8.5 year 2100 amounts relative to present day. Are we then sure the results of the emulator remain meaningful?

The importance of selecting an appropriate sampling distribution is addressed on P10L17-33 of the ACPD manuscript, and was motivated by observing the "NO_x saturation effect" at scaling factors greater than one. Given the overwhelming dominance of ozone precursors as drivers of tropospheric ozone variability ($\gtrsim$90% in all regions examined), we are confident that, were the analysis to be repeated with a constricted range of scaling factors, the overall results would remain unchanged.

19) P7: The final paragraph explains that physical/meteorological parameters are, by design, not investigated in the emulator experiments. Indeed there could be multiple reasons, besides chemistry, for SOCOL's particularly high ozone bias. This is explained well in the Discussion, but should also be made clear in the Introduction: the methodology used here does not explain (nor is it intended to explain) the entirety of the "remaining ozone bias in SOCOLv3.1" as stated on P3L20.

The following text has been added to the Introduction:

20) P8L2 and Section 3.2: Why not also show results for the global mean tropospheric ozone burden, given its discussion in the Abstract and elsewhere.

We have included the global-mean results in our analysis, and expanded Figures 6 and 8 (now Figures 3 and 5, since the emulator results have been moved to before the CCMI results, following the reviewer's suggestion above) to show the global-mean:

Revised Figure 6 (now Figure 3):
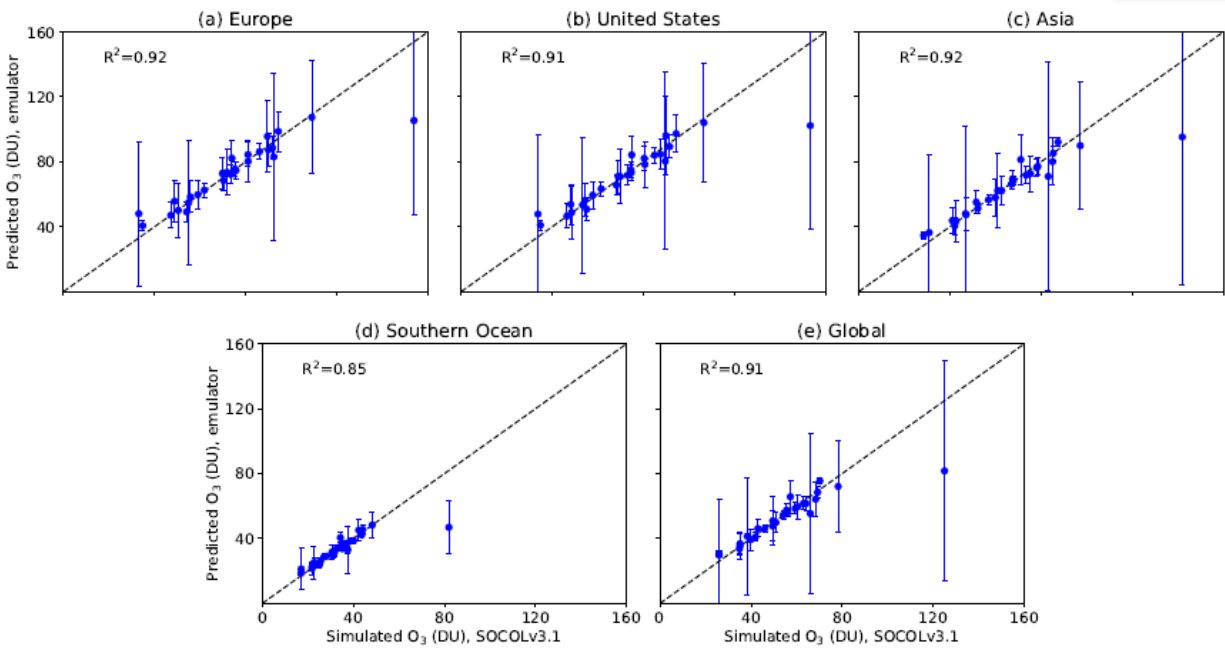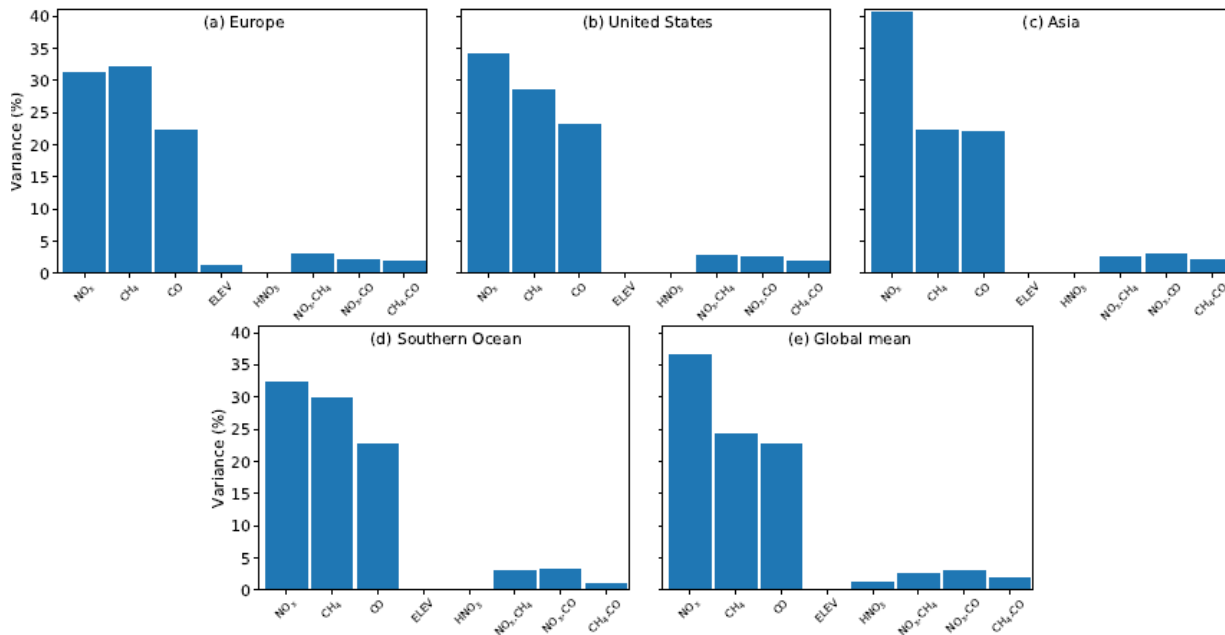
Revised Figure 8 (now Figure 5):

Figure 7 also shows global-mean results, with the corresponding plots for Europe, the US, Asia and Southern Ocean moved to the supplement.

21) P8L12: Reference Morgenstern et al. (2017) who discuss familial relationships between the CCMI models.

Done.

22) P8L22: I do not think you can say ECAM-L90 simulates a "better" representation here since there is no comparison to the observations yet.

This sentence has been relocated to after the following paragraph, once the comparison with observations has been introduced.

23) P9L16: Please provide the ACCMIP MMM global mean tropospheric ozone burden in DU for comparison with CCMI and CMIP5. Also state which, or at least how many, models were considered in the ACCMIP and CMIP mean.

ACCMIP: 30.8 DU calculated from 15 models. CMIP: 30.5 DU calculated from 18 models, as now noted in the text:

"The ACCMIP models simulated, on average, up to 30% more tropospheric column ozone compared with OMI/MLS at northern midlatitudes (Young et al., 2013). The global- annual-

mean tropospheric ozone column simulated by these models was 30.8 DU, calculated from 15 models. For the 18 CHEM models participating in CMIP5 (those models with interactive chemistry, i.e. ozone was calculated online and not prescribed from a climatology), the climatological-mean annual-mean MMM averaged over 2000-2005 was 30.5 DU (Eyring et al., 2013), which is similar to the MMMs calculated here. The CMIP5 and ACCMIP MMMs also show a stronger interhemispheric gradient than OMI/MLS observations do, consistent with our findings."

24) P9: The CCMI/ACCMIP/CMIP5 comparison is brief. This is fine for the present study, but perhaps the authors could highlight the potential for more detailed future investigation (see also General Comment #2). It would be interesting to see the extent of agreement - or lack thereof - between the different model intercomparisons' simulation of tropospheric ozone, given their different aims and formulations (e.g. a focus on stratosphere-troposphere interactions in the CCMI models vs atmosphere-ocean coupling in CMIP5).

We have included comments on this in the Discussions and conclusions:

"Although ACCMIP, CMIP5 and CCMI all used the same emissions inventories, it is nevertheless interesting that they all produced very similar global-mean 10 tropospheric ozone abundances (approximately 30 DU), given the different foci of the different model intercomparison activities; CCMI focussed on models coupling the stratosphere and troposphere, while CMIP5 focussed on coupling the atmosphere and ocean."

25) Figures 2 and parts of Figure 4, 5: The continuous scale in these figures makes it difficult to distinguish numerical differences between the sub-plots. I recommend a discrete scale as in Figures 3 and 4c, 4f, 5c, 5f.

Changed as suggested.

26) P9L30: Do you mean regionally not globally?

Yes – corrected in the text.

27) P9L33: From Figure 3, it looks like several of the CCMI models also show this bias over the Southern Ocean. Do they share the Wesely deposition scheme?

No, from Morgenstern et al. (2017) they use a variety of schemes – some online, some offline.

28) P10L6: State where this maximum bias occurs.

Done – continental regions in the Northern Hemisphere and Southeast Asia.

29) P10L9, Figure 6: Am I right in thinking that two conditions need to be satisfied in order for the emulator to perform well: having a high R squared value and having the points falling on a 1:1 line? Please clarify.

Yes, and this has now been clarified in the text.

30) P10L10: See earlier comment about using inputs outside feasible ranges, which is acknowledged on P10L30. Do these extremes need to be tested?

We did perform some testing on the extremes, described on P10L17-33 of the ACPD manuscript, and discussed above (point 18).

31) P10L20: Can we explain this? Does it reflect a NOx titration effect?

That is our thinking, yes, and we have added some text to clarify this in the revised manuscript.

32) P10L17, Figure 7: I am a little confused on what to take from this figure: is the "sensitivity" of tropospheric ozone to each parameter determined by the slopes of the sub-plots? If so, why compare the different sensitivities? To determine which parameters are more "important" for tropospheric ozone variability, it makes more sense to compare the variance explained by each parameter (Figure 8). Finally, what does the uncertainty in Figure 7 signify? I may be missing the obvious! Please explain Figure 7 clearly or consider removing.

Figure 7 is useful because it shows whether ozone increases or decreases in response to an individual forcing – this information can't be obtained from Figure 8. Also yes, the slopes can be used to get an indication of how sensitive tropospheric ozone is to a particular forcing. This section has been rewritten (noting that Figure 7 is called Figure 4 in the revised manuscript):

"Figure 4 displays the sensitivity of global-mean tropospheric ozone to each parameter, obtained by averaging over all other parameters, and indicates whether tropospheric ozone increases or decreases in response to an individual forcing/parametrization. Greater uncertainty is indicated where the lines diverge (appearing as a thicker line – i.e., the emulator is less well constrained). Tropospheric ozone exhibits a strong sensitivity to its precursor gases (Fig. 4a-c), and while the correlation between CH_4 and CO+NMVOCs is approximately linear, for NO_x there appears to be a saturation effect for scaling factors greater than one, likely due to the "NO_x titration effect" (Thornton et al., 2002)."

33) P10L17: "Figure 7 displays the sensitivity of global-mean tropospheric ozone..." but the figure caption suggests the mean is over the Asian region only.

It should have read that it was for the Asian region in the text. We have now replaced Figure 7 with a plot for the global-mean. Individual plots for Asia, Europe, the US and Southern Ocean are shown in the Supplement.

34) Figure 8: Remove "9 variables" from the figure caption since all 9 variables are not shown.

Done.

35) Figure 8: Could you also show a panel for the global mean burden?

Yes, and now done (shown above, point (20)).

36) Figure 8: Could you explain why the relative importance of CH4 and CO is smaller over Asia than Europe or the US? It would be better to use the same scale on all the panels.

In Europe and the US, the ratio of NO_x:CO is high (i.e. there is relatively more NO_x than CO) – see Revell et al. 2015 (www.atmos-chem-phys.net/15/5887/2015/), their Figure 2 and 3d. This would mean that over Asia, where NO_x is relatively less abundant compared with CO (because CO emissions are so large), NO_x would become more important for driving ozone variability. Discussion of this has been added to the text:

"Over Asia, where CO emissions are larger than over Europe and the United States, the ratio of NO_x:CO is also lower than it is over Europe and the United States (Revell et al., 2015). NO_x emissions therefore become more important as a driver of ozone variability over Asia (Fig. 5c)."

37) P11L6: "up to 8 DU regionally"

Changed as suggested.

38) P11L12: "up to ~30 DU regionally"

Changed as suggested.

39) Discussions and Conclusions: I very much like this section! I would only conclude with some remarks on the novelty of the emulation technique within this field and its potential future value in the study of ozone biases (see General Comment #2).

This section now concludes:

"Given the results of our multi-model intercomparison as well as previous multi-model studies, our results highlight the need for careful validation of emissions inventories used by global models. However, the way in which emissions are handled by the models also appears to result in biased ozone abundances, and further work is needed to address the challenges of simulating sub-grid processes of importance to tropospheric ozone, in SOCOLv3 as well as in other CCMs. GP emulation may prove a useful tool for such studies, and we have demonstrated its usefulness for understanding tropospheric ozone biases. GP emulation is a powerful tool, and should be considered for use by those wanting to perform detailed sensitivity analyses at low computational cost."