September 12, 2018.

Dear Editor,

We have substantially revised this manuscript following the recommendations of both reviewers. Below, please find point-by-point responses to the reviewer comments, which describe the modifications made to the manuscript, followed by a marked-up copy of the revised manuscript.

Best regards,

Laura Revell on behalf of all co-authors.

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 Reviewer Edmund Ryan for his very helpful comments, especially regarding the statistical methods used in this paper. I hope that the description of these methods in the revised manuscript reads more clearly for both statisticians and non-statisticians.

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.

E. Ryan (Referee)

edmund.ryan@lancaster.ac.uk

Received and published: 1 August 2018

General comments

This is a nice paper. As a statistician, I focused mainly on the emulator and statistical part of the manuscript. So my first comment is that it's great to see emulators appearing in atmospheric chemistry modelling research for the purposes of doing statistical analyses such as global sensitivity analysis which would be too computationally burdensome without emulators. These papers are still fairly rare, so you're encouraging others in the atmospheric chemistry modelling community to consider these methods (which is awesome!). GEM-SA is a great tool developed by statisticians at the University of Sheffield, to make it easier for applied scientists to carry out this type of statistical analysis with minimal understanding of the statistics. The implementation of GEM-SA appears to be done correctly and so I'm satisfied that the results are all fine. However there are a large number of issues that need addressing, So although there is nothing major that needs changing, I've indicated major corrections to give you enough time to address these large number of comments, some of which made need a lot of thought. Feel free to e-mail me if you need me to clarify any of these comments.

Major Comments

[1] page 6, lines 20-21. The sentence starting "The output variable . . ." sits uncomfortably with me. While we are technically "fitting", I would use this word here as the non-statistical reader may infer from this that you're using measurements. The phrase "uncertainties are calculated with a covariance function" is also too vague. Finally, you say that "each output point has a normal distribution." This is incorrect. A GP emulator that the guys at Sheffield

developed is built within a Bayesian framework, where prior is a GP, the likelihood function is a multivariate Normal distribution and the resulting derived posterior is a student-t distribution. I suggest you drastically reword this sentence. You can still keep parts of it, but the parts above that I mention need to be changed. I suggest you use these few lines to actually define what a GP emulator is. In Tony O'Hagan's paper he defines it by two properties: (1) an interpolator such that at inputs the emulator is trained at, the emulated outputs must be the same as the simulator outputs; (2) for inputs the emulator is not trained at, the emulated outputs have a probability distribution specified by a mean function and a covariance function. In my paper that recently got accepted (Ryan et al., in review; https://www.geosci-modeldev-discuss.net/gmd-2017-271/), I give a definition like this and other details. You may want to refer to this to help with this part of your methods section.

This section has been rewritten:

"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, such as a CCM (Le Gratiet et al., 2017). 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.

[2] Page 7, line 24. I feel uncomfortable about you using the words "not necessarily feasible" here. For a sensitivity analysis study, justifying he mins and maxs of your inputs is important because if your range covers values of a particular input that are not feasible this could give misleading results in the sensitivity analysis. In other words, suppose the range for an input is (2,4) and you find that the output is not sensitive to the changes in that input. Now suppose you were to repeat the analysis with a range of that input as (1,4) and suppose that the output is now quite sensitive to the changes in that input. Well, this means that the results of the sensitivity analysis are "sensitive" to the value you used for the minimum value. This won't always be the case, but I feel it's important to justify why the choices of mins and maxs of each input are appropriate.

We have removed this sentence, and expanded Table 1 to add further description of the ranges for the sensitivity analysis. The revised Table 1 is shown below. Some of the ranges were chosen based on past experience with SOCOL – for example, previous sensitivity tests have indicated that halving the NO_x emissions leads to close agreement between modelled and observed tropospheric column ozone. Here the range of 0.25 to 4 was selected to cover a larger uncertainty space. For ELEV and CLEV, the maximum of 6 levels (~2.5 km) corresponds to the maximum boundary layer height at mid-latitudes, which is where most

emissions occur; however most (if not all) models prescribe emissions only at the surface, which is the recommended approach.

Table 1. Range of the sensitivity forcings/parametrizations. **P** and **L** indicate whether the variable is of relevance to ozone production and/or loss, respectively.

	Minimum	Maximum	Descriptions
(1) NO_x emissions (**P**)	0	4	The surface NO_x emissions field as a function of latitude and longitude was multiplied by a scaling factor between 0 and 4, to explore the sensitivity of tropospheric ozone to a range of NO_x emissions.
(2) CH_4 concentrations (**P**)	0	4	The global-mean CH_4 mixing ratio was multiplied by a scaling factor between 0 and 4, to explore the sensitivity of tropospheric ozone to a range of CH_4 concentrations.
(3) CO+NMVOC (**P**) emissions	0	4	As for (1), but the scaling factor was applied to CO and NMVOC emissions simultaneously.
(4) ELEV for NO_x and CO+NMVOCs (**P**)	1	6	Emissions were prescribed on the lowermost 1–6 levels (between (the surface and ~2.5 km, to test whether the number of levels is important for tropospheric ozone abundances.
(5) CLEV for CH_4 (**P**)	1	6	CH_4 concentrations were prescribed on the lowermost 1–6 levels (between the surface and ~2.5 km, similar to (4).
(6) CMF (**P+L**)	0.25	1	1 implies clear-sky photolysis, whereas 0 would imply no photolysis. As photolysis rates of 0 do not occur during daytime, we selected a lower bound of 0.25 to represent cloudy sky conditions.
(7) HNO_3 washout (**L**)	0	0.5	To test the sensitivity of tropospheric ozone to HNO_3 removal, we removed between 0–50% of tropospheric gas-phase HNO_3 at each chemical time step.
(8) N_2O_5 hydrolysis (**L**)	0.001	0.3	The probability of N_2O_5 hydrolysis occurring. Since the default is 0.1, we explored the sensitivity of tropospheric ozone to a range from 0.001-0.3.
(9) O_3 dry deposition (**L**)	0	1	A specific reactivity of 0 stands for a nearly non-reactive gas, while 1 stands for a gas similarly reactive to ozone.

[3] Page 8/9 (section 3.1). In your methods, I found only one line where you talk about incorporating other models in this study, but then in your results you have four figures (figs. 2-5) of results before getting onto the results from the sensitivity analysis. I am unsure how section 3.1 and figures 2-5 fit into this analysis. Please can you explain this? Have figures 2-5 been reported elsewhere? I can understand why you may want to include one or two of figures 2-5 in your methods and motivation for doing the sensitivity analysis, but I don't think they should be part of your results. Reading your abstract, it seems that your paper is split into two parts: (1) introducing a new version to the SOCOL model; (2) carrying out the sensitivity analysis. So I could understand if figures 2-5 and section 3.1 were devoted to validating or testing SOCOL v3.1, but including the other CCMI models in your "results" section seems problematic. If you do justify leaving in section 3.1 and figs 2-5 then at the very

least I feel that you need to talk a lot more about these CCMI models in your methods and what research questions you're answering. Looking at the end of your introduction (where research questions are normally stated), the only things I read, that state what the paper will be about, are: (1) some results from SOCOL v3.1 and (2) the sensitivity analysis. Do you see my confusion?

The CCMI aspect of the study is an important one, as this is the first time that global distributions of tropospheric ozone from the CCMI models have been presented and compared with observations. Following Reviewer 2's suggestion, we have shuffled the order of material in the Results and Discussion a little, so that the emulator results are presented before the comparison of the CCMI models.

In the revised manuscript, the CCMI comparison is described in:

- Abstract, lines 2-6:

  "We investigate annual-mean tropospheric column ozone in 15 models participating in the SPARC/IGAC (Stratosphere-troposphere Processes and their Role in Climate/International Global Atmospheric Chemistry) Chemistry-Climate Model Initiative (CCMI). These models exhibit a positive bias, on average, of up to 40–50% in the Northern Hemisphere compared with observations derived from the Ozone Monitoring Instrument and Microwave Limb Sounder (OMI/MLS), and a negative bias of up to ~30% in the Southern Hemisphere."
- Introduction, P3L30-P4L2:

  "SOCOLv3.0 participated in phase 1 of the Chemistry-Climate Model Initiative (CCMI) (Eyring et al., 2013; Morgenstern et al., 2017), which is a joint activity of SPARC (Stratosphere-troposphere processes and their role in Climate) and IGAC (International Global Atmospheric Chemistry), and is the successor activity to phase 2 of the Chemistry-Climate Model Validation activity, CCMVal-2 (SPARC CCMVal, 2010). Unlike CCMVal-2, which focussed on stratospheric processes and composition, CCMI includes many models with comprehensive representations of the troposphere, and aims to additionally address aspects of tropospheric chemistry and circulation. Here, we examine tropospheric column ozone in SOCOLv3.0 and 14 other CCMI models. This is the first time that global distributions of tropospheric ozone have been examined in the CCMI models, and results are presented in Section 3.3."
- Methods, section 2.1 ("CCM simulations to compare with observations.")

Minor Comments

[1] In the abstract (page 2, lines 1-2), you talk about the reduction in ozone bias due to the inclusion of the N2O5 hydrolysis process. Is this reduction in bias at the cost of an increase in bias for other variables (e.g. CH4 lifetime) when compared with observations? This isn't necessarily something you need to change in the abstract, but the inclusion of an extra sentence in the manuscript which addresses this comment would be useful.

If anything, calculated quantities such as the CH_4 lifetime should improve due to reductions in OH abundances (CH_4 + OH being the primary CH4 oxidation reaction). Historically SOCOLv3's simulated OH abundance has been too high, since ozone is the primary source of OH. Revell et al. (2015, www.atmos-chem-phys.net/15/5887/2015/) showed that this leads

to approximately 40 ppbv too little CO in the Northern Hemisphere compared with observations, because too much OH means too much CO is oxidised by CO + OH. Similarly, SOCOLv3's CH_4 lifetime was historically shorter than that calculated by other models. While the appropriate chemical reactions to calculate the CH_4 lifetime were not saved from our simulations, the simulated CO abundance has improved (the bias of -40 ppbv c.f. observations shown by Revell et al. (2015) has weakened to only -20 ppbv), and we have included a paragraph on that in the Discussion:

"Reducing SOCOL's tropospheric ozone bias is expected to lead to improvements in the simulated abundance of species which are oxidised by the hydroxyl radical, such as CO and CH_4, since ozone is the primary source of OH. Revell et al. (2015) showed that CO in SOCOLv3 was up to 40 ppbv too low in the Northern Hemisphere compared with observations from TES, due to the tropospheric ozone bias. In SOCOLv3.1, the Northern Hemisphere CO bias is reduced by approximately a factor of 2 (not shown)."

[2] Page 2, line 6. "More than 90%"? Adding up the first three columns of figure 8, it looks more like 80-90%.

When the joint interaction terms (NOx.CH4, NOx.CO and CH4.CO) are included, it comes to over 90% for all regions shown in Figure 8 (now Figure 5).

[3] In the title and elsewhere in the manuscript you mostly refer to the emulator as a "Gaussian emulator" (I found five mentions of this but there may be more). Please change all occurrences of this phrase to "Gaussian process emulator" (or "GP emulator" once GP is defined) since this is what you've implemented. A Gaussian (Normal) distribution is related to but is also quite different to a Gaussian process, so it's important to make this distinction. I'm guessing that you used 'Gaussian process' because of GEM-SA being short for 'Gaussian Emulation Machine for Sensitivity Analysis'. 'Gaussian emulation' was probably used here to make the acronym work, but it's unfortunately also caused confusion.

Thanks for explaining this! It has been changed to GP emulator throughout the manuscript.

[4] Page 3, lines 20-26. You've got to be a bit careful about the language used here. You imply that it's the GP emulator that doing the Global Sensitivity Analysis (GSA). The point is that you need to do 1000s of runs to the GSA, so the emulator (trained with only 90 simulator runs) is much more computationally efficient. I know that you probably know this, but at the moment this isn't clear to me when I read these lines.

This has been re-worded:

"Because thousands of simulations are required to perform a sensitivity analysis, and this would be computationally inefficient with a CCM, we supplement SOCOLv3.1 with a GP emulator. This allows a sensitivity analysis to be performed at low computational cost. Variance-based sensitivity analysis evaluates a suite of model input parameters, and their relationship to the variable of interest, simultaneously."

[5] Page 3, line 26. The word "non-linear" is the probably the wrong word to use here. I think what you're referring to are the sensitivity indices computed due the interaction of two inputs. If this is what you mean that I suggest you replace non-linear with "interacting".

Replaced as suggested.

[6] First line of section 2.4 (page 6). Please change the start of the sentence to "Variance-based global sensitivity analysis . . ."

Replaced as suggested.

[7] Page 3, line 30. Oliver Wild's group at Lancaster University are also using emulators for their work with the FRSGC and GISS models. A paper of theirs which has been accepted and will be published shortly is (Ryan et al., 2018; https://www.geosci-modeldev-discuss.net/gmd-2017-271/). Please add the following to the end of this sentence on line 30: ". . . and to chemical transport modelling (Ryan et al., 2018)" or something to that effect.

Done and thanks for the pointer to your paper.

[8] page 6, line 19 – what do you mean "supplement" here? Following the comma I suggest you replace the text with "..., a type of statistical model called Gaussian process emulator can be used as a surrogate for the input-output relation of the a complex model (Le Gratiet et al, 2017)." There are many other references from the statistics literature that could be included as well as the Le Gratiet ref.

Replaced as suggested.

[9] Page 7, 18. Can I suggest that you split this sentence beginning "90" into two sentences. The bit in brackets concerning the 10*n rule would be good to be taken out of the brackets and form the first sentence. Please also use the Loeppky et al. (2009) ref to justify the 10*n rule.

Replaced as suggested:

"Typically 10*n* simulations are recommended for training a GP emulator, where *n* is the number of parameters under investigation (Loeppky et al., 2009). Hence we performed 90 SOCOLv3.1 "training'" simulations, and used the resulting annual-mean tropospheric ozone column to construct the GP emulator in several geographical regions (Europe, United States, Asia, the Southern Ocean and the global mean).

[10] Page 7, line 21. Replace "statistical method called" with "design" since this is what a Maximin LHD is.

Replaced as suggested.

[11] Page 7, line 22-23. On the line that follows, replace "approach" with "design". What do you mean by "near random sample"? This seems incorrect to me. Also the phrase "maximizing the uncertainty space" doesn't sit comfortably with me either. A Maximin LHD

is a space filling design. It is an efficient design for sampling form a multidimensional parameter / input space in terms of being space filling but not requiring many samples. On page 169 of the pdf of my PhD thesis (given as page 155 in the footer) (Ryan et al., 2013), I give a fuller description if that'll help.

This sentence has been changed:

"For each of the 90 training simulations, the 9 input variables were scaled simultaneously, with the scaling factors determined using a "maximin" Latin hypercube design, which generates a random sample of parameter values from a multidimensional distribution and fills the uncertainty space of the parameters (McKay et al., 1979)."

[12] Page 7, lines 21-23. How did you generate the Maximin LHD? I haven't used GEM-SA in a long time, so I can't remember if it has a feature which generates the design for you?

Yes, GEM-SA can generate Latin hypercubes, and this is what was done here. This has now been noted in the text.

[13] Page 7, lines 21-24. I notice that some of your inputs are continuous (e.g. inputs 1-3) and some are discrete (e.g. input 4). Whenever I've built emulators, all of my inputs are continuous. Indeed, I think this is the norm when using a maximin LHD. For the statistical individuals like me reading this, please can you add in a line stating how you used this design for the inputs that are discrete? E.g. did you just round to the nearest whole number? Rounding to the nearest who number might be okay but it might not be. You might want to survey the literature a bit and what others have done.

Added: "The Latin hypercube was generated using GEM-SA. For the discrete input parameters (e.g. (4) and (5) in the list above), the scaling factor was rounded to the nearest whole number."

[14] Page 9, line 26 – page 10, line 7. The first two paragraphs and start of the third paragraph of section 3.2 aren't anything to do with emulation or sensitivity analysis so please move to a different section or create a new section.

Created a new section, "Tropospheric ozone in SOCOLv3.1."

[15] page 10, line 9. Please don't use the word "correlation". Correlation is represented by 'r' and takes values between -1 and 1. R^2 is a measure of "goodness of fit" (takes values 0-1) which in this case refers to how well the emulated outputs compare with the simulator outputs at the validation inputs.

This has been corrected.

[16] Page 10, lines 17/18. You state here ". . . assuming all other parameters are held constant." This is wrong. This is what happens with one at a time sensitivity analysis. With variance based global sensitivity analysis, we average over the other inputs. See slide 9 of:

https://view.officeapps.live.com/op/view.aspx?src=http://www.tonyohagan.co.uk/academic/GEM/SensitivityAnalysis.ppt

This has been corrected.

[17] Page 11, line 33. You mention Young et al. (2018). From memory this is one of the TOAR papers where the chemistry models are compared with observations from the newly formed TOAR network. If you are going to keep figs 2-5 in their current form, then it seems that Young et al. (2018) is a key paper that you need to refer to a lot earlier in the paper (e.g. intro and methods).

This is now cited in the Introduction, as also requested by Reviewer 2:

"Most chemistry-climate models (CCMs), which are used to understand chemistry-climate interactions and project future atmospheric composition, overestimate tropospheric ozone in the Northern Hemisphere compared with observations (Young et al., 2013; Parrish et al., 2014; Young et al., 2018)."

[18] Page 13. Data availability section. For the benefit of reproducibility, please can you make the matrix of inputs and outputs that were used in GEM-SA to generate your sensitivity analysis results.

Certainly; these are now available in the supplement.

[19] Figure 1: When we do variance based global sensitivity analysis, the inputs are normalized to all be between 0 and 1. I think GEM-SA does this automatically. I mention this because it would look a lot better if the y-axis on figure 1 referred to the normalized inputs. By normalized I mean: x_norm = (x – xmin)/(xmax-xmin). This would make the points in figure 1 appear more randomly scattered as opposed to the larger gaps for the higher values of the inputs because of only some of the inputs extend to 4 or 6.

This has been changed as suggested.

[20] Figure 1. Are all of your inputs scaling factors? It seems not since for example input 4 is the "number of vertical levels . . .". If you agree, please change the y-axis label to "Inputs" or "parameters".

Changed to "Inputs."

[21] Figure 6. The caption is quite short here. I know that in the methods you described the simulator runs for validation of the emulators as "test simulations". But at some point in this caption you need to explain that these runs correspond to running the emulators and simulators at each of the 27 validation inputs. You also need to explain what each of the

panels refer to? I know it might seem obvious, but my view is that I should be able to understand everything about each figure without having to refer to the manuscript text. You also need to describe what R^2 is (not correlation, look it up on Wikipedia).

The caption has been changed to:

"Tropospheric column ozone as predicted by the GP emulator, vs. the amount simulated in SOCOLv3.1 "test" simulations (i.e., the simulations used to validate the emulator). The errorbars indicate the uncertainty on the GP emulator output, and the 1:1 line and coefficient of determination (R^2 value) are also shown. These simulations correspond to running the GP emulator and the simulator (SOCOLv3.1) at each of the 27 validation inputs, for: (a) Europe (37-60° N), 0-42° E); (b) United States (32-52° N, 67-124° W); (c) Asia (6-49° N, 70-146° E), (d) the Southern Ocean (45–60° S, all longitudes); and (e) globally."

[22] Figure 7. In the caption please replace "assuming the other variables are constant" with "averaging over the other inputs." You state this plot is representative of the other regions. Please can you put the equivalent plots for the other regions in the supplemental material.

Changed as suggested, and the equivalent plots are now in the supplement.

[23] Figure 8. Why not show the sensitivity indices for all nine inputs? I know that you say that you're not including the missing two because they are less than 1%, but for completeness (and given that it's only an extra two bars), I think it's worth including them.

We also show the joint interaction terms (e.g. NOx.CH4), making 45 possible terms to show in total – hence the decision to limit the number of terms plotted.

[24] Table 1. Is it accurate to describe all the inputs has "scaling". E.g. input 4 is not a scaling since it's the no. of levels.

Good point – it has been re-labeled as "range of the sensitivity forcings/parametrizations."

[25] Table 1. You have a "Comments" column, but I think that replacing this with "Descriptions" and giving a full definition of what each input is would be better.

Done, as suggested. The revised table is shown above.

*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".

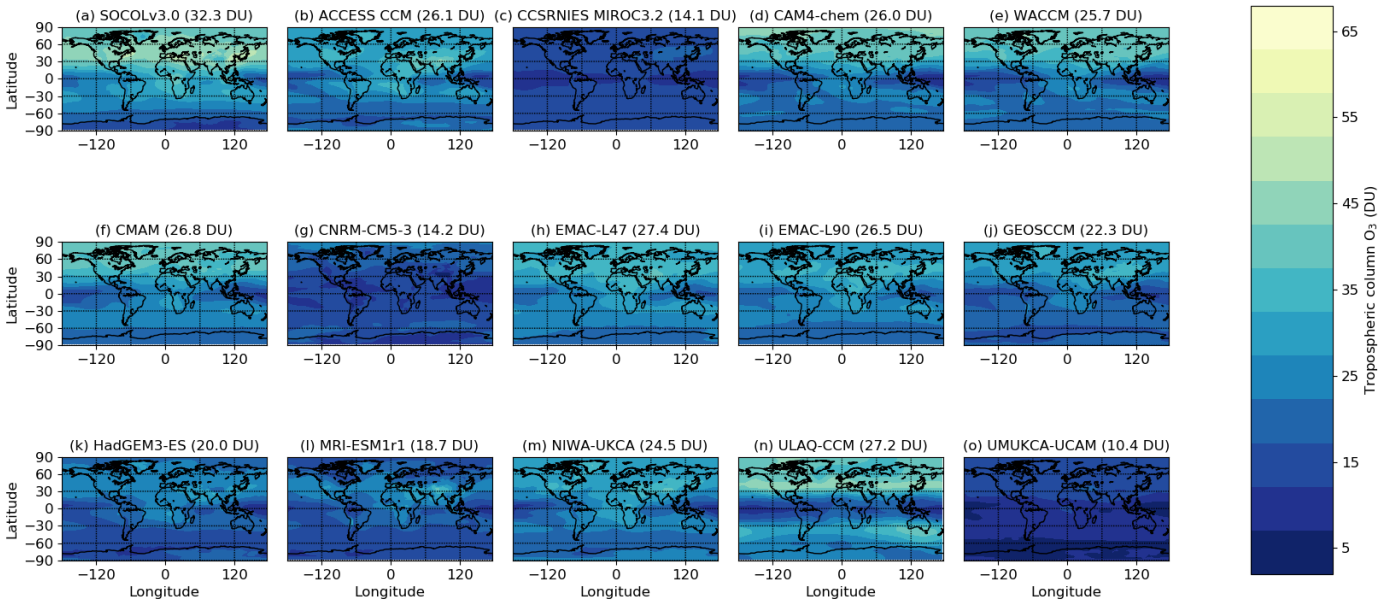
12) P5L16: Naively, I would not expect methane-induced ozone production to be reduced upon prescribing methane on one level versus multiple levels since it is well mixed in the troposphere.

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.

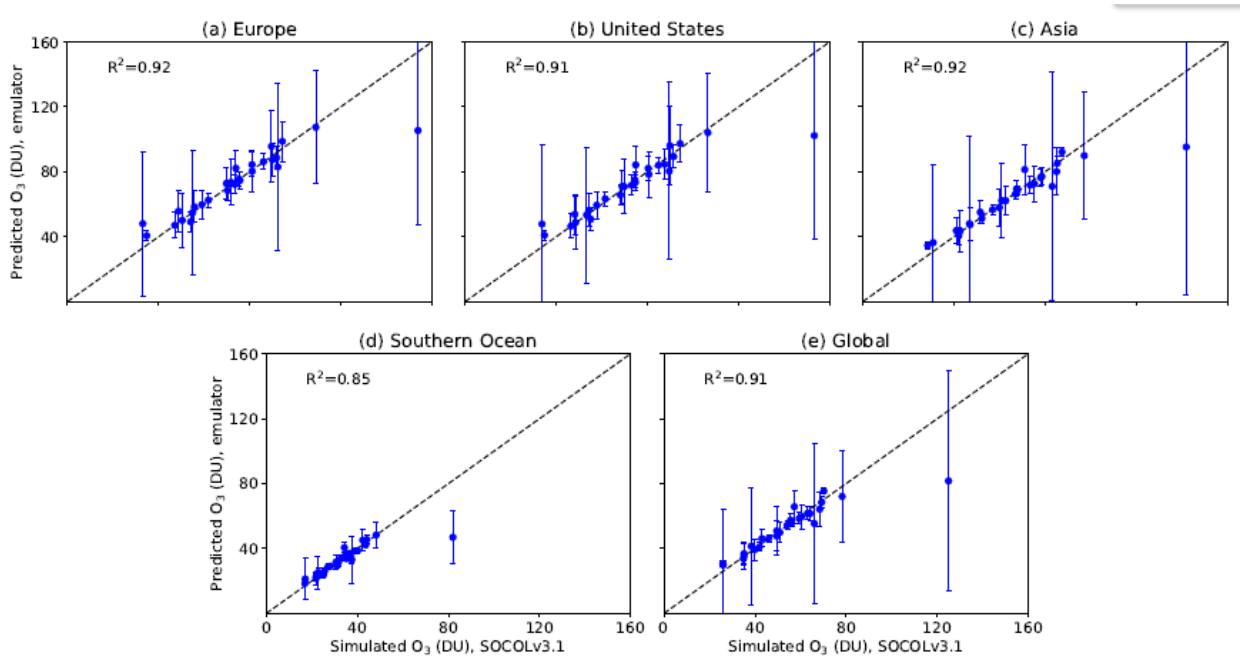
The following text has been added to the Introduction:

"Our GP emulator experiments have been designed to focus on recent developments surrounding SOCOL's tropospheric chemistry scheme, however the methodology has the potential to be expanded to also include meteorological parameters."

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):

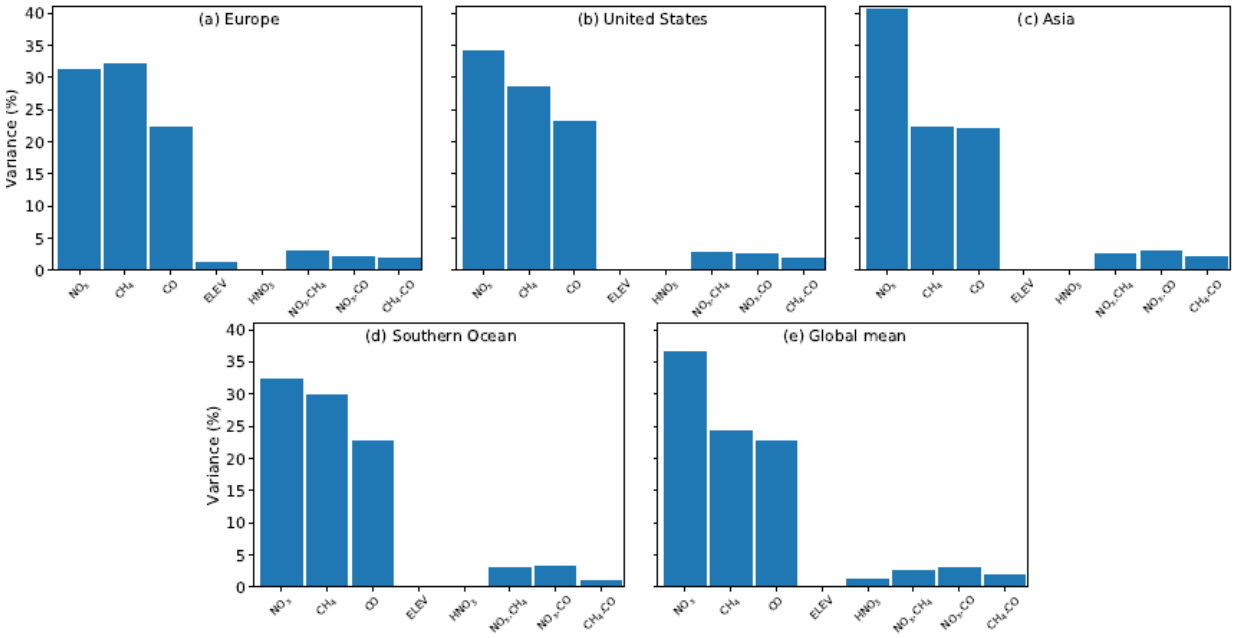


Figure 7 also shows global-mean results, with the corresponding plots for Europe, the US, Asia and Southern Ocean moved to the supplement.

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."

Tropospheric ozone in CCMI models and Gaussian process emulation to understand biases in the SOCOLv3 chemistry-climate model

Laura E. Revell[1,2,3], Andrea Stenke[2], Fiona Tummon[2,4], Aryeh Feinberg[2], Eugene Rozanov[2,5], Thomas Peter[2], N. Luke Abraham[6,7], Hideharu Akiyoshi[8], Alexander T. Archibald[6,7], Neal Butchart[9], Makoto Deushi[10], Patrick Jöckel[11], Douglas Kinnison[12], Martine Michou[13], Olaf Morgenstern[14], Fiona M. O'Connor[9], Luke D. Oman[15], Giovanni Pitari[16], David A. Plummer[17], Robyn Schofield[18,19], Kane Stone[18,19,20], Simone Tilmes[12], Daniele Visioni[16], Yousuke Yamashita[8,21], and Guang Zeng[14]

[1]School of Physical and Chemical Sciences, University of Canterbury, Christchurch, New Zealand
[2]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
[3]Bodeker Scientific, Christchurch, New Zealand
[4]Now at: Biosciences, Fisheries, and Economics Faculty, University of Tromsø, Norway
[5]Physical-Meteorological Observatory/World Radiation Center, Davos, Switzerland
[6]Department of Chemistry, University of Cambridge, Cambridge, UK
[7]National Centre for Atmospheric Science (NCAS), UK
[8]National Institute of Environmental Studies (NIES), Tsukuba, Japan
[9]Met Office Hadley Centre (MOHC), Exeter, UK
[10]Meteorological Research Institute (MRI), Tsukuba, Japan
[11]Institut für Physik der Atmosphäre, Deutsches Zentrum für Luft- und Raumfahrt (DLR), Oberpfaffenhofen, Germany
[12]National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA
[13]CNRM UMR 3589, Météo-France/CNRS, Toulouse, France
[14]National Institute of Water and Atmospheric Research (NIWA), Wellington, New Zealand
[15]National Aeronautics and Space Administration Goddard Space Flight Center (NASA GSFC), Greenbelt, Maryland, USA
[16]Department of Physical and Chemical Sciences, Universitá dell'Aquila, L'Aquila, Italy
[17]Environment and Climate Change Canada, Montréal, Canada
[18]School of Earth Sciences, University of Melbourne, Melbourne, Victoria, Australia
[19]ARC Centre of Excellence for Climate System Science, University of New South Wales, Sydney, Australia
[20]Now at: Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA
[21]Now at: Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokohama, Japan

Correspondence: Laura Revell (laura.revell@canterbury.ac.nz)

Abstract. Previous multi-model intercomparisons have shown that chemistry-climate models exhibit significant biases in tropospheric ozone compared with observations. We investigate annual-mean tropospheric column ozone in 15 models participating in the SPARC/IGAC (Stratosphere-troposphere Processes and their Role in Climate/International Global Atmospheric Chemistry) Chemistry-Climate Model Initiative (CCMI). These models exhibit a positive bias, on average, of up to 40–50% in the Northern Hemisphere compared with observations derived from the Ozone Monitoring Instrument and Microwave Limb Sounder (OMI/MLS), and a negative bias of up to ∼30% in the Southern Hemisphere. SOCOLv3.0 (version 3 of the Solar-Climate Ozone Links CCM), which participated in CCMI, simulates global-mean tropospheric ozone columns of 40.2 DU –

approximately 33% larger than the CCMI multi-model mean. Here we introduce an updated version of SOCOLv3.0, "SOCOLv3.1", which includes an improved treatment of ozone sink processes, and results in a reduction in the tropospheric column ozone bias of up to 8 DU, mostly due to the inclusion of N_2O_5 hydrolysis on tropospheric aerosols. As a result of these developments, tropospheric column ozone amounts simulated by SOCOLv3.1 are comparable with several other CCMI models. We apply Gaussian process emulation and sensitivity analysis to understand the remaining ozone bias in SOCOLv3.1. This shows that ozone precursors (nitrogen oxides (NO_x), carbon monoxide, methane and other volatile organic compounds) are responsible for more than 90% of the variance in tropospheric ozone. However, it may not be the emissions inventories themselves that result in the bias, but how the emissions are handled in SOCOLv3.1, and we discuss this in the wider context of the other CCMI models. Given that the emissions data set to be used for phase 6 of the Coupled Model Intercomparison Project includes approximately 20% more NO_x than the data set used for CCMI, further work is urgently needed to address the challenges of simulating sub-grid processes of importance to tropospheric ozone in the current generation of chemistry-climate models.

1 Introduction

Ozone is a key trace gas in the atmosphere. In the stratosphere, it absorbs UV-B ($280<\lambda<320$ nm) radiation and thus protects life at the surface. However in the troposphere, where approximately 10% of the total atmospheric ozone burden resides, ozone is a greenhouse gas and air pollutant, with adverse affects on human health and crop yields (Myhre et al., 2013; Stevenson et al., 2013; Silva et al., 2013, 2017). Approximately 90% of tropospheric ozone results from a series of photochemical reactions which are initiated by the reaction of NO_x (nitrogen oxides, NO_x = NO+NO_2) and either CO (carbon monoxide), CH_4 (methane) or a NMVOC (non-methane volatile organic compound) (Denman et al., 2007). These ozone precursors are emitted from, amongst other sources, fossil fuel burning, industrial processes and agriculture. Ozone can also be transported from the stratosphere in stratosphere-troposphere exchange (STE) events. 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%, ~~but~~**and** show that during individual STE events, over 10% of tropospheric ozone may be directly transported from the stratosphere.

Due to its **global tropospheric lifetime of** ∼22 day**s** ~~lifetime in the troposphere~~, ozone is subject to intercontinental transport (Auvray and Bey, 2005), and this is modulated by decadal climate variability (Lin et al., 2014). Ozone is lost from the troposphere either by dry deposition or photochemical destruction.

Most chemistry-climate models (CCMs), which are used to understand chemistry-climate interactions and project future atmospheric composition, overestimate tropospheric ozone in the Northern Hemisphere compared with observations (Young et al., 2013; Parrish et al., 2014; Young et al., 2018). In particular, version 3.0 of the SOCOL (Solar-Climate Ozone Links) CCM (Section 2.2) contains notable positive tropospheric ozone biases. Revell et al. (2015) identified that ozone concentrations in SOCOLv3.0 are up to 50% too high in the Northern Hemisphere mid-troposphere (500 hPa) compared with observations from the Tropospheric Emission Spectrometer (TES). The reasons underlying SOCOLv3.0's tropospheric ozone bias were not completely clear to Revell et al. (2015), who noted that, while SOCOLv3.0 could accurately simulate the general geographic

distribution of tropospheric ozone, the actual magnitude was wrong and likely to be "a source issue (that is, emissions), a sink issue (HNO_3 washout), or a combination of the two."

Staehelin et al. (2017) showed that the mean tropospheric ozone burden in SOCOLv3.0 is 413 Tg, which is approximately 80 Tg larger than the multi-model mean burdens reported for the ACCENT (Atmospheric Composition Change: the European Network of Excellence (Stevenson et al., 2006)) and ACCMIP (Atmospheric Chemistry and Climate Model Intercomparison Project (Young et al., 2013)) activities, of 337 and 336 Tg, respectively. Furthermore, SOCOLv3.0 overestimates both the tropospheric ozone production and destruction rates compared to the multi-model means from ACCENT and ACCMIP (Staehelin et al., 2017). While SOCOLv3.0's production rates are overestimated by 34% compared to ACCENT and 41% compared to ACCMIP, the destruction rates are overestimated only by 20% (ACCENT) and 31% (ACCMIP).

~~SOCOLv3.0 participated in phase 1 of the Chemistry-Climate Model Initiative (CCMI) (Eyring et al. 2013b, Morgenstern et al. 2017), which is a joint activity of SPARC (Stratosphere-troposphere processes and their role in Climate) and IGAC (International Global Atmospheric Chemistry), and is the successor activity to phase 2 of the Chemistry-Climate Model Validation activity, CCMVal-2 (CCMVal-2). Unlike CCMVal-2, which focussed on stratospheric processes and composition, CCMI includes many models with comprehensive representations of the troposphere, and aims to additionally address aspects of tropospheric chemistry and circulation. Here, we examine tropospheric column ozone in SOCOLv3.0 and 14 other CCMI models in Section x.~~

Recently a newer version of SOCOL has been developed, "SOCOLv3.1", which remediates obvious deficiencies in SOCOLv3.0's representation of tropospheric processes (Section 2.3). We compare tropospheric column ozone in SOCOLv3.0 and 3.1 with observations derived from OMI/MLS, the Ozone Monitoring Instrument/Microwave Limb Sounder (Section 3.1), and use Gaussian process **(GP)** emulation and sensitivity analysis to investigate the remaining ozone bias in SOCOLv3.1 (Section 3.2). **Because thousands of simulations are required to perform a sensitivity analysis, and this would be computationally inefficient with a CCM, we supplement SOCOLv3.1 with a GP emulator. This allows a sensitivity analysis to be performed at low computational cost. Variance-based sensitivity analysis evaluates** ~~the uncertainties to be evaluated simultaneously at low computational cost. GP emulation and variance-based sensitivity analysis allows~~ a suite of model input parameters, and their relationship to the variable of interest~~,~~ **simultaneously.**~~to be evaluated over a range of uncertainties of the inputs simultaneously at low computational cost.~~

Here, we apply ~~it~~ **GP emulation and variance-based sensitivity analysis** to the SOCOLv3.1 tropospheric ozone budget to understand causes of the bias. In contrast to one-at-a-time testing, which investigates the model response to varying one input parameter while holding all others constant, G**P** ~~aussian~~ emulation allows all parameters to be evaluated simultaneously and covers more of the parametric uncertainty space than one-at-a-time testing. G**P**~~aussian~~ emulation is computationally efficient and allows the ~~non-linear~~ **interacting** effects of the uncertainties on different input parameters to be accounted for. It also generates much more information than one-at-a-time testing – typically the same level of information as a Monte Carlo approach, but requiring a fraction of the model simulations (O'Hagan, 2006). **GP emulation has been used by the** ~~Within the~~ global atmospheric modelling community **only in the last few years, in applications such as** ~~GP emulation has previously been applied to~~ cloud and aerosol microphysics modelling (Lee et al., 2011, 2012; Carslaw et al., 2013; Johnson et

al., 2015) and chemical transport modelling (Ryan et al., 2018). **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.**

SOCOLv3.0 participated in phase 1 of the Chemistry-Climate Model Initiative (CCMI) (Eyring et al., 2013; Morgenstern et al., 2017), which is a joint activity of SPARC (Stratosphere-troposphere processes and their role in Climate) and IGAC (International Global Atmospheric Chemistry), and is the successor activity to phase 2 of the Chemistry-Climate Model Validation activity, CCMVal-2 (SPARC CCMVal, 2010). Unlike CCMVal-2, which focussed on stratospheric processes and composition, CCMI includes many models with comprehensive representations of the troposphere, and aims to additionally address aspects of tropospheric chemistry and circulation. Here, we examine tropospheric column ozone in SOCOLv3.0 and 14 other CCMI models. ~~in Section x.~~ **This is the first time that global distributions of tropospheric ozone have been examined in the CCMI models, and results are presented in Section 3.3**.

2 Computational and statistical methods

2.1 CCM simulations to compare with observations

We use the ensemble mean of three free-running SOCOLv3.0 simulations of the recent past to compare with observations (ETH-PMOD, 2015). These simulations were performed for CCMI, and conform to REF-C1 specifications (Eyring et al., 2013). The simulations cover the period 1960–2010, following a 10-year spin-up period. 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 (WMO, 2011), and stratospheric aerosol surface area densities and optical parameters were prescribed from the SAGE-4λ data set (Arfeuille et al., 2013; Luo, 2013).

We also examine annual-mean tropospheric ozone in REF-C1 simulations performed by models participating in CCMI, described by Morgenstern et al. (2017) and references therein. Using the simulated ozone volume mixing ratio and WMO-defined tropopause height from each model, tropospheric ozone columns were calculated for the year 2005 by integrating ozone between the surface and WMO-defined tropopause. The WMO definition of the tropopause was selected to be consistent with the OMI/MLS tropospheric ozone product (Ziemke et al., 2006). Between 2010–2014, the average tropospheric ozone burden derived from OMI/MLS was 300 Tg, which is very close to the multi-instrument mean of five satellite products over the same period, of 301 Tg (Gaudel et al., 2018).

Where multiple ensemble members ('realisations') of the REF-C1 simulation were submitted to the CCMI archive, the ensemble mean is shown. The exception is NIWA-UKCA, which submitted three realisations of the REF-C1 simu-

lation, however only the first realisation is shown as ozone precursor emissions were erroneously fixed at 1960 levels for the other two realisations (Morgenstern et al., 2017). The EMAC simulations used road traffic emissions which were updated every year rather than every month. Therefore when we examine year 2005 tropospheric column ozone in Section 3.3, the EMAC simulations used road traffic emissions for August 1954. Jöckel et al. (2016) show that this error results in tropospheric ozone columns that are ∼2 DU lower than if the correct emissions were used. The UMUKCA-UCAM simulations used CCMVal-2 REF-B2 emissions for NO_x aircraft emissions and NO_x, CO and HCHO surface emissions.

2.2 The SOCOLv3.0 chemistry-climate model

The SOCOL **CCM**~~model~~ was developed in Switzerland at ETH Zurich and PMOD/WRC (the Physical Meteorological Observatory Davos/World Radiation Center). Version 3.0 of SOCOL (Stenke et al., 2013; Revell et al., 2015) consists of the middle atmosphere version of the ECHAM5 (European Centre Hamburg Model) atmosphere-only general circulation model (Roeckner et al., 2003) coupled to the MEZON (Model for Ozone Trends) chemistry transport model (Egorova et al., 2003). The chemical solver takes into account 41 chemical species, 140 gas-phase reactions, 46 photolysis reactions and 16 heterogeneous reactions. The oxidation of isoprene, an important NMVOC for the tropospheric ozone budget, is accounted for with the Mainz Isoprene Mechanism (MIM-1), which comprises 16 organic degradation products of isoprene and a further 44 chemical reactions (Pöschl et al., 2000). Global isoprene emissions are estimated to range from 440 to 660 Tg(C)/yr, which is comparable to the annual amount of CH_4 emissions (Guenther et al., 2006). About two thirds of the annual global emissions of volatile organic compounds (VOCs) are accounted for in SOCOLv3.0 by isoprene and methane. Apart from isoprene and formaldehyde, other NMVOCs are not included explicitly in the model but their contribution to CO is accounted for via the addition of a certain fraction of NMVOC emissions to CO. For anthropogenic, biomass burning and biogenic NMVOC emissions the conversion factors to CO are 1.0, 0.31 and 0.83, respectively (Ehhalt et al., 2001).

Clear-sky photolysis rates are calculated ~~online~~ using a look-up-table (LUT) approach, which provides photolysis rates as a function of overhead ozone and oxygen columns (Rozanov et al., 1999). Variability of solar irradiance is included in the LUTs. Cloud impacts on photolysis are accounted for in the troposphere by the inclusion of a cloud modification factor following the parametrization described by Chang et al. (1987). From a recent intercomparison of photolysis rates simulated by different CCMI models we learned that SOCOLv3.0 overestimates tropospheric NO_2 photolysis by roughly a factor of 2 compared to other models (Nicely et al., 2018). This overestimation is likely related to the treatment of backscattering from clouds in the calculations of the photolysis LUTs and the missing impact of aerosols. Both effects cannot be easily corrected by the implemented cloud modification factor, and so an online photolysis scheme is planned for future model versions.

Dry deposition velocities of O_3, CO, NO, NO_2, HNO_3 and H_2O_2 are based on Hauglustaine et al. (1994). This simplified approach assumes constant dry deposition velocities over land and ocean, without accounting for seasonal or geographical variability. The tropospheric wash-out of HNO_3 and H_2O_2 is calculated by using a constant removal rate of 4×10^{-6} s^{-1}, irrespective of precipitation occurrence. At every chemical time step, i.e., every two hours, 2.8% of tropospheric HNO_3 and H_2O_2 below 160 hPa are removed. Boundary conditions for the ozone precursor gases NO_x, CO and NMVOCs are imple-

mented as surface emission fluxes. ~~Methane is prescribed as a global average surface mixing ratio.~~ For this study, both SOCOL configurations were run with 39 vertical levels between Earth's surface and 0.01 hPa (∼80 km) and T42 horizontal resolution (grid cells approximately 2.8° × 2.8°).

2.3 Upgraded model version SOCOLv3.1

SOCOLv3.1 was developed to address SOCOLv3.0's representation of processes relevant to tropospheric ozone chemistry, with the aim of improving the model's large tropospheric ozone bias as shown by Revell et al. (2015).

First, we implemented heterogeneous hydrolysis of N_2O_5 on tropospheric aerosol, as this is an important removal process for atmospheric NO_x and was not included in SOCOLv3.0. As SOCOLv3.0 does not explicitly simulate tropospheric aerosols, the new scheme makes use of the ECHAM5 internal tropospheric aerosol climatology considering aerosol properties of 11 Global Aerosol Data Sets types (Köpke et al., 1997). The reaction probabilities for the different aerosol types are calculated following the parametrization by Evans and Jacob (2005).

Second, the simplified treatment of dry deposition was replaced by a more sophisticated scheme in SOCOLv3.1 based on the surface resistances approach for the estimation of dry deposition velocities proposed by Wesely (1989). This takes into account actual meteorological conditions, different surface types and trace gas properties like solubility and reactivity. Further details of this scheme are given by Kerkweg et al. (2006).

Third, we adjusted how methane is prescribed in the model. In previous versions of SOCOL, methane was prescribed as a global surface average mixing ratio on the six lowermost model levels (covering approximately 2.5 km). This was changed to only the surface level in SOCOLv3.1. While the amount of methane entering the atmosphere is the same in both configurations, prescribing it on one level instead of six means that methane-induced ozone production in the mid-to-upper troposphere is reduced. Because SOCOLv3 has a high OH bias compared to the ACCMIP models (Staehelin et al., 2017), ozone production from methane oxidation is amplified by the continuous re-supply of methane due to the mixing ratio boundary condition when methane is prescribed on six levels instead of one. An interhemispheric gradient and seasonal cycle in methane have also been implemented in SOCOLv3.1, however these were not used in this study and instead methane was prescribed as a global average surface mixing ratio to test the general sensitivity of tropospheric ozone to surface methane concentrations.

Finally, because the LUTs used in SOCOLv3.0 cause tropospheric NO_2 photolysis to be overestimated due to the treatment of backscattering from clouds (Section 2.2), we recalculated LUTs for SOCOLv3.1. While the SOCOLv3.0 LUTs were calculated assuming 0.5 cloud coverage and a surface albedo of 0.3, the SOCOLv3.1 LUTs were based on clear-sky conditions and also used a surface albedo of 0.3.

~~We use the ensemble mean of three free-running SOCOLv3.0 simulations of the recent past to compare with observations (ETH, 2015). These simulations were performed for CCMI, and conform to REF-C1 specifications (Eyring et al., 2015b). The simulations cover the period 1960–2010, following a 10-year spin-up period. 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 (WMO 2011), and stratospheric aerosol surface area densities and optical parameters were prescribed from the SAGE-4λ data set (Arfeuille et al., 2013; Luo 2013).

We also examine annual-mean tropospheric ozone in REF-C1 simulations performed by models participating in CCMI, described by (Morgenstern et al., 2017) and references therein. Using the simulated ozone volume mixing ratio and WMO-defined tropopause height from each model, tropospheric ozone columns were calculated for the year 2005 by integrating ozone between the surface and WMO-defined tropopause. The WMO definition of the tropopause was selected to be consistent with the OMI/MLS tropospheric ozone product (Ziemke et al., 2006). Between 2010–2014, the average tropospheric ozone burden derived from OMI/MLS was 300 Tg, which is very close to the multi-instrument mean of five satellite products over the same period, of 301 Tg (Gaudel et al., 2018).

Where multiple ensemble members ('realisations') of the REF-C1 simulation were submitted to the CCMI archive, the ensemble mean is shown. The exception is NIWA-UKCA, which submitted three realisations of the REF-C1 simulation, however only the first realisation is shown as ozone precursor emissions were erroneously fixed at 1960 levels for the other two realisations (Morgenstern et al., 2017). The EMAC simulations used road traffic emissions which were updated every year rather than every month. Therefore when we examine year 2005 tropospheric column ozone in Section x, the EMAC simulations used road traffic emissions for August 1954. Jockel et al. (2016) show that this error results in tropospheric ozone columns that are ~2 DU lower than if the correct emissions were used. The UMUKCA-UCAM simulations used CCMVal-2 REF-B2 emissions for NO_x aircraft emissions and NO_x, CO and HCHO surface emissions.

2.5 SOCOLv3.1 simulations for GP emulator training and testing

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** a GP emulator can be used to supplement a complex model with a statistical 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** The output variable of interest (here tropospheric column ozone) is fitted with a mean function, and uncertainties are calculated with a**nd** covariance function **(O'Hagan, 2006).**, assuming that each unknown output point has a normal distribution. **Here, we use tropospheric ozone columns from SOCOLv3.1 to train the emulator. Interacting**Non-linear 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 G**P**aussian 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.

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 ~~include:~~ **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.**

1. **Natural and anthropogenic** NO_x emissions (Denoted in figures as 'NO_x').
2. ~~Surface m~~**M**ethane concentrations ('CH_4').
3. CO **emissions (natural and anthropogenic), and+** NMVOC emissions **(anthropogenic, biogenic and biomass burning)** ('CO').
4. the number of vertical levels NO_x and CO+NMVOC emissions were prescribed on in the model ('ELEV').
5. the number of vertical levels CH_4 concentrations were prescribed on in the model ('CLEV').
6. the impact of clouds on photolysis rates, via the cloud modification factor ('CMF').
7. the rate of HNO_3 washout ('HNO_3').
8. the N_2O_5 uptake coefficient, which represents the probability of N_2O_5 hydrolysis occurring ('N_2O_5').
9. the specific reactivities for ozone dry deposition ('O_3DD'), which are used to estimate the dry deposition velocity.

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.** ~~SOCOL contains only two NMVOCs, isoprene and formaldehyde, and other NMVOCs are represented as additional CO in the model (Section 2.2), hence for (3), CO and NMVOC emissions were varied simultaneously.~~

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 ~~parameters~~**variable** ~~(4) and~~ (5) w**as**~~ere~~ included in our analysis to investigate the sensitivity of tropospheric ozone to this implementation ~~for all ozone precursors~~. 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.**

~~Parameter (6) was chosen b~~**B**ecause ozone production and destruction reactions are mostly photochemical, i.e. they occur in the presence of sunlight, ~~and~~ **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**. ~~Parameter (7) was selected because~~ 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**. ~~Parameter (8) is similarly important as h~~**H**eterogeneous 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.** ~~Parameter~~ **Finally, variable** (9) was chosen to test the sensitivity of tropospheric ozone to the newly-implemented dry deposition parametrization (Section 2.3).

Typically 10*n* simulations are recommended for training a GP emulator, where *n* is the number of variables under investigation (Loeppky et al., 2009). Hence we performed 90 SOCOLv3.1 "training" simulations ~~(i.e. 10*n*, where *n* is the number of input parameters under investigation)~~, and **used** the resulting annual-mean tropospheric ozone column ~~was used~~ to construct the G**P**~~aussian process~~ emulator in several geographical regions (Europe, United States, Asia, the Southern Ocean and the global mean). For each of the 90 training simulations, the 9 input variables were scaled simultaneously, with the scaling factors determined using a ~~statistical method called a~~ "maximin" Latin hypercube **design**~~approach~~, which generates a ~~near-~~random sample of parameter values from a multidimensional distribution **and fills** ~~while also maximising~~ the uncertainty space of the parameters (McKay et al., 1979). **The Latin hypercube was generated using GEM-SA. For the discrete input parameters (e.g. (4) and (5) in the list above), the scaling factor was rounded to the nearest whole number.** Table 1 summarises the minimum and maximum scalings applied to each of the 9 variables. ~~These selected ranges are not necessarily feasible, but were selected to cover a range of parametric uncertainties.~~ This is discussed further in Section 3.2. Figure 1 shows the experimental design for the 90 training simulations, and is explained in further detail by Supplementary Video 1.

SOCOLv3.1 training simulations were performed for the year 2005 (following a common model spin-up period of 10 years, which was discarded from our analysis). The feedback between chemistry and radiation was switched off to keep internal variability as small as possible. Switching off the chemistry–radiation feedback means that all simulations have the same meteorology (given that they started from the same initial conditions and ran with the same dynamical boundary conditions), despite having different chemistry. Therefore, we can be confident that the differences between the simulations, ~~for example due to tropospheric ozone,~~ are caused by differences in chemistry and not dynamics.

The emulator was constructed using tropospheric ozone columns calculated between the surface and the WMO-defined tropopause. We focus on four regions, namely Europe (37-60° N, 0-42° E), the United States (32-52° N, 67-124° W), Asia (6-49° N, 70-146° E) and the Southern Ocean (45–60° S, all longitudes), where different chemical regimes may dominate, e.g. Sillman et al. (1990).

After constructing the **GP** emulator, the next step is to validate it by comparing emulator-predicted ozone with SOCOL-simulated ozone. This was done by performing a further 27 (i.e. 3*n*) SOCOLv3.1 "testing" simulations. The set-up for these simulations was similar to the training simulations, with a new Latin hypercube generated **by GEM-SA** to supply the scaling factors.

3 Results

~~**Tropospheric ozone in the CCMI models** Figure x shows tropospheric ozone in the CCMI models, and illustrates the diversity amongst the models. Despite most of the models using ozone precursor emissions following the REF-C1 recommendations (see Section x), they simulate vastly different representations of tropospheric ozone. A few of the models are closely related; for example the CESM1 models, WACCM and CAM4-chem, are essentially the same model in terms of tropospheric ozone. They differ only in the height of the model lid, which is 140 km for WACCM and 40 km for CAM4-Chem. ACCESS and NIWA-UKCA can also be considered the same model for the REF-C1 experiment; although a coupled ocean was used for most of NIWA-UKCA's CCMI simulations, for the REF-C1 experiment they used the same prescribed sea surface conditions (temperature and ice coverage) as ACCESS. Differences between ACCESS and NIWA-UKCA in the REF-C1 simulation, therefore, are likely related to issues with the different compilers used which may induce small differences in stochastic physics and tropospheric age of air (Dietmuller et al., 2018). The EMAC L47 and L90 models are also very similar; both have a model lid at 0.01 hPa (~80 km), but they differ in the number of model levels between the surface and 0.01 hPa (47 and 90, respectively). They also use different time steps. Interestingly, EMAC-L90 simulates a better representation of tropospheric column ozone than EMAC-L47, despite the fact that EMAC-L90 has three fewer model levels between the surface and 300 hPa than EMAC-L47 and a longer time step. The difference in tropospheric column ozone between the two models likely results from the increased vertical resolution around the tropopause in EMAC-L90, which has 11 levels between 300–100 hPa compared with 7 in EMAC-L47, meaning that EMAC-L90 better simulates stratosphere-troposphere exchange.~~

Figure 7 shows the difference in tropospheric ozone between each of the CCMI models and OMI/MLS, and the root-mean-square error (RMSE) for the model-OMI/MLS difference. Alongside Fig. 6, Fig. 7 indicates clear outlying models in terms of tropospheric ozone. UMUKCA-UCAM simulates the smallest amount of tropospheric ozone (14.9 DU in the global mean, Fig. 6o), however it only contains one NMVOC (formaldehyde) and does not 'lump' NMVOCs together in the way that many other CCMI models do. This means that additional NMVOC source gases are not considered by substituting with represented species, such as e.g. in SOCOLv3, whereby additional NMVOCs are included in the form of CO. Of the CCMI models, SOCOLv3.0 simulates the largest global-mean tropospheric ozone column, of 40.2 DU (Fig. 6a). SOCOLv3.0's tropospheric ozone bias is investigated further in Section 3.2. In ULAQ-CCM, the zonal bands of large ozone abundances at northern and southern midlatitudes are related to the model's coarse horizontal resolution ($5.6^\circ \times 5.6^\circ$), which affects surface fluxes and tropospheric transport (Orbe et al., 2018). **Interestingly, EMAC-L90 simulates a better representation of tropospheric column ozone than EMAC-L47, despite the fact that EMAC-L90 has three fewer model levels between the surface and 300 hPa than EMAC-L47 and a longer time step. The difference in tropospheric column ozone between the two models likely results from the increased vertical resolution around the tropopause in EMAC-L90, which has 11 levels between 300–100 hPa compared with 7 in EMAC-L47, meaning that EMAC-L90 better simulates stratosphere-troposphere exchange.**

Figure 8 shows multi-model means (MMM) and standard deviations. The MMM in Fig. 8a was calculated for all models, while the MMM in Fig. 8d was calculated only for models with a RMSE less than 10, as indicated in Fig. 7 – i.e., all models

except SOCOLv3.0, ACCESS CCM, EMAC-L47, ULAQ-CCM and UMUKCA-UCAM. The CCMI models simulate a global-mean tropospheric ozone abundance of 31.1 DU (Fig. 8a), and 30.2 DU (Fig. 8d), depending on the MMM definition applied. Both global-mean MMMs are close to the OMI/MLS global mean of 28.6 DU (Fig. 2b). However, the MMMs differ markedly from OMI/MLS in terms of the global tropospheric ozone distribution.

Compared to OMI/MLS, the models overestimate tropospheric column ozone almost everywhere between 60° N–60° S (the region where OMI/MLS data are available), regardless of the MMM definition. The exception is at southern midlatitudes, where the models underestimate tropospheric ozone compared to OMI/MLS. When the MMM is calculated for all models, the positive bias is up to 50%, and the negative bias reaches up to -33% (Fig. 8c). When models with an RMSE>10 are discarded from the MMM, the negative bias is largely unchanged at -32%, but the positive bias is reduced, and reaches up to 40% (Fig. 8f). These results broadly agree with models evaluated as part of ACCMIP (Young et al., 2013), and phase 5 of the Coupled Model Intercomparison Project (CMIP5) (Eyring et al., 2013). The**se** ~~ACCMIP~~ models used the same ozone precursor emissions as for CCMI**. The ACCMIP models** ~~and~~ 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 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.

The standard deviation on the MMM is up to 11.3 DU when calculated for all models (Fig. 8b), and reduces to a maximum of 9.5 DU when calculated for only the "RMSE<10" models (Fig. 8e). **The variability between models is largest at northern midlatitudes, and in the continental outflow region off the west coast of Africa.**

3.1 Tropospheric ozone in SOCOLv3.1

Figure 2 compares annual-mean tropospheric column ozone as simulated by SOCOLv3.0 and 3.1 with observations derived from OMI/MLS. Although SOCOLv3.0 captures the spatial distribution of tropospheric ozone fairly well in a qualitative sense, i.e. elevated ozone in the Northern Hemisphere and a minimum over the tropical Western Pacific (Fig. 2a), it overestimates tropospheric column ozone between 60° N–40° S by up to 30 DU – approximately a factor of 2 (Fig. 2c). The improved treatment of ozone sink processes in SOCOLv3.1 means that tropospheric ozone columns are reduced **regionally**~~globally~~ by up to 8 DU compared with SOCOLv3.0 (Figs. 2d-e). Individual sensitivity tests (not shown) indicate that this is due mostly to the inclusion of heterogeneous N_2O_5 hydrolysis on tropospheric aerosol.

Both SOCOLv3.0 and 3.1 show a small negative bias in tropospheric ozone over the Southern Ocean. This was also visible in the SOCOLv3.0 and TES comparison presented by Revell et al. (2015). Recent work **by Luhar et al. (2017)** has indicated that the Wesely (1989) dry deposition scheme overestimates the observed ozone deposition velocity by a factor of 2-4 in the Southern Ocean, where SSTs are low and chemical reactions are slow ~~(Luhar et al., 2017)~~. Further upgrades to the model's deposition scheme may therefore improve comparisons of simulated and observed tropospheric ozone in cold oceanic regions.

The global-mean tropospheric ozone column in SOCOLv3.1 is 36.4 DU (Fig. 2d), which is still at the upper end of the range of the CCMI models (Fig. 6), but comparable to other models such as ACCESS (36.3 DU), EMAC-L47 (37.3 DU) and MRI-ESMr1 (35.7 DU). Despite the improvements to SOCOLv3.1, a large bias in tropospheric ozone of approximately 20 DU compared with OMI/MLS remains (Fig. 2f). **The bias maximises over continental regions in the Northern Hemisphere, and over Southeast Asia.**

3.2 GP~~aussian~~ emulation and sensitivity analysis in SOCOLv3.1

To understand the drivers of th~~eis~~ remaining **tropospheric ozone** bias **in SOCOLv3.1**, we constructed a G**P**~~aussian~~ emulator from the 90 SOCOLv3.1 "training" simulations (Section 2.5). Tropospheric ozone predicted by the emulator is compared with SOCOLv3.1 test simulations in Figure 3. In all ~~four~~ geographical regions shown, the **goodness of fit**~~correlation~~ between emulated and simulated tropospheric ozone is high ($R^2 \geq 0.85$) **and the points fall mostly along the 1:1 line**, indicating that the emulator performs well in these regions. The point with the largest simulated tropospheric ozone column corresponds to a simulation in which two ozone loss processes, HNO_3 washout and ozone dry deposition, were set to zero and large scalings (4.00 and 3.54) were applied to the ozone precursors NO_x and CH_4, respectively, following the Latin hypercube design (Fig. 1). The emulator underestimates tropospheric ozone for this point in all ~~four~~ regions **examined**, indicating that it may not be well constrained at the extreme ends of the parameter uncertainty space – noting however, that within the uncertainty range this point agrees with the 1:1 line in all regions except the Southern Ocean.

Figure 4 displays the sensitivity of global-mean tropospheric ozone to each parameter**, obtained by averaging over all other parameters,** ~~assuming all other parameters are held constant.~~ **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)**. In our calculations a uniform sampling distribution was applied when generating the Latin hypercube, which means that in 25% of our training simulations the NO_x (and CH_4, CO and NMVOC) scaling factors are less than one, while in the other 75% of simulations they are larger than one.

To test whether the emulator may be biased due to the sampling distribution used, we calculated tropospheric column ozone as a function of NO_x and CO+NMVOCs using the gradients in Fig. 4a and c. Assuming a uniform sampling distribution between 0 and 4, as per the Latin hypercube design used here, the sensitivity indices for NO_x and CO+NMVOCs are 0.68 and 0.32, respectively. If we assume a piecewise uniform distribution, so that 50% of the points are between 0 and 1 and 50% are between 1 and 4, the sensitivity indices are 0.72 for NO_x and 0.28 for CO+NMVOCs. That is, the differences are negligible, implying that the type of sampling distribution used doesn't bias the result. However, given the NO_x saturation effect above one (Fig. 4a), if we assume a uniform distribution between 0 and 2 instead of 0 and 4, the NO_x sensitivity index increases to 0.86, while the CO index decreases to 0.14. This shows the importance of selecting an appropriate range for the

parameter uncertainty space. However, the conclusions of our emulator analysis – that ozone precursors are the dominant driver of tropospheric ozone variability – remain unchanged.

Figure 5 shows the percentage of variance that each parameter contributes to in each geographic region, either jointly or alone. In all ~~four~~ regions **examined**, ozone precursors – CH_4, NO_x, CO and NMVOCs – account for **more than 90**~~91–94~~% of the variance in tropospheric column ozone. In other words, changing these ozone source input parameters has a far larger impact on tropospheric ozone abundances than changing ozone sink parameters does, and this applies to both polluted regions (Europe, the United States and Asia) and relatively pristine environments (the Southern Ocean). NO_x emissions are generally the dominant driver of variability (in the European region they are approximately equal to the contribution from CH_4, Fig. 5a). **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).**

In all regions, joint~~Joint~~ interactions between NO_x, CH_4 and CO+NMVOCs play a relatively minor role compared with the individual influences of these species. Although updating SOCOLv3.1 with regards to N_2O_5 hydrolysis, HNO_3 washout, LUTs and ozone dry deposition results in a reduction in tropospheric ozone of **up to** 8 DU **regionally** (Fig. 2e), as drivers of tropospheric ozone variability in SOCOLv3.1 they are insignificant compared with ozone precursors. However, we cannot discount the possibility that it is not the ozone precursor emissions themselves that are responsible for SOCOLv3's tropospheric ozone bias, but rather the way in which the emissions are handled by the model; this is **considered**~~discussed~~ further in the Discussion and conclusions.

3.3 Tropospheric ozone in the CCMI models

We now consider SOCOL's tropospheric ozone bias in the context of the CCMI models. Figure 6 illustrates the diversity in simulated tropospheric ozone amongst the CCMI models. Despite most of the models using ozone precursor emissions following the REF-C1 recommendations (Section 2.1), they simulate vastly different representations of tropospheric ozone. A few of the models are closely related, as discussed by Morgenstern et al. (2017); for example the CESM1 models, WACCM and CAM4-chem, are essentially the same model in terms of tropospheric ozone. They differ only in the height of the model lid, which is 140 km for WACCM and 40 km for CAM4-Chem. ACCESS and NIWA-UKCA can also be considered the same model for the REF-C1 experiment; although a coupled ocean was used for most of NIWA-UKCA's CCMI simulations, for the REF-C1 experiment they used the same prescribed sea surface conditions (temperature and ice coverage) as ACCESS. Differences between ACCESS and NIWA-UKCA in the REF-C1 simulation, therefore, are likely related to issues with the different compilers used which may induce small differences in stochastic physics and tropospheric age of air (Dietmüller et al., 2018). The EMAC L47 and L90 models are also very similar; both have a model lid at 0.01 hPa (∼80 km), but they differ in the number of model levels between the surface and 0.01 hPa (47 and 90, respectively). They also use different time steps. ~~Interestingly, EMAC-L90 simulates a better representation of tropospheric column ozone than EMAC-L47, despite the fact that EMAC-L90 has three fewer model levels between the surface and 300 hPa than EMAC-L47 and a longer time step. The difference in tropospheric column ozone between~~

~~the two models likely results from the increased vertical resolution around the tropopause in EMAC-L90, which has 11 levels between 300–100 hPa compared with 7 in EMAC-L47, meaning that EMAC-L90 better simulates stratosphere-troposphere exchange.~~

Figure 7 shows the difference in tropospheric ozone between each of the CCMI models and OMI/MLS, and the root-mean-square error (RMSE) for the model-OMI/MLS difference. Alongside Fig. 6, Fig. 7 indicates clear outlying models in terms of tropospheric ozone. UMUKCA-UCAM simulates the smallest amount of tropospheric ozone (14.9 DU in the global mean, Fig. 6o), however it only contains one NMVOC (formaldehyde) and does not 'lump' NMVOCs together in the way that many other CCMs do. This means that additional NMVOC source gases are not considered by substituting with represented species, such as e.g. in SOCOLv3, whereby additional NMVOCs are included in the form of CO. Of the CCMI models, SOCOLv3.0 simulates the largest global-mean tropospheric ozone column, of 40.2 DU (Fig. 6a). In ULAQ-CCM, the zonal bands of large ozone abundances at northern and southern midlatitudes are related to the model's coarse horizontal resolution (5.6° × 5.6°), which affects surface fluxes and tropospheric transport (Orbe et al., 2018). Interestingly, EMAC-L90 simulates a better representation of tropospheric column ozone than EMAC-L47, despite the fact that EMAC-L90 has three fewer model levels between the surface and 300 hPa than EMAC-L47 and a longer time step. The difference in tropospheric column ozone between the two models likely results from the increased vertical resolution around the tropopause in EMAC-L90, which has 11 levels between 300–100 hPa compared with 7 in EMAC-L47, meaning that EMAC-L90 better simulates stratosphere-troposphere exchange.

Figure 8 shows multi-model means (MMM) and standard deviations. The MMM in Fig. 8a was calculated for all models, while the MMM in Fig. 8d was calculated only for models with a RMSE less than 10, as indicated in Fig. 7 – i.e., all models except SOCOLv3.0, ACCESS CCM, EMAC-L47, ULAQ-CCM and UMUKCA-UCAM. The CCMI models simulate a global-mean tropospheric ozone abundance of 31.1 DU (Fig. 8a), and 30.2 DU (Fig. 8d), depending on the MMM definition applied. Both global-mean MMMs are close to the OMI/MLS global mean of 28.6 DU (Fig. 2b), however the MMMs differ markedly from OMI/MLS in terms of the global tropospheric ozone distribution.

Compared to OMI/MLS, the models overestimate tropospheric column ozone almost everywhere between 60° N–60° S (the region where OMI/MLS data are available), regardless of the MMM definition. The exception is at southern mid-latitudes, where the models underestimate tropospheric ozone compared to OMI/MLS. When the MMM is calculated for all models, the positive bias is up to 50%, and the negative bias reaches up to -33% (Fig. 8c). When models with an RMSE>10 are discarded from the MMM, the negative bias is largely unchanged at -32%, but the positive bias is reduced, and reaches up to 40% (Fig. 8f).

These results broadly agree with models evaluated as part of ACCMIP (Young et al., 2013), and phase 5 of the Coupled Model Intercomparison Project (CMIP5) (Eyring et al., 2013), which used the same ozone precursor emissions as for CCMI. 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.

The standard deviation on the MMM is up to 11.3 DU when calculated for all models (Fig. 8b), and reduces to a maximum of 9.5 DU when calculated for only the "RMSE<10" models (Fig. 8e). The variability between models is largest at northern midlatitudes, and in the continental outflow region off the west coast of Africa.

4 Discussion and conclusions

Despite using the ozone precursor emissions recommended for CCMI, SOCOLv3.0 simulates the largest global-mean tropospheric ozone abundance of all the CCMI models (Fig. 6), and exhibits a bias of ∼30 DU **regionally** compared with OMI/MLS observations (Fig. 2a). The CCMI MMM is biased high in the Northern Hemisphere and low in the Southern Hemisphere compared with OMI/MLS (Fig. 8c and f), consistent with previous studies ~~relying on the same emissions inventories~~ **(ACCMIP and CMIP5)**. **Although ACCMIP, CMIP5 and CCMI all used the same emissions inventories, it is nevertheless interesting that they all produced very similar global-mean 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.**

We have developed a new model version, SOCOLv3.1, which includes an upgraded treatment of tropospheric ozone sink processes. This results in a reduction in tropospheric ozone of up to 8 DU (Fig. 2e), which is mostly due to the inclusion of N_2O_5 hydrolysis on tropospheric aerosol. SOCOLv3.1 still exhibits a positive bias in tropospheric column relative to OMI/MLS (particularly in the Northern Hemisphere), but simulates tropospheric column ozone amounts that are much more comparable with the other CCMI models. **Reducing SOCOL's tropospheric ozone bias is expected to lead to improvements in the simulated abundance of species which are oxidised by the hydroxyl radical, such as CO and CH_4, since ozone is the primary source of OH. Revell et al. (2015) showed that CO in SOCOLv3 was up to 40 ppbv too low in the Northern Hemisphere compared with observations from TES, due to the tropospheric ozone bias. In SOCOLv3.1, the Northern Hemisphere CO bias is reduced by approximately a factor of 2 (not shown).**

We have quantified the contribution to tropospheric ozone variance in SOCOLv3.1 from 9 model forcings/parametrizations using G**P**~~aussian process~~ emulation and sensitivity analysis. By switching off the coupling between chemistry and radiation in the emulator experiments, we aimed to limit dynamical and meteorological variability. We did not consider stratosphere-troposphere exchange in our emulator experiments. Staehelin et al. (2017) showed that SOCOLv3.0's ozone burden due to stratospheric influx, when calculated from ozone origin tracers as described by Garny et al. (2011) and Revell et al. (2015), is close to the multi-model mean values from the ACCMIP and ACCENT ensembles. Therefore, STE is unlikely to be a major driver of SOCOLv3's tropospheric ozone bias. **To the best of our knowledge, this is the first time that GP emulation has been applied to global tropospheric ozone modelling. By selecting a relatively small number of model forcings/parametrizations and focussing largely on tropospheric ozone chemistry we aim to demonstrate the utility of the**

methodology, however it could also be extended to explore the variability in tropospheric ozone due to meteorological parameters.

Our G**P**~~aussian process~~ emulation experiments and sensitivity analysis illustrate that the ozone precursors NO_x, CH_4, CO and NMVOCs are responsible for more than 90% of the variance in tropospheric column ozone in the improved model version, SOCOLv3.1. While CH_4 is prescribed as a surface mixing ratio, the other ozone precursors are specified from emissions inventories. Collating emissions inventories is challenging as they are typically compiled using a bottom-up approach. Anthropogenic emissions must rely on accurate reporting, while for biogenic emissions there are no reporting requirements. Furthermore, emissions are generally prescribed in global models as monthly means, and thus do not reflect diurnal or weekly variability (Young et al., 2018). Hassler et al. (2016) identified that current global emissions inventories do not capture trends in the NO_x/CO ratio, and previous multi-model studies have also identified potential deficiencies with the inventories (Young et al., 2013; Parrish et al., 2014). Jena et al. (2015) and Zhong et al. (2016) showed that different NO_x emissions inventories can significantly alter simulated tropospheric ozone.

However, it may not be the emissions used for CCMI themselves that are incorrect, but rather problems in how they are handled in global models. Given the coarse grid sizes necessary to run a global model and still retain computational efficiency, resolution – horizontal, vertical and temporal – is likely important for simulating tropospheric ozone, especially in polluted regions where very large emissions in an urban environment may be spread over a model grid cell spanning thousands of square kilometers. In global models, polluted air coming from a point source is considered to be well-mixed throughout a large grid cell, which would generally lead to more efficient ozone production (Young et al., 2018). Horizontal and vertical resolution are difficult to test in an emulator sensitivity study as presented here, however by examining the CCMI models collectively (Morgenstern et al., 2017), we can derive some insights. For example, we note that GEOSCCM, HadGEM3-ES and the CESM1 models (CAM4Chem and WACCM), which simulate the smallest RMSEs relative to OMI/MLS (Fig. 7d,e,j,k), have fairly high horizontal resolution relative to other CCMs, of $2^\circ \times 2^\circ$, $1.875^\circ \times 1.25^\circ$ and $1.9^\circ \times 2.5^\circ$ degrees, respectively. Of the models analysed in this study, HadGEM3-ES also has the largest number of levels in the troposphere (48). Similarly, tropospheric ozone in the EMAC model with 90 levels (EMAC-L90) compares better with observations than the 47 level version (EMAC-L47) (Fig. 7h,i), which may be due to a more realistic simulation of the ozone gradient across the tropopause (Section 3.3).

SOCOLv3.0 uses T42 horizontal resolution (approx. $2.8^\circ \times 2.8^\circ$), which is also used by CCSRNIES MIROC 3.2 and EMAC. With 16 vertical levels, SOCOLv3.0 has the smallest number of vertical levels in the troposphere out of all the models analysed here, except CCSRNIES MIROC3.2, which has 15. CCSRNIES-MIROC3.2, CNRM-CM5-3 and CMAM do not include any NMVOCs, while SOCOLv3.0 includes only 2 NMVOCs – isoprene and formaldehyde. Models with complex NMVOC schemes tend to simulate tropospheric ozone favourably compared to OMI/MLS, such as the CESM1 models, with 19 NMVOCs, and GEOSCCM, with 13 explicit NMVOCs.

Another respect in which SOCOLv3.0 is an outlier amongst the CCMI models is its chemical time step of two hours. The other models analysed in this study have chemical time steps ranging from 6 minutes (CCSRNIES-MIROC3.2) to one hour (the models based on the UK Met Office Unified Model, i.e. HadGEM3-ES, NIWA-UKCA, ACCESS and UMUKCA-UCAM). In

a sensitivity test, SOCOLv3.0's chemical time step was reduced to 15 minutes, which reduced the ozone burden in polluted urban areas by approximately 5 DU (not shown). To test how SOCOL responds to prescribing a surface mixing ratio of NO_x rather than an emissions flux, we performed a further sensitivity simulation where surface NO_2 mixing ratios from the CESM1 WACCM REF-C1 simulation were prescribed instead of NO_x emissions. This also resulted in a reduction of tropospheric ozone of up to 5 DU. In reality there is likely no single solution for reducing SOCOLv3.0's excessive tropospheric ozone bias, however assuming that the prescribed emissions are correct, then increasing the model's spatial and temporal resolution within the bounds of computational efficiency will likely reduce the bias.

We have shown the importance of ozone precursor emissions for simulating the tropospheric ozone budget with SOCOLv3.1. This is in line with the findings of Revell et al. (2015), who analysed three SOCOLv3.0 simulations for the period 1960-2100: REF-C2 (based on RCP 6.0), SEN-C2-fEmis (NO_x, CO and NMVOC emissions fixed at constant 1960 levels) and SEN-C2-fEmis-fCH_4 (Similar to SEN-C2-fEmis but with surface methane concentrations also fixed at constant 1960 levels). They showed that future global ozone abundances are governed largely by changes in methane and NO_x, with methane causing an increase in tropospheric ozone that is approximately one-third of that caused by NO_x. Future work should investigate how tropospheric ozone evolves in future under the various CCMI sensitivity scenarios in all CCMI models.

Finally, phase 6 of the Coupled Model Intercomparison Project (CMIP6) will use the emissions data set described by Hoesly et al. (2018). In this data set, year 2000 NO_x emissions are $\sim$20% larger than the emissions used for CCMI (Lamarque et al., 2010). Therefore, simulated ozone biases by the current generation of CCMs will likely be amplified in CMIP6.

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.**

Data availability. The **CCM** data used here (except the CESM1 data) are held at the Centre for Environmental Data Analysis (CEDA, http://data.ceda.ac.uk/badc/wcrp-ccmi/data/CCMI-1/). CESM1 WACCM and CESM1 CAM4-chem data were downloaded from http://www.earthsystemgrid.org. For instructions for access to both archives see http://blogs.reading.ac.uk/ccmi/badc-data-access. GEOSCCM data were provided directly by L. Oman to replace the GEOSCCM data currently held in the CEDA archive. SOCOLv3.1 data are available by contacting L. Revell. **Matrices for training and testing the GP emulator are in the supplementary material.**

Competing interests. The authors declare no competing interests.

Acknowledgements. We acknowledge the modeling groups for making their simulations available for this analysis, the joint WCRP SPARC/IGAC Chemistry-Climate Model Initiative (CCMI) for organizing and coordinating the model data analysis activity, and the British Atmospheric Data Centre (BADC) for collecting and archiving the CCMI model output. The EMAC simulations were performed at the German Climate Computing Centre (DKRZ) through support from the Bundesministerium für Bildung und Forschung (BMBF). DKRZ and its scientific steering committee are gratefully acknowledged for providing the HPC and data archiving resources for this consortial project ESCiMo (Earth System Chemistry integrated Modelling). We acknowledge the UK Met Office for use of the MetUM. This research was partially supported by the NZ Government's Strategic Science Investment Fund (SSIF) through the NIWA programme CACV. OM acknowledges funding by the New Zealand Royal Society Marsden Fund (grant 12-NIW-006). The authors wish to acknowledge the contribution of NeSI high-performance computing facilities to the results of this research. New Zealand's national facilities are provided by the New Zealand eScience Infrastructure (NeSI) and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation and Employment's Research Infrastructure programme (https://www.nesi.org.nz). FT was supported by SNSF grant number 20F121_138017. ACCESS-CCM runs were supported by Australian Research Council's Centre of Excellence for Climate System Science (CE110001028), the Australian Government's National Computational Merit Allocation Scheme (q90) and Australian Antarctic science grant program (FoRCES 4012). The HadGEM3-ES simulations from the Met Office were supported by the Joint UK BEIS/Defra Met Office Hadley Centre Climate Programme (GA01101) and the European Commission's 7th Framework Programme StratoClim project (grant agreement 603557). CCSRNIES research was supported by the Environment Research and Technology Development Fund (2-1303 and 2-1709) of the Ministry of the Environment, Japan, and computations were performed on NEC-SX9/A(ECO) computers at the CGER, NIES. UMUKCA-UCAM model integrations were performed using the ARCHER UK National Supercomputing Service and MONSooN system, a collaborative facility supplied under the Joint Weather and Climate Research Programme, which is a strategic partnership between the UK Met Office and the Natural Environment Research Council. **The authors thank Edmund Ryan and one anonymous reviewer for their helpful and constructive comments.**

References

Arfeuille, F., Luo, B. P., Heckendorn, P., Weisenstein, D., Sheng, J. X., Rozanov, E., Schraner, M., Brönnimann, S., Thomason, L. W., and Peter, T.: Modeling the stratospheric warming following the Mt. Pinatubo eruption: Uncertainties in aerosol extinctions, Atmos. Chem. Phys., 13, 11,221-11,234, doi:10.5194/acp-13-11221-2013, 2013.

Auvray, M., and Bey, I.: Long-range transport to Europe: Seasonal variations and implications for the European ozone budget, J. Geophys. Res.: Atmos., 110, D11, doi:10.1029/2004JD005503, 2005.

Carslaw, K.S., Lee, L.A., Reddington, C.L., Pringle, K.J., Rap, A., Forster, P.M., Mann, G.W., Spracklen, D.V., Woodhouse, M.T., Regayre, L.A., and Pierce, J.R.: Large contribution of natural aerosols to uncertainty in indirect forcing, Nature, 503, 67-71, doi:10.1038/nature12674, 2013.

Chang, J.S., Brost, R.A., Isaksen, I.S.A., Madronich, S., Middleton, P., Stockwell, W.R., and Walcek, C.J.: A three-dimensional Eulerian acid deposition model: Physical concepts and formulation, J. Geophys. Res.: Atmos., 92, 14681-14700, doi:10.1029/JD092iD12p14681, 1987.

Denman, K.L., Brasseur, G., Chidthaisong, A., Ciais, P., Cox, P.M., Dickinson, R.E., Hauglustaine, D., Heinze, C., Holland, E., Jacob, D., Lohmann, U., Ramachandran, S., da Silva Dias, P. L., Wofsy, S.C., and Zhang, X.: Couplings between changes in the climate system and biogeochemistry, Chapter 7 in Climate Change 2007: the Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M., and Miller, H.L., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

Dietmüller, S., Eichinger, R., Garny, H., Birner, T., Boenisch, H., Pitari, G., Mancini, E., Visioni, D., Stenke, A., Revell, L., Rozanov, E., Plummer, D.A., Scinocca, J., Jöckel, P., Oman, L., Deushi, M., Kiyotaka, S., Kinnison, D.E., Garcia, R., Morgenstern, O., Zeng, G., Stone, K.A., and Schofield, R.: Quantifying the effect of mixing on the mean age of air in CCMVal-2 and CCMI-1 models, Atmos. Chem. Phys., 18, 6699-6720, 10.5194/acp-18-6699-2018, 2018.

Egorova, T.A., Rozanov, E.V., Zubov, V.A., and Karol, I. L.: Model for investigating ozone trends (MEZON), Izv. Atmos. Ocean. Phys., 39, 277–292, 2003.

Ehhalt, D., Prather, M., Dentener, F., Derwent, R., Dlugokencky, E., Holland, E., Isaksen, I., Katima, J., Kirchhoff, V., Matson, P., Midgley, P., and Wang, M.: Atmospheric chemistry and greenhouse gases, Chapter 4 in Climate Change 2001: The Scientific Basis. Contribution ofWorking Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Houghton, J.T., Ding, Y., Griggs, D.J., Noguer, M., van der Linden, P.J., Dai, X., Maskell, K., and Johnson, C.A, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2001.

ETH-PMOD: Swiss Federal Institute of Technology Zurich and the Physical-Meteorology Observatory Davos, Data, Part of the Chemistry-Climate Model Initiative (CCMI-1) Project Database, NCAS British Atmospheric Data Centre, available at: http://catalogue.ceda.ac.uk/uuid/1005d2c25d14483aa66a5f4a7f50fcf0 (28 September 2017), 2015.

Evans, M.J. and Jacob, D.J.: Impact of new laboratory studies of N_2O_5 hydrolysis on global model budgets of tropospheric nitrogen oxides, ozone and OH, Geophys. Res. Lett., 32, L09813, doi:10.1029/2005GL022469, 2005.

Eyring, V., Arblaster, J.M., Cionni, I., Sedláček, J., Perlwitz, J., Young, P.J., Bekki, S., Bergmann, D., Cameron-Smith, P., Collins, W.J., Faluvegi, G., Gottschaldt, K.D., Horowitz, L. W., Kinnison, D.E., Lamarque, J.F., Marsh, D.R., Saint-Martin, D., Shindell, D.T., Sudo, K., Szopa, S., and Watanabe, S.: Long-term ozone changes and associated climate impacts in CMIP5 simulations, J. Geophys. Res., 118, 5029-5060, 10.1002/jgrd.50316, 2013a.

Eyring, V., Lamarque, J.-F., Hess, P., Arfeuille, F., Bowman, K., Chipperfield, M.P., Duncan, B., Fiore, A., Gettelman, A., Giorgetta, M.A., Granier, C., Hegglin, M., Kinnison, D., Kunze, M., Langematz, U., Luo, B., Martin, R., Matthes, K., Newman, P.A., Peter, T., Robock, A., Ryerson, T., Saiz-Lopez, A., Salawitch, R., Schultz, M., Shepherd, T.G., Shindell, D., Staehelin, J., Tegtmeier, S., Thomason, L., Tilmes, S., Vernier, J.-P., Waugh, D.W., and Young, P.J.: Overview of IGAC/SPARC Chemistry-Climate Model Initiative (CCMI) Community Simulations in Support of Upcoming Ozone and Climate Assessments, SPARC Newsletter no. 40, ISSN 1245-4680, 48–66, 2013b.

Garny, H., Grewe, V., Dameris, M., Bodeker, G.E., and Stenke, A.: Attribution of ozone changes to dynamical and chemical processes in CCMs and CTMs, Geosci. Model Dev., 4, 271-286, 10.5194/gmd-4-271-2011, 2011.

Gaudel, A., Cooper, O.R., Ancellet G., Barret, B., Boynard, A., Burrows, J.P., et al.: Tropospheric Ozone Assessment Report: Present-day distribution and trends of tropospheric ozone relevant to climate and global atmospheric chemistry model evaluation, Elem. Sci. Anth., 6(39), 10.1525/elementa.291, 2018.

Greenslade, J.W., Alexander, S.P., Schofield, R., Fisher, J.A., and Klekociuk, A.K.: Stratospheric ozone intrusion events and their impacts on tropospheric ozone in the Southern Hemisphere, Atmos. Chem. Phys., 17, 10269-10290, 10.5194/acp-17-10269-2017, 2017.

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P. I., and Geron, C., Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature), Atmos. Chem. Phys., 6, 3181-3210, https://doi.org/10.5194/acp-6-3181-2006, 2006.

Hassler, B., McDonald, B.C., Frost, G.J., Borbon, A., Carslaw, D.C., Civerolo, K., Granier, C., Monks, P.S., Monks, S., Parrish, D.D., Pollack, I.B., Rosenlof, K.H., Ryerson, T.B., von Schneidemesser, E. and Trainer, M.: Analysis of long-term observations of NO_x and CO in megacities and application to constraining emissions inventories, Geophys. Res. Lett., 43(18), 9920-9930, doi:10.1002/2016GL069894, 2016.

Hauglustaine, D.A., Granier, C., Brasseur, G., and Megie G., The importance of atmospheric chemistry in the calculation of radiative forcing on the climate system, J. Geophys. Res., 99, 1173– 1186, 1994.

Hoesly, R.M., Smith, S.J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., Seibert, J.J., Vu, L., Andres, R.J., Bolt, R.M., Bond, T.C., Dawidowski, L., Kholod, N., Kurokawa, J. I., Li, M., Liu, L., Lu, Z., Moura, M.C.P., O'Rourke, P.R., and Zhang, Q.: Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the Community Emissions Data System (CEDS), Geosci. Model Dev., 11, 369-408, 10.5194/gmd-11-369-2018, 2018.

Jena, C., Ghude, S.D., Beig, G., Chate, D.M., Kumar, R., Pfister, G.G., Lal, D.M., Surendran, D.E., Fadnavis, S., and van der A, R.J.: Inter-comparison of different NO_x inventories and associated variation in simulated surface ozone in Indian region, Atmos. Env., 117, 61-73, https://doi.org/10.1016/j.atmosenv.2015.06.057, 2015.

Jöckel, P., Tost, H., Pozzer, A., Kunze, M., Kirner, O., Brenninkmeijer, C. A. M., Brinkop, S., Cai, D. S., Dyroff, C., Eckstein, J., Frank, F., Garny, H., Gottschaldt, K.-D., Graf, P., Grewe, V., Kerkweg, A., Kern, B., Matthes, S., Mertens, M., Meul, S., Neumaier, M., Nützel, M., Oberländer-Hayn, S., Ruhnke, R., Runde, T., Sander, R., Scharffe, D., and Zahn, A.: Earth System Chemistry integrated Modelling (ESCiMo) with the Modular Earth Submodel System (MESSy) version 2.51, Geosci. Model Dev., 9, 1153-1200, https://doi.org/10.5194/gmd-9-1153-2016, 2016.

Johnson, J.S., Cui, Z., Lee, L.A., Gosling, J.P., Blyth, A.M., and Carslaw, K.S.: Evaluating uncertainty in convective cloud microphysics using statistical emulation, J. Adv. Model. Earth Syst., 7, 162-187, doi:10.1002/2014MS000383, 2015.

Kerkweg, A., Buchholz, J., Ganzeveld, L., Pozzer, A., Tost, H., and Jöckel, P., Technical Note: An implementation of the dry removal processes DRY DEPosition and SEDImentation in the Modular Earth Submodel System (MESSy), Atmos. Chem. Phys., 6, 4617-4632, https://doi.org/10.5194/acp-6-4617-2006, 2006.

Köpke, P., Hess, M., Schult, I., and Shettle, E.P., Global Aerosol Data Set, Max-Planck-Institut für Meteorologie, Hamburg, Report No. 243, available at: https://www.mpimet.mpg.de/fileadmin/publikationen/Reports/MPI-Report_243.pdf (last access: 25 September 2017), 1997.

Lamarque, J.F., Bond, T.C., Eyring, V., Granier, C., Heil, A., Klimont, Z., Lee, D., Liousse, C., Mieville, A., Owen, B., Schultz, M.G., Shindell, D., Smith, S.J., Stehfest, E., Van Aardenne, J., Cooper, O.R., Kainuma, M., Mahowald, N., McConnell, J.R., Naik, V., Riahi, K., and van Vuuren, D.P.: Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application, Atmos. Chem. Phys., 10, 7017-7039, doi:10.5194/acp-10-7017-2010, 2010.

Lee, L.A., Carslaw, K.S., Pringle, K.J., Mann, G.W., and Spracklen, D.V.: Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters, Atmos. Chem. Phys., 11, 12253-12273, doi:10.5194/acp-11-12253-2011, 2011.

Lee., L.A., Carslaw, K.S., Pringle, K.J., and Mann, G.W.: Mapping the uncertainty in global CCN using emulation, Atmos. Chem. Phys., 12, 9739-9751, doi:10.5194/acp-12-9739-2012, 2012.

Le Gratiet, L., Marelli, S., and Sudret, B.: Metamodel-Based Sensitivity Analysis: Polynomial Chaos Expansions and Gaussian Processes. In: Ghanem R., Higdon D., Owhadi H. (eds), Handbook of Uncertainty Quantification, Springer, https://doi.org/10.1007/978-3-319-12385-1_38, 2017.

Lin, M., Horowitz, L.W., Oltmans, S.J., Fiore, A.M., and Fan, S.: Tropospheric ozone trends at Mauna Loa Observatory tied to decadal climate variability, Nature Geosci., 7, 136-143, doi:10.1038/ngeo2066, 2014.

Loeppky, J.L., Sacks, J., and Welch, W. J.: Choosing the sample size of a computer experiment: A Practical Guide, Technometrics, 51, 366-376, doi:10.1198/TECH.2009.08040, 2009.

Luhar, A.K., Galbally, I.E., Woodhouse, M.T., and Thatcher, M.: An improved parameterisation of ozone dry deposition to the ocean and its impact in a global climate–chemistry model, Atmos. Chem. Phys., 17, 3749-3767, 10.5194/acp-17-3749-2017, 2017.

Luo, B.: Stratospheric aerosol data for use in CCMI models, available at: ftp://iacftp.ethz.ch/pub_read/luo/ccmi/ (last access: 29 August 2018), 2013.

Masui, T., Matsumoto, K., Hijioka, Y., Kinoshita, T., Nozawa, T., Ishiwatari, S., Kato, E., Shukla, P.R., Yamagata, Y., and Kainuma, M.: An emission pathway for stabilization at 6 Wm-2 radiative forcing, Climatic Change, 109, 59, doi:10.1007/s10584-011-0150-5, 2011.

McKay, M., Conover, W., and Beckman, R.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics, 21, 239–245, doi:10.2307/1268522, 1979.

Morgenstern, O., Hegglin, M.I., Rozanov, E., O'Connor, F.M., Abraham, N.L., Akiyoshi, H., Archibald, A.T., Bekki, S., Butchart, N., Chipperfield, M.P., Deushi, M., Dhomse, S.S., Garcia, R.R., Hardiman, S.C., Horowitz, L.W., Jöckel, P., Josse, B., Kinnison, D., Lin, M., Mancini, E., Manyin, M.E., Marchand, M., Marécal, V., Michou, M., Oman, L.D., Pitari, G., Plummer, D.A., Revell, L.E., Saint-Martin, D., Schofield, R., Stenke, A., Stone, K., Sudo, K., Tanaka, T.Y., Tilmes, S., Yamashita, Y., Yoshida, K., and Zeng, G.: Review of the global models used within phase 1 of the Chemistry-Climate Model Initiative (CCMI), Geosci. Model Dev., 10, 639-671, doi:10.5194/gmd-10-639-2017, 2017.

Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestvedt, J., Huang, J., Koch, D., Lamarque, J.-F., Lee, D., Mendoza, B., Nakajima, T., Robock, A., Stephens, G., Takemura, T., and Zhang, H.: Anthropogenic and natural radiative forcing, Chapter 8 in Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Nicely, J.M., Hanisco, T.F., Deushi, M., Duncan, B.N., Haslerud, A.S., Jöckel, P., Josse, B., Kinnison, D.E., Klekociuk, A., Manyin, M.E., Morgenstern, O., Murray, L.T., Myhre, G., Oman, L.D., Pitari, G., Pozzer, A., Revell, L.E., Rozanov, E., Salawitch, R.J., Stenke, A.,

Stone, K., Strahan, S., Tilmes, S., Tost, H., Westervelt, D.M., and Zeng, G., Hydroxyl radical intercomparison between chemistry-climate model and chemical transport model simulations for CCMI-1, in prep., 2018.

O'Hagan, A.: Bayesian analysis of computer code outputs: A tutorial, Reliability Engineering & System Safety, 91, 1290-1300, doi:http://dx.doi.org/10.1016/j.ress.2005.11.025, 2006.

Orbe, C., Yang, H., Waugh, D.W., Zeng, G., Morgenstern, O., Kinnison, D.E., Lamarque, J.F., Tilmes, S., Plummer, D.A., Scinocca, J.F., Josse, B., Marecal, V., Jöckel, P., Oman, L.D., Strahan, S.E., Deushi, M., Tanaka, T.Y., Yoshida, K., Akiyoshi, H., Yamashita, Y., Stenke, A., Revell, L., Sukhodolov, T., Rozanov, E., Pitari, G., Visioni, D., Stone, K.A., Schofield, R., and Banerjee, A.: Large-scale tropospheric transport in the Chemistry–Climate Model Initiative (CCMI) simulations, Atmos. Chem. Phys., 18, 7217-7235, 10.5194/acp-18-7217-2018, 2018.

Parrish, D.D., Lamarque, J.F., Naik, V., Horowitz, L., Shindell, D.T., Staehelin, J., Derwent, R., Cooper, O.R., Tanimoto, H., Volz-Thomas, A., Gilge, S., Scheel, H.E., Steinbacher, M., and Fröhlich, M.: Long-term changes in lower tropospheric baseline ozone concentrations: Comparing chemistry-climate models and observations at northern midlatitudes, J. Geophys. Res.: Atmos., 119, 5719-5736, doi:10.1002/2013JD021435, 2014.

Pöschl, U., von Kuhlmann, R., Poisson, N., and Crutzen, P.J.: Development and Intercomparison of Condensed Isoprene Oxidation Mechanisms for Global Atmospheric Modeling, J. Atmos. Chem., 37, 29-52, doi:10.1023/a:1006391009798, 2000.

Rayner, N.A., Parker, D.E., Horton, E.B., Folland, C.K., Alexander, L.V., Rowell, D.P., Kent, E.C., and Kaplan, A.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, J. Geophys. Res.: Atmos., 108, D14, doi:10.1029/2002JD002670, 2003.

Revell, L.E., Tummon, F., Stenke, A., Sukhodolov, T., Coulon, A., Rozanov, E., Garny, H., Grewe, V., and Peter, T.: Drivers of the tropospheric ozone budget throughout the 21st century under the medium-high climate scenario RCP 6.0, Atmos. Chem. Phys., 15, 5887-5902, doi:10.5194/acp-15-5887-2015, 2015.

Riahi, K., Rao, S., Krey, V., Cho, C., Chirkov, V., Fischer, G., Kindermann, G., Nakicenovic, N., and Rafaj, P.: RCP 8.5—-A scenario of comparatively high greenhouse gas emissions, Climatic Change, 109, 33, doi:10.1007/s10584-011-0149-y, 2011.

Roeckner, E., Bäuml, G., Bonaventura, L., Brokopf, R., Esch, M., Giorgetta, M., Hagemann, S., Kirchner, I., Kornblueh, L., Manzini, E., Rhodin, A., Schlese, U., Schulzweida, U., and Tompkins, A.: The atmospheric general circulation model ECHAM 5. Part I: Model description, Max-Planck-Institut für Meteorologie, Hamburg, Report No. 349, available at: http://www.mpimet.mpg.de/fileadmin/publikationen/Reports/max_scirep_349.pdf (last access: 28 September 2017), 2003.

Rozanov, E., Schlesinger, M.E., Zubov, V., Yang, F., and Andronova, N.G.: The UIUC three-dimensional stratospheric chemical transport model: Description and evaluation of the simulated source gases and ozone, J. Geophys. Res., 104, 11755- 011781, 1999.

Ryan, E., Wild, O., Voulgarakis, A., and Lee, L.: Fast sensitivity analysis methods for computationally expensive models with multi-dimensional output, Geosci. Model Dev., 11, 3131-3146, https://doi.org/10.5194/gmd-11-3131-2018, 2018.

Sillman, S., Logan, J.A., and Wofsy, S.C.: The sensitivity of ozone to nitrogen oxides and hydrocarbons in regional ozone episodes, J. Geophys. Res.: Atmos., 95, 1837-1851, doi:10.1029/JD095iD02p01837, 1990.

Silva, R.A., West, J.J., Zhang, Y., Anenberg, S.C., Lamarque, J.-F., Shindell, D.T., Collins, W.J., Dalsoren, S., Faluvegi, G., Folberth, G., Horowitz, L.W., Nagashima, T., Naik, V., Rumbold, S., Skeie, R., Sudo, K., Takemura, T., Bergmann, D., Cameron-Smith, P., Cionni, I., Doherty, R.M., Eyring, V., Josse, B., MacKenzie, I.A., Plummer, D., Righi, M., Stevenson, D. S., Strode, S., Szopa, S., and Zeng, G.: Global premature mortality due to anthropogenic outdoor air pollution and the contribution of past climate change, Environ. Res. Lett., 8, 034005, doi:10.1088/1748-9326/8/3/034005, 2013.

Silva, R.A., West, J.J., Lamarque, J.-F., Shindell, D.T., Collins, W.J., Faluvegi, G., Folberth, G.A., Horowitz, L.W., Nagashima, T., Naik, V., Rumbold, S.T., Sudo, K., Takemura, T., Bergmann, D., Cameron-Smith, P., Doherty, R.M., Josse, B., MacKenzie, I.A., Stevenson, D.S., and Zeng, G.: Future global mortality from changes in air pollution attributable to climate change, Nature Clim. Change, 7, 647-651, doi:10.1038/nclimate3354, 2017.

SPARC CCMVal: SPARC Report on the Evaluation of Chemistry-Climate Models, edited by: Eyring, V., Shepherd, T. G., and Waugh, D. W., SPARC Report No. 5, WCRP-132, WMO/TD-No. 1526, available at: http://www.sparc-climate.org/publications/sparc-reports/sparc-report-no-5/ (last access: 31 May 2018), 2010.

Staehelin, J., Tummon, F., Revell, L.E., Stenke, A., and Peter, T.: Tropospheric ozone at northern mid-latitudes: modeled and measured long-term changes, Atmosphere, 8, 163, doi:10.3390/atmos8090163, 2017.

Stenke, A., Schraner, M., Rozanov, E., Egorova, T., Luo, B., and Peter, T.: The SOCOL version 3.0 chemistry-climate model: description, evaluation, and implications from an advanced transport algorithm, Geosci. Model Dev., 6, 1407-1427, doi:10.5194/gmd-6-1407-2013, 2013.

Stevenson, D.S., Dentener, F.J., Schultz, M.G., Ellingsen, K., van Noije, T.P.C., Wild, O., Zeng, G., Amann, M., Atherton, C.S., Bell, N., Bergmann, D.J., Bey, I., Butler, T., Cofala, J., Collins, W.J., Derwent, R.G., Doherty, R.M., Drevet, J., Eskes, H.J., Fiore, A.M., Gauss, M., Hauglustaine, D.A., Horowitz, L.W., Isaksen, I.S.A., Krol, M.C., Lamarque, J.-F., Lawrence, M.G., Montanaro, V., Müller, J.-F., Pitari, G., Prather, M.J., Pyle, J.A., Rast, S., Rodriguez, J.M., Sanderson, M.G., Savage, N.H., Shindell, D.T., Strahan, S.E., Sudo, K., and Szopa, S.: Multimodel ensemble simulations of present-day and near-future tropospheric ozone, J. Geophys. Res.: Atmos., 111, doi:10.1029/2005JD006338, 2006.

Stevenson, D.S., Young, P.J., Naik, V., Lamarque, J.F., Shindell, D.T., Voulgarakis, A., Skeie, R.B., Dalsøren, S.B., Myhre, G., Berntsen, T.K., Folberth, G.A., Rumbold, S.T., Collins, W.J., MacKenzie, I.A., Doherty, R.M., Zeng, G., van Noije, T.P.C., Strunk, A., Bergmann, D., Cameron-Smith, P., Plummer, D.A., Strode, S.A., Horowitz, L., Lee, Y.H., Szopa, S., Sudo, K., Nagashima, T., Josse, B., Cionni, I., Righi, M., Eyring, V., Conley, A., Bowman, K.W., Wild, O., and Archibald, A.: Tropospheric ozone changes, radiative forcing and attribution to emissions in the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP), Atmos. Chem. Phys., 13, 3063-3085, doi:10.5194/acp-13-3063-2013, 2013.

Thornton, J.A., Wooldridge, P.J., Cohen, R.C., Martinez, M., Harder, H., Brune, W.H., Williams, E.J., Roberts, J.M., Fehsenfeld, F.C., Hall, S.R., Shetter, R.E., Wert, B.P., and Fried, A.: Ozone production rates as a function of NO_x abundances and HO_x production rates in the Nashville urban plume, J. Geophys. Res.: Atmos., 107, ACH 7-1-ACH 7-17, doi:10.1029/2001JD000932, 2002.

Wesely, M., Parameterization of the surface resistances to gaseous dry deposition in regional-scale numerical models, Atmos. Environ., 23, 1293–1304, 1989.

World Meteorological Organization: Scientific Assessment of Ozone Depletion: 2010, WMO Global Ozone Research and Monitoring Project - Report No. 52, Geneva, Switzerland, 2011.

Young, P.J., Archibald, A.T., Bowman, K.W., Lamarque, J.F., Naik, V., Stevenson, D.S., Tilmes, S., Voulgarakis, A., Wild, O., Bergmann, D., Cameron-Smith, P., Cionni, I., Collins, W.J., Dalsøren, S.B., Doherty, R.M., Eyring, V., Faluvegi, G., Horowitz, L.W., Josse, B., Lee, Y.H., MacKenzie, I.A., Nagashima, T., Plummer, D.A., Righi, M., Rumbold, S.T., Skeie, R.B., Shindell, D.T., Strode, S.A., Sudo, K., Szopa, S., and Zeng, G.: Pre-industrial to end 21st century projections of tropospheric ozone from the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP), Atmos. Chem. Phys., 13, 2063-2090, doi:10.5194/acp-13-2063-2013, 2013.

Young, P.J., Naik, V., Fiore, A.M., Gaudel, A., Guo, J., Lin, M.Y., Neu, J.L., Parrish, D.D., Rieder, H.E., Schnell, J.L., Tilmes, S., Wild, O., Zhang, L., Ziemke, J., Brandt, J., Delcloo, A., Doherty, R.M., Geels, C., Hegglin, M.I., Hu, L., Im, U., Kumar, R., Luhar, A., Murray, L.,

Plummer, D., Rodriguez, J., Saiz-Lopez, A., Schultz, M.G., Woodhouse, M.T., and Zeng, G.: Tropospheric Ozone Assessment Report: Assessment of global-scale model performance for global and regional ozone distributions, variability, and trends, Elem. Sci. Anth., 6(10), doi:10.1525/elementa.265, 2018.

Zhong, M., Saikawa, E., Liu, Y., Naik, V., Horowitz, L.W., Takigawa, M., Zhao, Y., Lin, N.H., and Stone, E.A.: Air quality modeling with WRF-Chem v3.5 in East Asia: sensitivity to emissions and evaluation of simulated air quality, Geosci. Model Dev., 9, 1201-1218, 10.5194/gmd-9-1201-2016, 2016.

Ziemke, J.R., Chandra, S., Duncan, B.N., Froidevaux, L., Bhartia, P.K., Levelt, P.F., and Waters, J.W.: Tropospheric ozone determined from Aura OMI and MLS: Evaluation of measurements and comparison with the Global Modeling Initiative's Chemical Transport Model, J. Geophys. Res.: Atmos., 111, D19, doi:10.1029/2006JD007089, 2006.

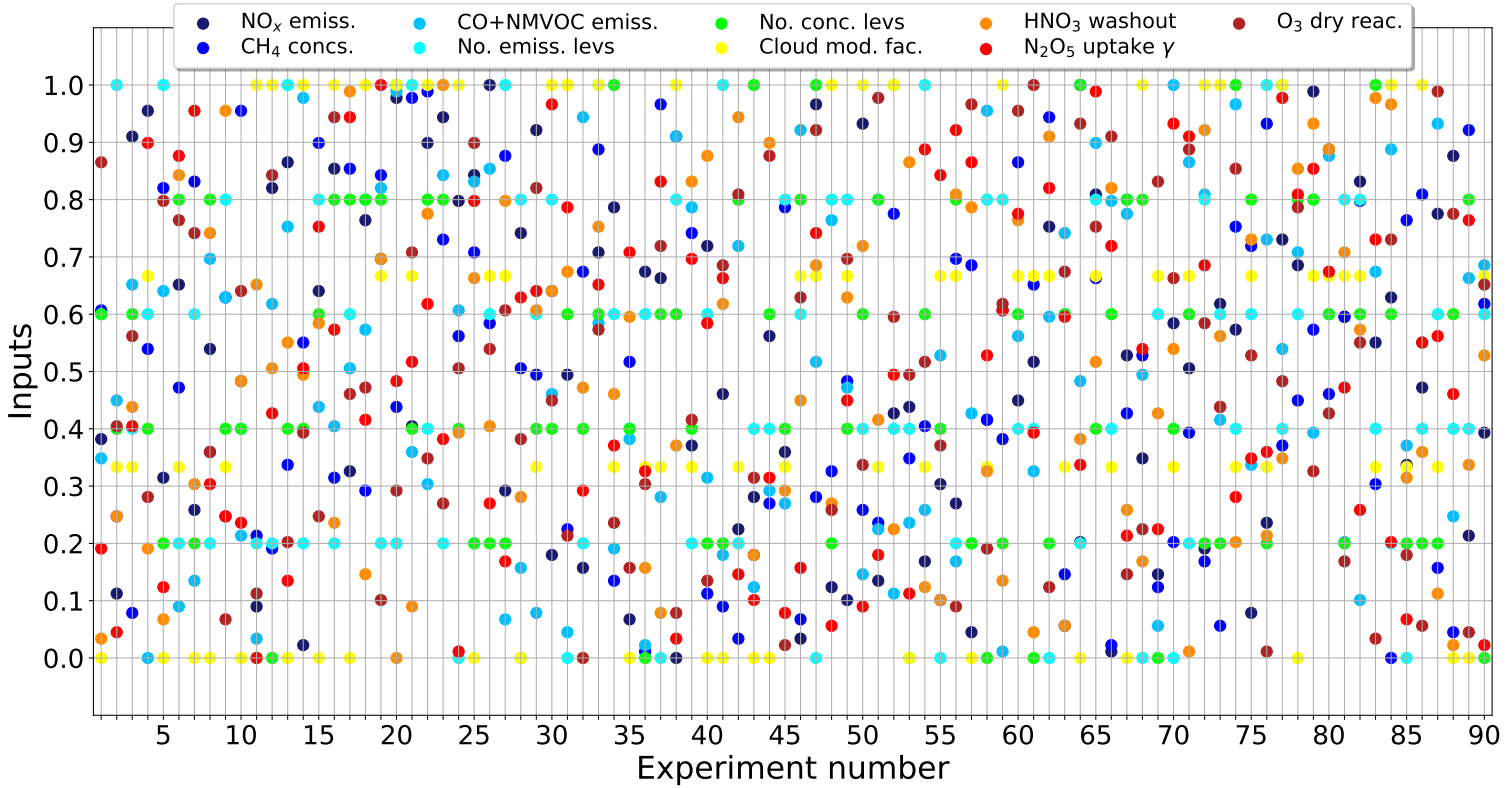


Figure 1. Experimental design for the 90 SOCOLv3.1 simulations performed to train the GP emulator. Each column of dots indicates the scaling applied to each of the 9 variables – see Table 1 for more details. For clarity the ~~N_2O_5 hydrolysis scaling factors have been multiplied by 10 here, and the HNO_3 washout scaling factors have been multiplied by 12~~ **inputs have been scaled between 0 and 1**. See also Supplementary Video 1, which explains this figure in greater detail.

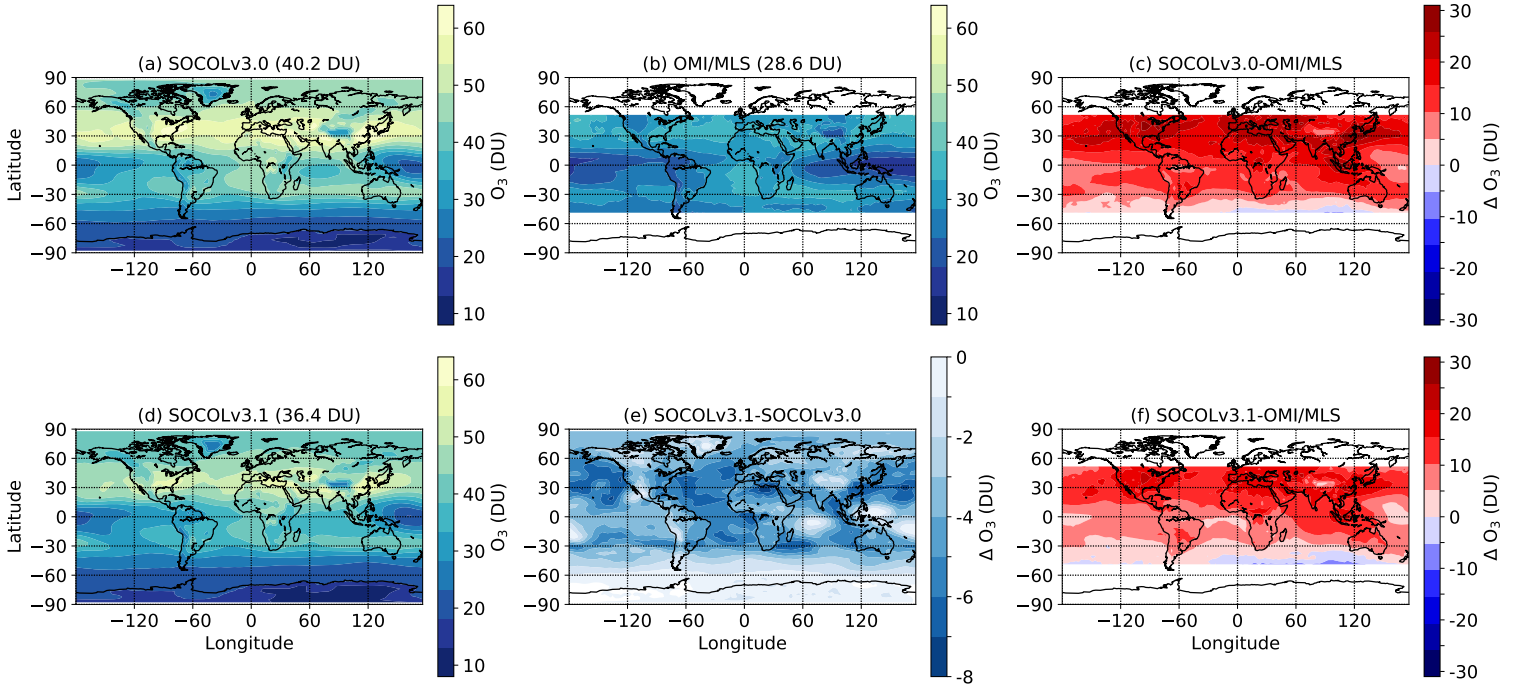


Figure 2. Annual-mean year 2005 tropospheric column for: (a) SOCOLv3.0; (b) OMI/MLS observations; (c) The difference between SOCOLv3.0 and OMI/MLS; (d) SOCOLv3.1; (e) The difference between SOCOLv3.1 and SOCOLv3.0; (f) The difference between SOCOLv3.1 and OMI/MLS. The global-mean tropospheric column ozone amount is indicated in the title for (a), (b) and (d).

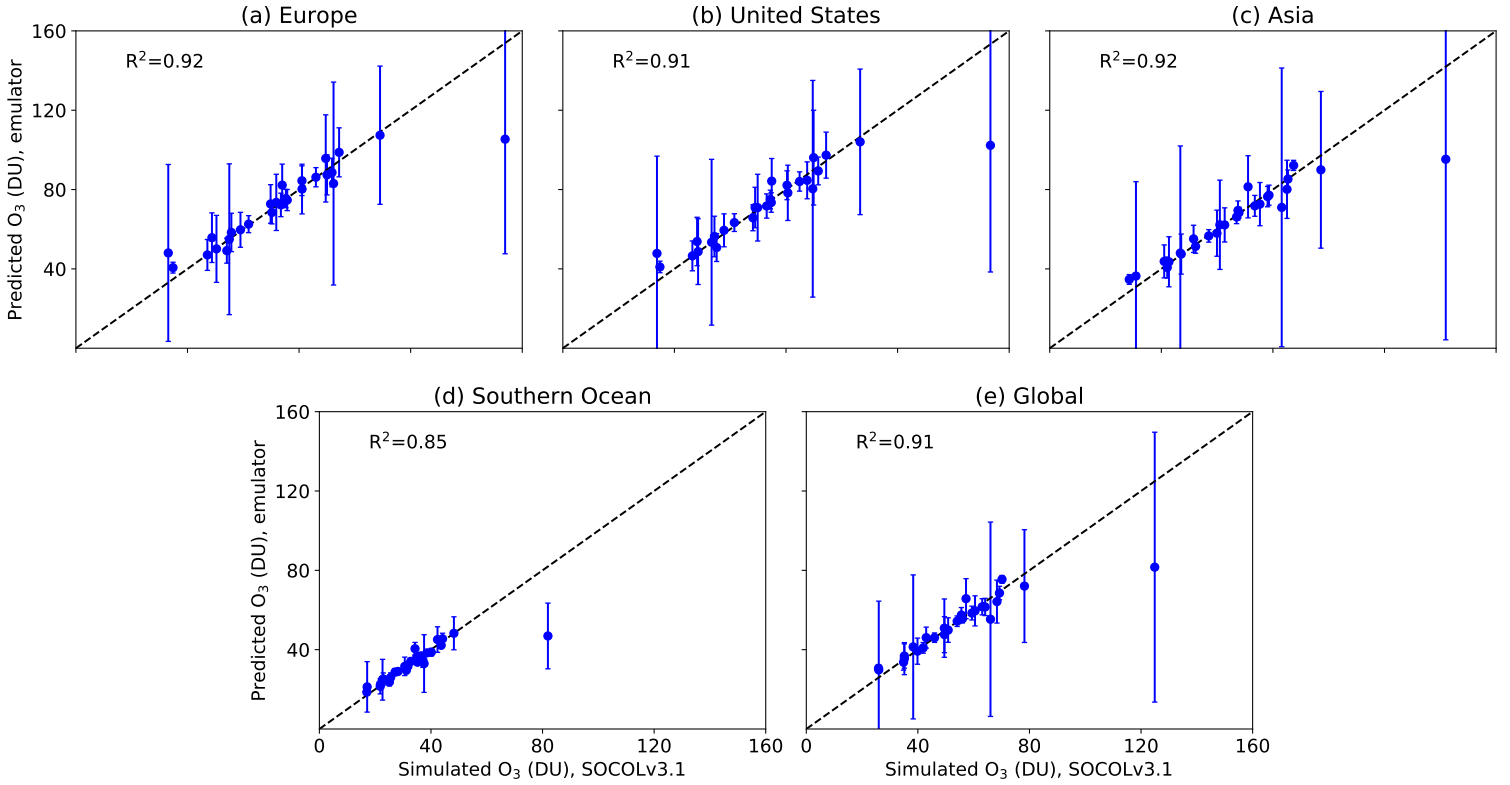


Figure 3. Tropospheric column ozone as predicted by the GP emulator, vs. the amount simulated in SOCOLv3.1 "test" simulations (i.e., the simulations used to validate the emulator). The errorbars indicate the uncertainty on the GP emulator output, and the 1:1 line and coefficient of determination (R^2 value) are also shown. These simulations correspond to running the GP emulator and the simulator (SOCOLv3.1) at each of the 27 validation inputs, for: (a) Europe (37-60° N, 0-42° E); (b) United States (32-52° N, 67-124° W); (c) Asia (6-49° N, 70-146° E), (d) the Southern Ocean (45–60° S, all longitudes); and (e) globally.

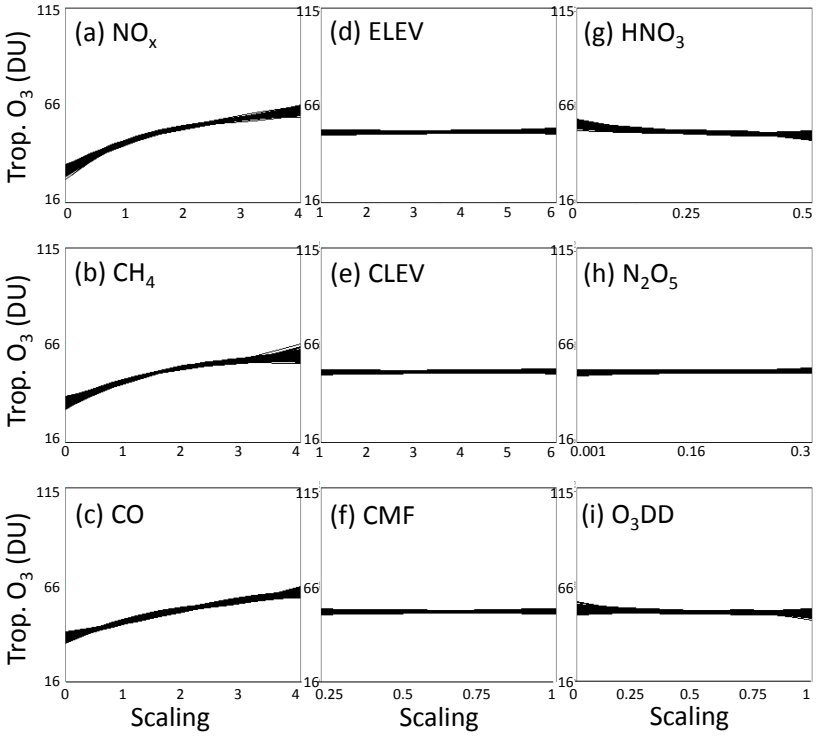


Figure 4. Sensitivity of annual-, global-mean tropospheric column ozone in 2005 to each of the 9 sensitivity forcings/parametrizations listed in Table 1, averaging over the other inputs. The horizontal axis shows the range of scaling factors applied to each variable. Plots for individual regions (Europe, the United States, Asia and the Southern Ocean are in the supplementary material.

.

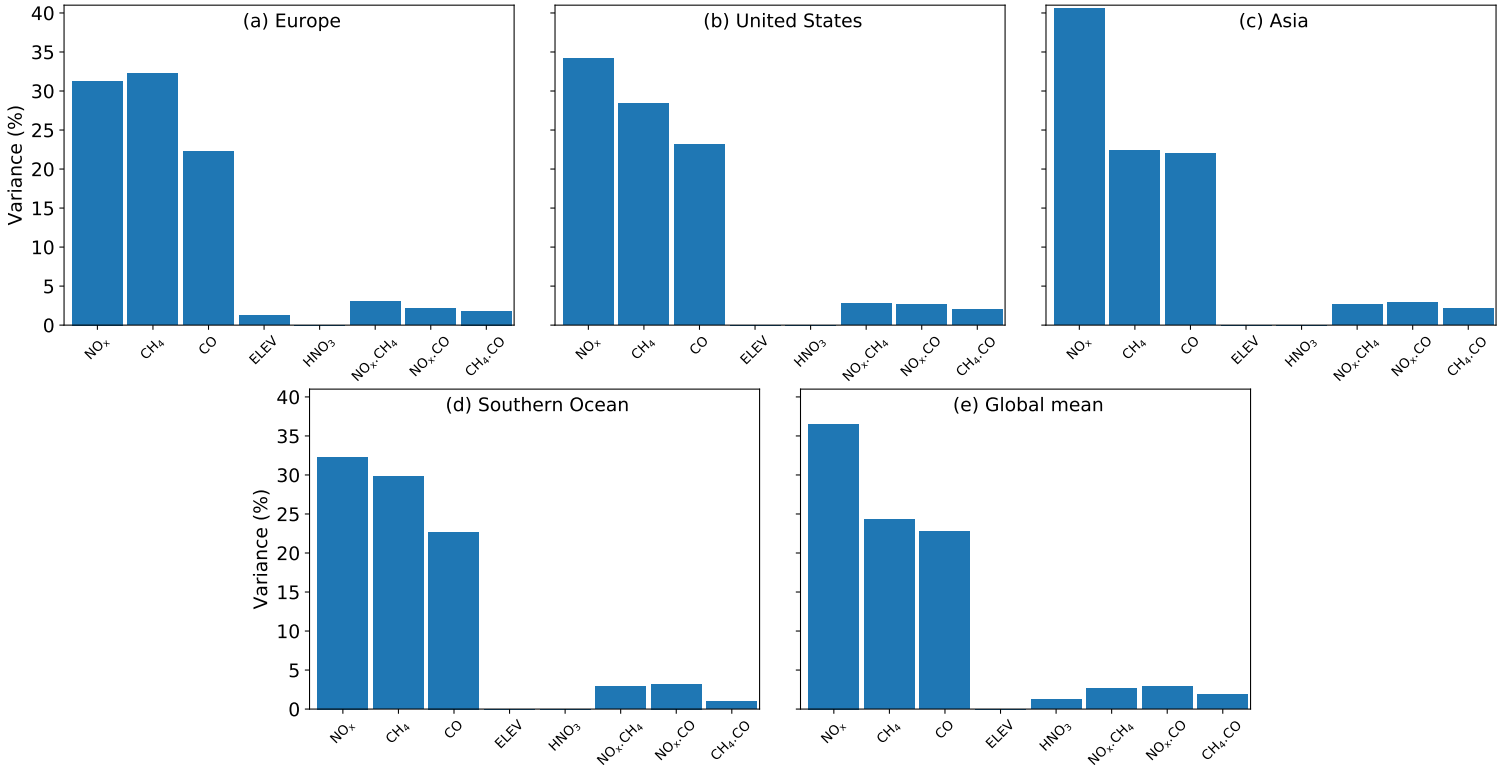


Figure 5. Contributions to variance from the **sensitivity forcings/parametrizations applied (Table 1)~~9 variables listed in Table 1~~, for the same regions shown in Figure 6). For clarity only those which contribute at least 1% are shown. NO_x = NO_x emissions; CH_4 = CH_4 concentrations; CO = CO+NMVOC emissions; ELEV = the number of vertical model levels that NO_x, CO and NMVOC emissions are prescribed on. Joint interactions, indicated by e.g. NO_x.CH_4 are also indicated where these contribute at least 1% to the variance.**

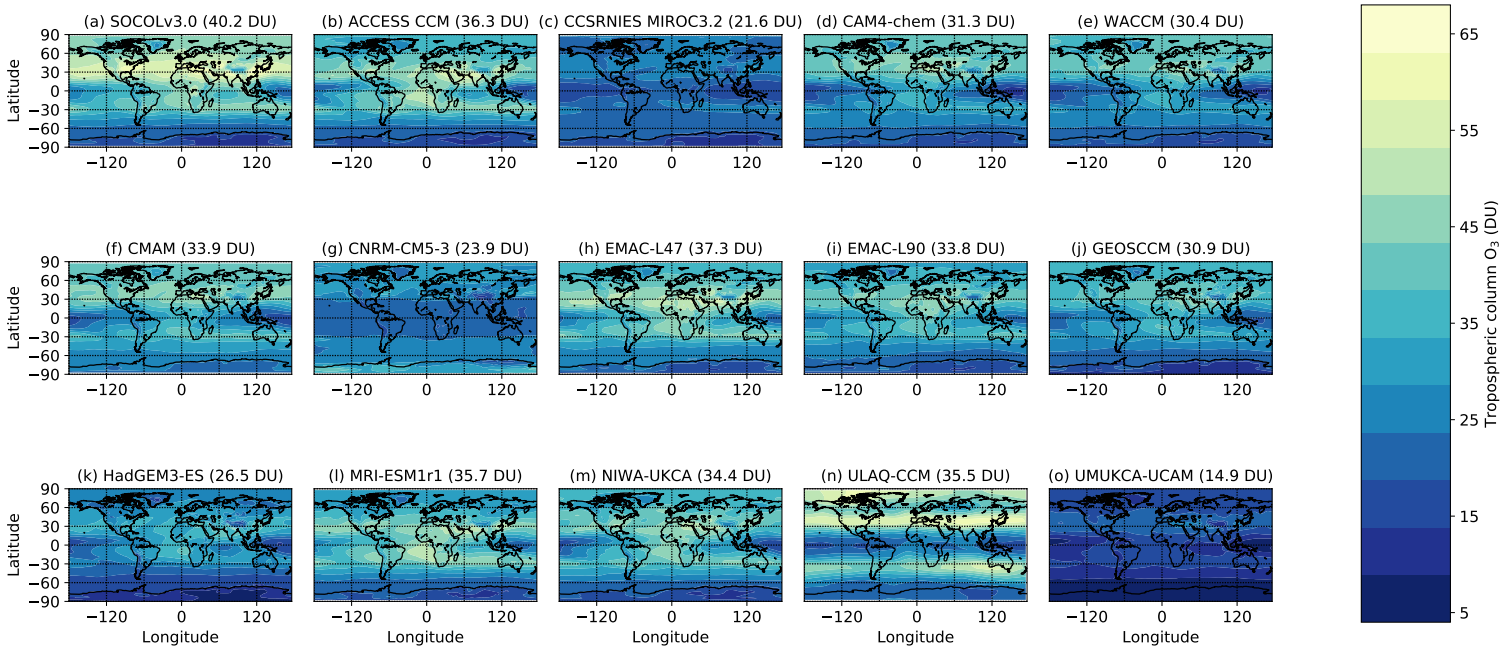


Figure 6. Annual-mean year 2005 tropospheric ozone columns in REF-C1 simulations from CCMI models (calculated relative to the WMO-defined tropopause pressure for each model). The global-mean tropospheric column ozone amount for each model is indicated in the title.

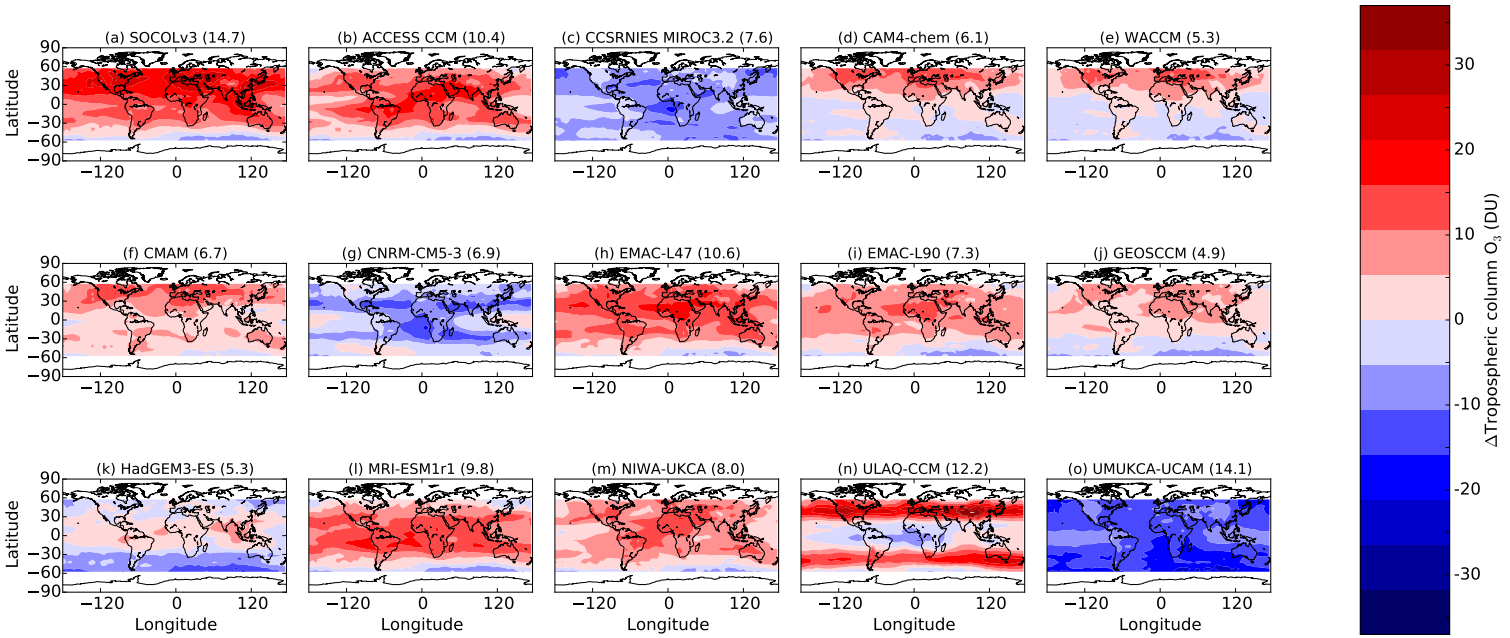


Figure 7. Difference between annual-mean year 2005 tropospheric column ozone in CCMI models compared with OMI/MLS, i.e. model minus OMI/MLS. The root-mean-square error for each model compared with OMI/MLS is indicated in the title.

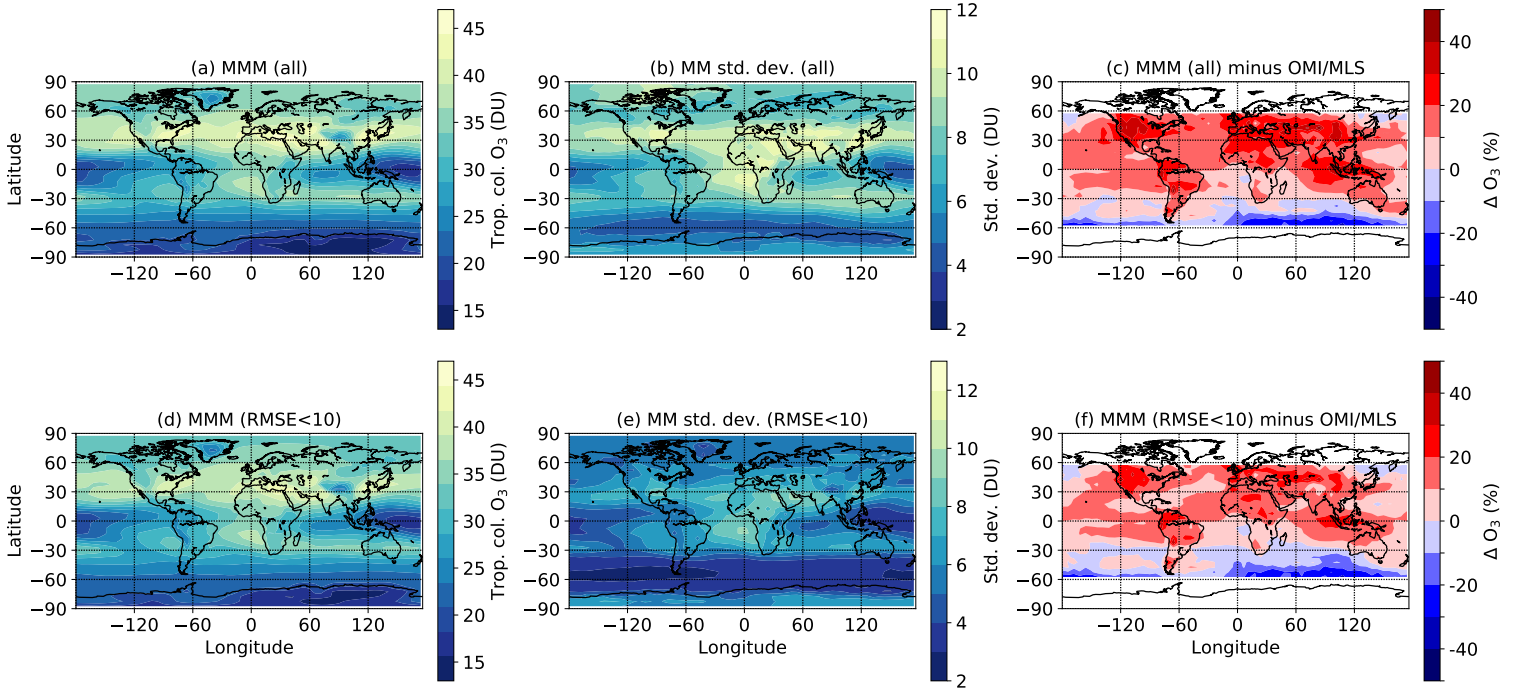


Figure 8. Annual-mean year 2005 tropospheric column ozone. (a) The multi-model mean (MMM) of all CCMI models; (b) multi-model standard deviation for the models shown in (a); (c) percent difference between the MMM in (a) and OMI/MLS (MMM minus OMI/MLS); (d) MMM for a subset of CCMI models – those with a root-mean-square error (RMSE) less than 10 when compared with OMI (see Fig. 7); (e) multi-model standard deviation for the models shown in (d); (f) percent difference between the MMM in (d) and OMI/MLS (MMM minus OMI/MLS).

Table 1. Revised Table 1: Range of the sensitivity forcings/parametrizations. **P** and **L** indicate whether the variable is of relevance to ozone production and/or loss, respectively.

	Minimum	**Maximum**	**Descriptions**
(1) NO_x emissions **(P)**	0	4	The surface NO_x emissions field as a function of latitude and longitude was multiplied by a scaling factor between 0 and 4, to explore the sensitivity of tropospheric ozone to a range of NO_x emissions.
(2) CH_4 concentrations **(P)**	0	4	The global-mean CH_4 mixing ratio was multiplied by a scaling factor between 0 and 4, to explore the sensitivity of tropospheric ozone to a range of CH_4 concentrations.
(3) CO+NMVOC **(P)** emissions	0	4	As for (1), but the scaling factor was applied to CO and NMVOC emissions simultaneously.
(4) ELEV for NO_x and CO+NMVOCs **(P)**	1	6	Emissions were prescribed on the lowermost 1–6 levels (between (the surface and $\sim$2.5 km, to test whether the number of levels is important for tropospheric ozone abundances.
(5) CLEV for CH_4 **(P)**	1	6	CH_4 concentrations were prescribed on the lowermost 1–6 levels (between the surface and $\sim$2.5 km, similar to (4).
(6) CMF **(P+L)**	0.25	1	1 implies clear-sky photolysis, whereas 0 would imply no photolysis. As photolysis rates of 0 do not occur during daytime, we selected a lower bound of 0.25 to represent cloudy sky conditions.
(7) HNO_3 washout **(L)**	0	0.5	To test the sensitivity of tropospheric ozone to HNO_3 removal, we removed between 0–50% of tropospheric gas-phase HNO_3 at each chemical time step.
(8) N_2O_5 hydrolysis **(L)**	0.001	0.3	The probability of N_2O_5 hydrolysis occurring. Since the default is 0.1, we explored the sensitivity of tropospheric ozone to a range from 0.001-0.3.
(9) O_3 dry deposition **(L)**	0	1	A specific reactivity of 0 stands for a nearly non-reactive gas, while 1 stands for a gas similarly reactive to ozone.