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.

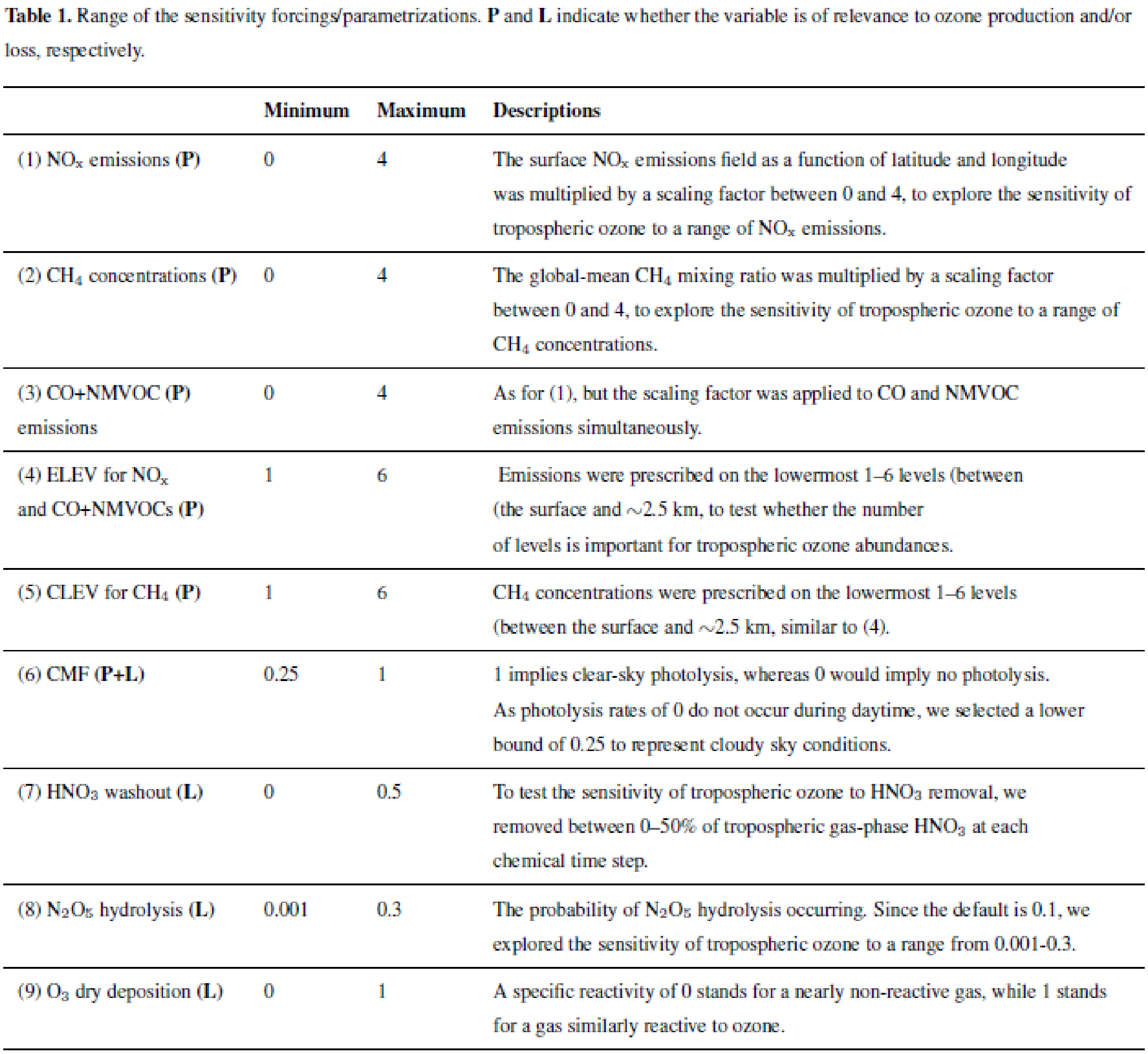

Table 1. Range of the sensitivity forcings/parametrizations. **P** and **L** indicate whether the variable is of relevance to ozone production and/or loss, respectively.

	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.