

## ***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 Univer-

C1

sity 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-model-dev-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.

[2] Page 7, line 24. I feel uncomfortable about you using the words “not necessarily feasible” here. For a sensitivity analysis study, justifying the mins and maxs of your inputs is important because if your range covers values of a particular input that are

C2

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.

[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?

#### 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 N<sub>2</sub>O<sub>5</sub> hydrolysis process. Is this reduction in bias at the cost of an increase in bias for other variables (e.g. CH<sub>4</sub> lifetime) when compared with

C3

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.

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

[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.

[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.

[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”.

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

[7] Page 3, line 30. Oliver Wild’s group at Lancaster University are also using emulators

C4

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-model-dev-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.

[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.

[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.

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

[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. An Maximin LHD is a space filling design. It is an efficient design for sampling from a multi-dimensional 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.

[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?

C5

[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 whole number might be okay but it might not be. You might want to survey the literature a bit and what others have done.

[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.

[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.

[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/GI>

[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).

[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.

C6

[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:  $x_{norm} = (x - x_{min}) / (x_{max} - x_{min})$ . 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.

[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”.

[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).

[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.

[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.

[24] Table 1. Is it accurate to describe all the inputs has “scaling”. E.g. input 4 is not a

C7

scaling since it’s the no. of level.s

[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.

#### Reference

O’Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering & System Safety*, 91(10-11), pp.1290-1300.

Loeppky, J. L., Sacks, J., & Welch, W. J. (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, 51(4), 366-376.

Ryan, E. (2013). *The Limitations and Robustness of Data Assimilation in Terrestrial Ecosystem Modelling* (Doctoral dissertation, University of Sheffield). <http://etheses.whiterose.ac.uk/4293/>

Ryan, E., Wild, O., O’Connor, F., Voulgarakis, A., and Lee, L. (accepted, 2018): Fast sensitivity analysis methods for computationally expensive models with multidimensional output, *Geosci. Model Dev. Discuss.*, <https://doi.org/10.5194/gmd-2017-271>.

Interactive comment on *Atmos. Chem. Phys. Discuss.*, <https://doi.org/10.5194/acp-2018-615>, 2018.

C8