Comment on acp-2021-801

Anonymous Referee #2

Referee comment on "Assessing the potential for simplification in global climate model cloud microphysics" by Ulrike Proske et al., Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2021-801-RC2, 2021

General Comments:

This study makes use of the technique of variance-based sensitivity analysis on an emulated perturbed parameter ensemble of the global aerosol-climate model ECHAM-HAM to understand the impact of perturbing the effect of selected cloud-ice microphysical processes on the models' output. For each process considered, a 'phasing parameter' is implemented to perturb the strength of the generated effect of a process (from 0 to 200%), and the study uses this 'phasing' as a proxy for the effect of process simplification.

This is a novel use of the emulation and sensitivity analysis approach to assess model behaviour under uncertainty. The paper definitely falls within the scope of ACP and EGU, and is written to a high standard. However, there are several points that I believe need clarification (see specific comments below). In particular, I am concerned that the design and sampling for the PPE simulations does not provide the required coverage of the actual phasing effect for a completely robust analysis. The parameters are multiplicative factors but they are not treated as such in the PPE design, so the PPE has a very skewed coverage over the effect of 'phasing out' a process (see specific comment at Line 210-214, below). Because of this, the PPE looks to have very low coverage of training data where the 'phasing out effect' is strongest and the model response is likely to be greatest/more erratic (as the η_i parameters move towards zero), and a much denser coverage of training points where the phasing out effect is weaker ($0.5<\eta_i<1$) or there is over-estimation of a process due to an inaccurate description ($\eta_i>1$). Given the low amount of training information for the emulator where the phasing out is strong, I'm not convinced that the emulator can properly capture this response in any kind of detail. For a more robust conclusion, I would recommend (if possible - this would be a major revision) a re-design of the PPE input combinations to properly cover the parameter space for emulation and provide a more even sampling of the 'phasing out effect' for the sensitivity analysis. Once this and the further issues/comments below are addressed, I would recommend the publication of the manuscript in ACP.

Specific Comments:

- Line 7 (in abstract): 'The response to the phasing of a process **thereby** serves as a proxy for the effect of a simplification'. This sentence is confusing me in two ways. Firstly, what is meant by 'the phasing of a process'? – this is unclear (I realise this might be explained in the paper, but people will read the abstract first, so it's not clear at this point.). Also, the use of the word 'thereby' in this sentence is confusing – it suggests that the information in this sentence follows as a result of the sentence before it, but I don't think it is – it is a separate point with new information about the method/assumptions made. Please remove 'thereby' and re-phrase to clarify.

- Line 13 (in abstract): Is this really a 'new framework'? Statistical emulation and sensitivity analysis have been used in several studies to assess process impacts in complex models of clouds, the atmosphere and the climate, as you have stated in the paper e.g. Line 239-240: 'This approach is similar to Johnson et al. (2015)…,' Please re-phrase.

- Line 42-44: 'Finally, the detail of … increases computational demand and thereby costs or inhibits other advancements such as the move towards higher resolution…'. Do parameterisations at a lower resolution still always hold at a higher resolution? Is it not the case that for a move to higher resolution, we need more detailed representations of processes? – So, to a reasonable extent, doesn't higher resolution and more detailed process descriptions have a dependence? – I don't think they are quite as independent as this sentence is suggesting – please clarify.

- Line 63: '… a simplified model **equifinal to** a more complex model'. What does 'equifinal to' mean here, when the difference is between 2 different models (simplified and complex)? – I'm not sure if this is the same as the definition for equifinality on page 2, where different parts of a model's parameter space lead to a similar observed state? Please clarify.

- Line 68: '…The influence of CMPs has been shown to dominate over that of aerosol schemes…' I'm not sure this has always been the case for a GCM? For example, Regayre at al (2018) [Figure 9] showed that both aerosol and physical atmosphere (cloud-related) model parameters are both important sources of uncertainty in aerosol ERF in the GCM HadGEM3-UKCA. Please update the text here to reflect this.

- Line 89: '…variations in input as well as…': The word 'input' should be plural. Also, should this be '…variations in **independent** input**s**…'? Most global sensitivity analysis techniques (especially variance-based sensitivity analysis) assume independence between

inputs.

 - Line 112: 'By phasing we mean that we vary the effectiveness of a given process, going from using 0 to 200% of a process's effect in the model'. This is quite a difficult concept to understand here in terms of how this can be done – I don't think all processes within a model could be easily 'phased'. What is meant by 'effectiveness'? How is it defined and is this 'effectiveness' the same or does it differ between inputs / processes? I realise that the next section (2) will bring more detail on this, but giving a small (brief) example or a little more detail here could provide a bit more clarity for the reader as a starting point.

 - Line 202 (which also connects the point for Line 112): 'From the response of model output to variations in η_i, we can extract how accurately a process i needs to be represented in the model.' How can you extract this? Does process accuracy actually directly correspond to the effect of 'phasing' in/out a process like this? From the abstract: 'The response to the phasing of a process serves as a proxy for the effect of simplification' – But, is a less accurate / simplified process necessarily going to produce a reduced change in the additional 'delta' component in equation 1? Couldn't a simplified process potentially make that component larger? Or, have any effect on what that value is? I cannot work out if it really is realistic to treat a process in this way. Please give a clearer description as to how/why the phasing feeds through to inference on process accuracy and process simplification.

 - Line204 (connects to the point for Line 202): What is a 'sigmoidal function'? [Will a general reader know?] From google, a 'sigmoid function' has a loose 'S' shape? [like the 'logistic function: $f(x) = 1/(1+\exp(-x))$]. So, it's gradient can be steep or shallow depending where you are on the curve? Hence, how can you know that some detail can easily be left out? This needs more clarity – how this parameter for each process can inform the need for model complexity is a key message from the study, so understanding how to interpret it is very important, yet it seems to be skipped over here. A diagram to help the reader picture what you mean (maybe with several different options as to how the parameter η_i could be interpreted for a process) would be helpful, as it is not clear to me that the statement in this example (lines 203-205) is true, or how the parameter in general will inform us.

 - Lines 210-214: The scaling of the 'η' parameters here treats them as linear factors, but I don't think they are. I think each η_i is a multiplicative factor, and as such, the phasing effect is not varying evenly over the η ranges, with it likely that there will be a much more significant effect on the model behaviour with very small values as an η approaches 0. [I think this is also an aspect of the cause of the large outlier at the very low η_{aggr} in Fig 3?].

This is tricky to explain, but within your range of $0<\eta<2$, $0.5<\eta_i<1$ corresponds to a scaling of the given process by 1 times (1x) to a half times (0.5x), covering a 'phasing reduction' of the process by up to 2 times (2x) smaller than its default effect over a range in η values of 0.5. But, lower down the η range, say $0.01<\eta<0.1$, this covers a more significant reduction of 10 times smaller (0.1x) to 100 times smaller (0.01x) than the default effect, but within a much smaller range on η of size 0.09. Because of this, your

PPE looks to have very low coverage of training data where the 'phasing out effect' is strongest and the model response is likely to be greatest/more erratic (as the η parameters move towards zero), and a much denser coverage of training points where the phasing out effect is relatively weaker ($0.5<\eta_i<1$) or where you consider over-estimation ($1<\eta_i<2$). In fact, designing the training points linearly between 0 and 2 leads to having approx. 50% of simulations with $\eta>1$ for each η parameter – so really concentrating on the parts of the ranges / 4-d parameter space that is to 'imitate an overestimation of a given process due to an inaccurate description' (Line 214). Is that what you intended? As, my understanding is that this area of the space isn't really the focus of the study (to understand how sensitive the model responses are to phasing out processes), so why sample it the most?

I think this is a significant error in your PPE design. And this will also feed through to affect how you sample the phasing effects for the sensitivity analysis (concentrated away from a strong phasing out, and highly focussed on $\eta_i>1$, if sampling uniformly). In most PPE studies, parameters like this are varied on a $\log_{10}$ scale to account for the multiplicative behaviour. However, including zero in your range means a $\log_{10}$ transform is difficult here (as $\log_{10}(0) = -inf$) – it might be better to only vary the parameters down to a small value close to zero (e.g. 0.001) so that a $\log_{10}$ scaling could be used to even out the phasing effect over the η ranges. Given the low amount of training information for the emulator where the phasing out is strong, I'm not convinced that the emulator can properly capture this response in any kind of detail. If possible, for a more robust conclusion, I would recommend a re-design of the PPE input combinations in this way to properly cover the parameter space for emulation, provide a more even sampling of the 'phasing out effect' for the sensitivity analysis, and also provide more detail on how the model response changes as a process is phased out. If this is not possible, please at least acknowledge the assumption that has been made here – that you treat the η parameters as linearly varying factors – and note/describe/discuss here and in the results and discussion section how this is affecting your analysis and results.

- Line 215: The phrase 'sets of simulation input' is unclear. I think you mean 'the set of input parameter combinations (η_1, η_2, η_3, η_4) to be simulated with the model'. Please clarify the text.

- Line 236: 'As kernel, an additive combination of the linear, polynomial, bias and exponential kernel was used (Duvenaud, 2014)' What does 'as kernel' mean? Is this the function that describes the covariance between points in the Gaussian process (GP), and so control the smoothness of the GP response surface? This additive combination seems rather complex – why is this chosen/used?

- Line 237: 'The input data was centred and whitened prior to emulation'. What exactly does that mean? Why is this needed, and how does it affect the emulator / surrogate model? Please give more detail. [There isn't enough detail here for someone to be able to replicate the analysis.]

- Line 245: '1-out validation'. This is an unusual term to describe this approach. Please

change to 'Leave-one-out validation', here and elsewhere.

- Line 246-247: In the brackets, please use the notation as it is in the formula. So, '(with **Y_{sim}** and **Y_{emu}** the output of the ECHAM-HAM simulations and the emulated output respectively, and **V_{emu}** the emulator variance)'

- Lines 257-266: I think this could also result in part from the PPE design and low coverage of large changes in the phasing amount at the low end of the parameter ranges (see comment [L210-214] above), which should also be acknowledged here.

- Line 299 (and in the paragraphs that follow): '…inflicted by the **inhibition** of the other three processes.' I don't think 'inhibition' is the right word to use here – what do you mean by a process' 'inhibition'? Do you mean it has very little effect? Or just the process of 'phasing out'? – it's not clear. [When I google it, I don't find a relevant meaning for this context.] Please re-phrase and remove this term 'inhibition' throughout the manuscript.

- Line 305: 'The shape of the model response to the gradual phasing of the processes holds additional information: while the generated model response is mostly gradual, for low η_{aggr} the response is more abrupt.' – This is also, in part, the effect of the uneven distribution of the 'phasing effect' over the parameter range (see comment [L210-214] above), which should be acknowledged here.

- Line 310: 'As can be seen from Fig. 1 it (aggregation) is the only process that generates snow flakes. Accretion and riming need the snow flakes to be able to act upon them.' Does this mean that there is a dependence between the phasing parameters here? Is this a strong dependence? i.e. without quite a high value of η_{aggr}, you cannot have (it isn't realistic to have) a high value of η_{accr} or η_{rim}? – or can you have (is it realistic to have) a high value of η_{rim} when η_{aggr} is pretty small (just not zero)? If it is a strong dependence, then this would invalidate the assumptions of the variance-based sensitivity analysis (Sections 2.5,3.3) which assumes independence between inputs for the breakdown of the variance into its component parts.

- Figure 6 / Figure 7 captions: It isn't correct to label the panels in these figures as 'correlation matrix'. They are not plots of correlations – they are 2-d projections of the sampling of the 4-d response surface for a given output. Please amend the captions.

- Line 353-355: 'The observed sensitivities are different from what Bacer et al. (2021) find…. In our analysis, the influence of aggregation dwarfs that of accretion in terms of sensitivity indices as well as for the process rates...' Are you comparing 'like-for-like' here? Or are you seeing a larger effect for aggregation because you vary the process more – to phasing it out completely? Please clarify.

- Line 363-364: 'This was excluded from the sensitivity analysis as only the input parameter space with $\eta_{aggr} \geq 0.5$ was taken into consideration...' Is it not more appropriate to consider the sensitivity analysis (SA) for a range of η_i that doesn't go all the way down to zero anyway? Is it not the case that processes need to still be accounted for (they still need to be included in the model), but that you are investigating just how detailed or not (phased in or out) that representation needs to be?

Why did you choose 0.5 here? Also, is the focus of the SA in terms of space sampled more on $\eta_i > 1$? Could this be biasing the SA results away from the effect that you really want to consider? (i.e. is it focussing much less on phasing out from the current full complexity at $\eta=1$, and more on the effects of increasing complexity / overestimation of a process?). How might this feed through to affect the inferences and conclusions made?

- Line 388: should this say '...representation of **global annual mean** IWP and LWP, the....'?

- Line 392: '...analysis of grid-point level data is tedious and error prone...' I agree that you have to create a lot of emulators to do grid-point analysis, so reducing the dimensions as you do in this section is definitely advantageous. But, why is the grid-point analysis 'error-prone'? Given the quite high level of uncertainty in emulator prediction from the emulators of the spherical harmonics here (Line 429 and Appendix D Figure D1 part d, where the points are rather widely scattered about the line of equality), is it really fair to infer that this approach is less error-prone?

Also, is it possible to generate a plot (map) of how the sensitivities of parameters vary over the globe via the method used here, like you might with a grid-point level analysis? (I'm not suggesting you necessarily do this here – I'm just wondering if it is possible to do this...)

- Line 409: Should the sum term here have '(l)' at the end, as it does on line 406?

- Line 427 (end): should this say '**LWP and** CDNC **are** dominated by....'?

- Line 443: '...as with the 48th member the computational constraint was too tight for the emulator' I don't understand what you mean by 'computational constraint was too tight'. Please clarify.

- Figures 10 and 11: Why do you only show the total sensitivity index? Isn't the First

order effect more informative? With the total sensitivity index, if there are large interactions, or if the model output is quite noisy which can sometimes induce interaction in the indices that is not real, this index on its own can give a potentially skewed/false impression of the true sensitivity. To conclude robustly on the sensitivity, really you should present both indices. Can the first order index also be shown in Figures 10 and 11?

Technical Corrections:

 - Line 9 (in abstract): For clarity, change '…on snow influences mostly the liquid phase.' to '…on snow **has most influence on** the liquid phase.'

 - Line 30: Change 'More complexity has also its downsides:…' to 'More complexity **also has** its downsides:…'

 - Line 65: Change '…in face of…' to '…in **the** face of…'

 - Line 136: Missing bracket at the end of the sentence: '…see Neubauer et al. (2019)**)**.'

 - Line 186: '…in-cloud ICNC concentration and snow…' Remove the work 'concentration' here, as it is already in the acronym ICNC.

 - Line190: Change '…Table A1 in the Appendix.' to '…Table A1 in Appendix A.'

 - Line 227: 'Each simulation included a 3 months spin up…'. Change to either 'Each simulation included 3 months **of** spin up…' or 'Each simulation included a 3 month spin up…'

 - Line 230: 'Fig. 1e' should be 'Fig. **2**e'. Please correct.

 - Line 256: Change '…bounds of the emulator crossing…' to '…bounds o**n** the emulator **predictions** crossing…'

- Figure 3 caption, line 1: Remove 'according to Bastos and O'Hagan (2009).' from the first sentence. (Bastos and O'Hagan do not validate this emulator…)

 - Line 349: Missing word: '…allows **us** to quantify…'.

 - Line 362: Missing word: '…space and not due **to** the threshold behaviour…'

 - Line 382: Remove the word 'as' before 'e.g.'.

 - End of figure 9 caption: Change 'truncated' to 'truncate'.

 - End of figure 10 caption: Change 'is missing here' to 'are missing here'.

 - Line 608: Check the details of the reference 'Hawker et al (2021a)' – This paper has now been accepted in ACP and should be published soon, so the exact reference might be available?