Review of “Identifying climate model structural inconsistencies allows for tight constraint of aerosol radiative forcing”

Regayre al. (2023), submitted to Atmospheric Chemistry and Physics

March 29, 2023

This analysis proposes a new interpretation of model structural inconsistencies and their effect on aerosol radiative forcing constraint. The authors sampled uncertainty in 37 model parameters related to aerosols, clouds, and radiation in a perturbed parameter ensemble (PPE) of the UK Earth System Model and evaluated 1 million emulated model variants against satellite-derived observations over several cloudy regions. They argue that incorporating observations associated with model internal inconsistencies weakens the forcing constraint because they require a wider range of parameter values to accommodate conflicting information. They propose an estimated aerosol forcing range based on the maximum feasible constraint using a structurally imperfect model and the chosen observations. I was particularly interested by the first steps of their search for potential structural model inconsistencies. These steps allow to rule out variable constraints when the emulator uncertainty is too high, when only extreme model behaviour aligns with them or when the “pairwise comparison” reveals an impossibility to match them consistently with a larger set of variables (because of model trade-offs and structural inconsistencies). As discussed in the specific comments, I am less convinced by the definition of the “optimal constraint” on aerosol forcing, based on the tightest constraint achievable. Overall, I think this paper tackles an important issue in climate model tuning : the difficulty to design a relevant multi-variate metric for the model evaluation and the need for structural systematic bias identifications across a large number of model variants.

Specific comments :

I think this paper is very relevant and I do not have major issues to point out, but I have listed some remarks that I would like to discuss with the authors.

1. My main issue is the lack of some figures evaluating the Gaussian Process (GP) prediction skills. Indeed, you emulate a lot of variables, with both regional and global means, monthly, annual and seasonal means and some cloud-specific fields … I doubt the GP would be able to evenly perform in the prediction of all of these outputs. You acknowledge that by setting a criteria to rule out some of the observational constraints, when the GP uncertainty is too high, but I think it would be interesting to show the GP skills for the different fields (maybe a simple

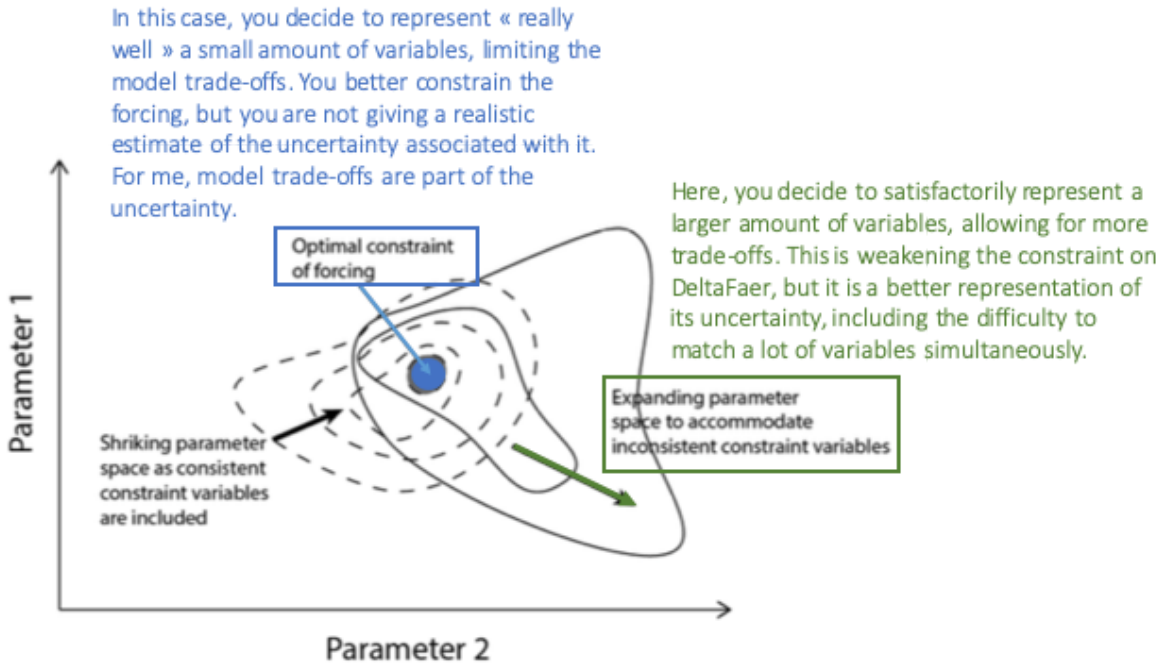
test and a table with the different skill scores). This way, we can know more about the GP skills when the constraint is ruled out or when the constraint is kept. I would also be very curious to see if the GP has the same skill in predicting the observational constraints that look consistent in the pair-wise plot (Figure 5) and the inconsistent ones. I have the feeling that the most inconsistent ones might be the ones with the weakest variability in the PPE, the weakest dependance to parameters and the most difficult to emulate. This is just an intuition that the GP could give us more information about the parametric dependance of these variables and the need to keep them in the final constraint or to rule them out.

2. Also, you refer to the model-observation differences as “RMSE”' and “NRMSE”, which is still confusing to me. This led me to assume that you were considering 2D-fields of model outputs and observations, to be able to compute RMSEs within the regions detailed in SI Table S2. But I am pretty sure I misunderstood and you are actually taking the model and observation regional means before computing the differences. Could you confirm that all of your observational constraints are scalars (2D fields averaged over time and space) and that you are taking the absolute differences between two points |model.(1d) - obs.(1d)| ? If this is the case, I would recommend explicitly writing the words ``averaged over time and space” somewhere at the beginning of Section 2.2. I would also recommend not to use the term “RMSE”, which implies that you are doing a mean of squared differences across the grid points, the time steps or the PPE members. You could refer to your model-observation differences as euclidian distances or absolute differences between averaged fields, something like that.

3. Just to clarify a point that I am not sure to understand. **line 252** - *“However, the multi-stage design of the present PPE leaves gaps in the parameter space that limits the interpretability of variance-based methods.”* → How would the multi-stage design leave gaps in the parameter space ? Is it because the NROY space allows for discontinuity in the parameter values ? The PDF of your model parameters after constraint in SI Fig. S11 and S12 do not expose any gaps - there are parts of the space completely ruled out by the constraint (high values of AI in Fig S11 for example), but I do not see discontinuity or gaps. I thought the idea of the multi-stage approach was to define a new plausible parameter space (NROY) in order to do a new sampling of this space and to run a new wave. In this case, I do not see any differences between applying the variance-based method in your first parameter space and in your second parameter space - in both cases you explore only the parameter space you defined as “plausible”, whether it is based

on prior knowledge of the parameter values or implausibility tests and NROY space definition. Could you develop this a bit more ?

4. My last point is more of an open discussion. “Our estimated aerosol forcing range is the maximum feasible constraint using our structurally imperfect model and the chosen observations” → Aren’t you afraid of over-constraining the range ? I am not sure I would call this the “optimal constraint”, because I would tend to view things the other way around : we can not rule out options as soon as the model performs well giving a set of observations. I think your approach is really interesting, because you try to define your performance metric as relevant for the problem as possible. I like the idea of identifying inconsistencies in the model and looking for a multi-variate metric that really represents what calibration can improve, rather than being polluted by some inachievable observational constraint that even the best calibrated model would not reach because of structural inadequacies. That is why I really like how you rule out constraints based on the emulator uncertainty, the observations being outside of the PPE distribution and the pairwise comparison (which is, in my opinion, a really interesting analysis and the highlight of the paper). But I am less convinced about the use of a criteria based on the amount of constraint you reach on DeltaFaer. I am not sure about using the word “optimal” in this case, because the difficulty to tune a model to match a large number of variables is part of the uncertainty. With your method, you reach the tightest constraint, but I am not sure that it is an optimal constraint and it is probably not a realistic estimate of DeltaFaer uncertainty. That said, I think your point is really interesting and is worth noting. I do not have a better way of deciding where to stop when adding the observational constraints, as for me it is a choice between a good representation of a small number of variables or a satisfactorily representation of a larger amount (see figure). I do not know if we can reach an optimal constraint and how to decide which observational variables should be ruled out. But over-constraining the forcing could lead to an under-estimate of the model uncertainties, which is not desirable. I tend to see the PPE as a tool to explore the diversity of model error trade-offs andtheir impact on forcing, feedback and climate sensitivity values.



Minor comments :

line 30 - "our analysis of a very large set of model variants exposes model internal inconsistencies that would not be apparent in a small set of model simulations" → I get how enlarging the size of the ensemble improves the identification of such inconsistencies, but there is no Figure comparing the inconsistencies in the initial 221 PPE with the inconsistencies in the 1 million emulated ensemble. Did you compare them ? Do you, indeed, need 1 million emulated simulations to expose the inconsistencies ? I feel like the same inconsistencies could be present in both ensembles.

line 80 - "suggesting that parametric uncertainties in DeltaFaer are as important as structural model differences" → this sentence assumes that considering a multi-model ensemble allows the quantification of structural model differences. I would argue this is not true : the multi-model ensemble shows a mix of structural and parametric differences, only multi-model multi-PPE ensembles could help identifying purely structural differences between models.

Figure 1 - great and super helpful flowchart, I really appreciated it ! You could add a comment about how you went from 1st stage PPE to 2nd stage PPE : "Identification of NROY space through history matching" or something like that.

line 131-135 - "Horizontal wind fields above around 2 km in our simulations (model vertical level 17) were nudged towards ERA-Interim values for the period December 2016 to November 2017. Nudging is intended to remove the effects of differences in

large-scale meteorology between our PPE members, meaning we can attribute differences between model variants to perturbed parameter values. We do not nudge winds within the boundary layer, as many of our parameters are intended to affect meteorological conditions, in particular cloud adjustments, in this part of the atmosphere." → I am not very familiar with the nudging techniques : is it intended to reduce the effect of internal variability in your PPE ? By nudging the simulations toward observations, aren't you afraid to also reduce the effect of parametric variability ?

line 137-139 - "We calculated ΔFaer as the difference in top-of-the-atmosphere radiative fluxes between these two periods. We accounted for above-cloud aerosol in our calculation of the components of ΔFaer (Ghan et al., 2016) and aerosol-cloud interactions (Grosvenor and Carslaw, 2020)." → Could you explain how you calculate DeltaFaci and DeltaFari and introduce the terms here ? I have read them for the first time in the caption of **SI Fig. S1**, noted **line 218-219**, and I was not familiar with the terms yet. I think a few sentences about DeltaFaci, DeltaFari and how you compute them are missing.

line 198-200 - "For the second (final) stage, we identified the model variant closest to the centre of the not-ruled-out parameter space, then iteratively identified 220 additional parameter combinations with the greatest Euclidean distance from existing points, until we had a new and diverse set of 221 members that spanned the uncertain parameter space retained from the first stage." → What is the difference between this approach and drawing a new LHS from the NROY space ? I thought the LHS were already designed to sample the space as evenly as possible, isn't it the same goel as computing the euclidian distances from existing points ? Is it because you want to make sure you sample the model variant closest to the center of the NROY space ?

line 218-219 - "We evaluated constraint variables at the regional level, since there are no clear relationships between aerosol forcing and observations of global mean values (SI Fig. S1)." → At this moment we look at **SI Fig. S1** and we don't know yet what DeltaFaci and DeltaFari are, you should either introduce them earlier, or describe them quickly in the Figure caption.

2.4.3 Emulator uncertainty - This Section is really short, I would like to know more about the emulator uncertainty and how you decided which variables to rule out based on the emulator uncertainty (see General comments). I also feel like the **Section title** is not very adequate, since you also rule out constraints when their observed value is outside of the 90% CI of corresponding values in the sample : something that I found really interesting and that could be more developed in the paper. I suggest something like "Selecting and emulating meaningful constraints".

line 323 - "and repeated until ΔFaci could not be not constrained further."

line 327 - "We tested the how the order of introducing … "

line 337 - “the strength of constraint and the bounds of constrained DeltaFaer are insensitive to the number of model variants retained”. → Looking at table S4, I see the strongest constraint when 1000 model variants are retained and the constraint strength seems to decrease as you retain more model variants. This is something I would expect : by retaining less model variants, you strengthen the constraint. But the sentence (**line 337**) is in opposition with this idea, I do not understand why.

line 334 - “The number of constraint variables needed to optimally constrain DeltaFaer does vary with the number of model variants retained (SI Fig. S13 and table S4)” → On the other hand, I feel like the link between number of model variants retained and number of constraint variables needed is less obvious. I do not see a clear relationship between them in table 4.

line 344 - “These positive DeltaFaci and DeltaFari values arise from individually plausible parameter values that produce seemingly implausible model output when combined.” → Is it expected ? Does this reveal structural inadequatie in the model ? Or is it because some of the perturbed parameters should depend on other parameter values rather than being tuned independently ?

3.3.1 Detection of potential structural model inadequacies - I really like your approach to select a “sub-set of observations for which the model-observation comparison is not affected by structural model inadequacies”. I especially loved the “pairwise” comparison. I think this is the most interesting step in your method and a highlight of the paper.

Figure 6 - I do not think you describe how you compute the synthetic examples (blue and purple curves), this is really missing ! I would suggest explicitly describing this part in the text and in the figure caption. Also, you could use a logarithmic scale to show all 450 constraint variables. Or, if it is more convenient to show only up to the 140 first constraint variables retained, I would recommend putting the 52% and 37% arrows outside of the graph. Their values do not correspond to the numbers on the x axis and I found that a little bit confusing.

line 575-598 - I am a bit uncomfortable with the definition of the “optimal constraint” (see general comments).

line 596-598 - “However, we did not anticipate the optimal constraint to include so few constraint variables. These results suggest across 1 million variants, the model is structurally incapable of matching more than a handful of our chosen observations simultaneously (Fig. 6 and SI table S4).” → This is an interesting result and I think I agree with it overall. But, here, you decided to define as “optimal set of constraint variables” the ones that do not losen the constraint on DeltaFaer. It is actually a choice between keeping a small amount of really well represented variables or a larger amount of satisfactorily represented ones (which might losen the constraint on DeltaFaer,

because there is a real uncertainty about it). I think this is linked to my general comments about using a criteria on DeltaFaer constraint to identify the optimal constraint variables sub-set … This makes the results a little difficult to interpret.

line 632-634 - "At present, 97% of variables weaken the optimal constraint. If we could make these variables consistent with the model, for example by altering the structure of the model, then they would instead add to the constraint by further defining parameter relationships that were not constrained by the 3%" → I would remove the word "optimal". I am also not sure about the second sentence. The model can already well represent some of these variables, the difficulty comes from representing all of them simultaneously. Do you suggest that, with a perfect model, this would not happen ? There is no perfect model and with a realistic model, we could hope that improving the structure would improve our ability to well represent multiple fields simultaneously, but I do not know if we can be sure about it.