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 13, 2023

Regayre et al. construct a large, emulated PPE of uncertain process parameters, which they aim to constrain with observations. They identify common sensitivities to parameters between ΔF_{aer} and the observation variables as simulated by the model. Adding constraining variables one after another (going by which one reduces uncertainty most), they find that the maximal reduction in ΔF_{aer} is achieved using only 13 of the 420 observational variables. That is because the usage of additional variables for constraint leads the plausible parameter space of already constrained parameters to expand again. The authors conclude that this points to structural inconsistencies in the model.
This work bravely embarks on a thorough pathway to reducing forcing uncertainty by actually stringently targeting it. This is a welcome deviation from the vague reference to uncertainty that is used to motivate (or only justify?) much aerosol cloud research. Hence I think that the paper is a valuable addition to the research community, allowing to raise the discussion of how to address model uncertainty to a higher level. The methodology is complex and admirably stringently developed and thought through. To me, the introduction and discussion of results are the most interesting and give ample food for thought. This is conceptually also the most demanding, so this is also where most of my comments target and I would be delighted to receive clarifications upon those.

1 Conceptual issues

I'm being tough here and playing devil's advocate, but that's because I believe it's promising, want to know your thoughts and believe that this will benefit the clarity of the paper. Maybe not all of these thoughts need to be addressed in the paper, but I believe the public discussion in the review process is still useful.

1. The generalisability of the results does not become clear from reading the paper, but it would be helpful in order to interpret your conclusions.

    - How much are your results dependent on your scheme/model/model version? Which conclusions are generalizable? To me it seems like qualitative results, like which observables share constraining possibility with ΔF_{aer}, or the finding of a large structural inconsistency between model and observations are generalizable, both across model generations and other models. However, quantitative results to me seem to be bound to your specific model version. New parameterization additions or structural changes to the code would likely change the consistent variables and the range of ΔF_{aer} (as ll. 393 - 396 exemplify).

      You mention the model-bound of your results sometimes throughout your manuscript (also e.g. l. 34, ll. 494 - 497, l. 655), but never state it explicitly and clearly. In contrast, in l. 108 it sounds as if you're deriving not one model-based estimate but working towards the "final best value" (similar in l. 625).

    - Relation to reality: Do we learn anything about a range of ΔF_{aer} outside of your model (version)? Since you're not explicitly stating the opposite, I am assuming that you're implying that your newly constrained ΔF_{aer} bears some resemblance to the ΔF_{aer} in the real climate system. Please make the relation to reality that you assume for your results more explicit.

      Personally, I find this a thorny issue, also regarding the points mentioned below. Reading through the paper I found myself questioning more and more whether this relates to a forcing estimate for reality. Of course, this does not mean to say that this study isn't useful. It for sure is super interesting and important for model development and for updating forcing estimates

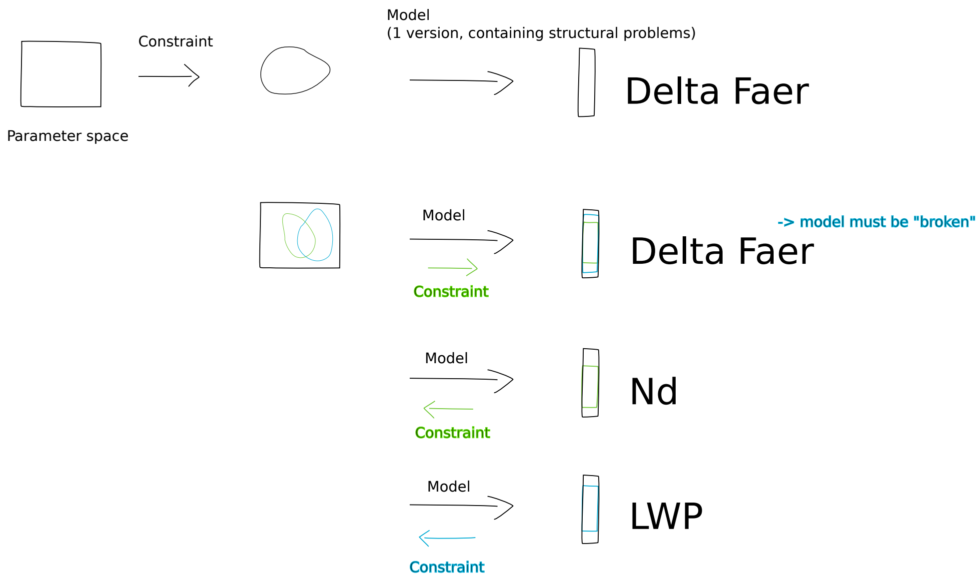

Figure 1: Sketch to clarify my understanding of your work's approach. Observations are used to constrain the parameter space and thus reduce the range of ΔF_{aer}. If they fail to reduce that range, the observations are seen to point to structural inconsistencies in the model and not taken into account in the estimate of ΔF_{aer}.

that themselves rely on (probably similarly structurally inconsistent) models. However, I'm curious to hear your reasoning for how the results, especially the forcing estimate, has relations to or implications for reality beyond (your) model world.

2. I find myself confused by the fact that you point out so heavy structural inconsistencies, but at the same time optimistically derive a new constrained estimate for ΔF_{aer}. In this light, some of your formulations sound too promising to me. E.g. you call your tightest constraint the "optimal" one and also the title boldly promises such a "tight" constraint.

   - How meaningful is any constraint when there are so many structural problems in the model? Structural inconsistencies could also act to lead you to a false constraint. In fact, in l. 62 you argue that model agreement with observations supports trust in the model to produce estimates of ΔF_{aer}, but since you show that model and observations don't agree in most cases, why do you still trust any estimate from this model?
   - Your logic/assumption is that any variable can serve as a constraint if it makes the constraint tighter but if it makes it wider, it should be ignored. Cynically said, this sounds like picking the berries off the cake. I'm unsure whether this is logically justified. I see that by excluding inconsistent variables from the constraint, your constraint is tighter if you naively take the remaining (very few!) variables to be completely meaningful as a constraint. Thus, you "show that it is possible to reduce parametric uncertainty" (l. 29) to constrain global mean aerosol forcing, but is that constraint meaningful? Thus, very critically put, one may question you calling the tightest constraint possible the "optimal constraint" and doubt that it is a real constraint at all.
   - Similarly, in l. 107 you call your "optimal" constraint an "internally consistent constraint". I'd rather say these are combinations that you haven't seen to be inconsistent. For example, if I understand correctly, you may agree with any combination of variables where none widens the ΔF_{aer} estimate of the previous ones as internally consistent and thus different variable combinations could be internally consistent. Thus, one could imagine different, somewhat contradictory variable combinations to give you multiple equally plausible constraints but that means that no one of them can claim internal consistency as that implies inconsistency of the others. This might be nit-picky.
   - I see that I'm being very critical here of what one can even do with structurally uncertain models and the amount of questions might hit you unjustifiedly, but since your work points

out the problems so clearly, it makes me question the positive attitude towards the "classical" constraint that you achieve. Of course, this questions all other constraint work equally. I'm just asking the questions here since you target structural uncertainty and make up this distinction between constraining and inconsistency-indicating variables.

3. The identification of the constraining variables that you use raises questions on their meaning to me. I tried to summarize my understanding of your approach in a simple sketch in Fig. 1. Albeit it is very rough and focused on the point that interests me here, is that understanding correct? There might be a misunderstanding on my side that could explain why I have difficulty following your approach's meaning.

   - Switching the order of variables you may identify different variables as constraining or indicating a "broken"/structurally inconsistent model, so it's not even defined which ones are the ones that indicate structural uncertainty (or does your ordering of variables translate into a clear definition?). Do I understand that correctly?
   - To get a constraint your assumption is that variables can function as useful constraints but showing that this assumption is broken for the vast majority, why does it hold for the rest (the 3%)?
   - Could it be that you're just finding 13 variables that constrain like this by coincidence, same as you'll find some random correlation when you're looking at enough variables?

4. I find it very promising that your work addresses structural inconsistencies and points towards resolving those. Even though I know that a thorough investigation is outside the scope of this work, I would find it helpful to the reader if you could elaborate more on what these inconsistencies could look like and how they could be addressed?

   - Does the model have too many or too few free parameters or both at the same time? The point that there is a range of ΔF_{aer} to be reduced points to too many, but the inconsistencies point to too few (or at something else entirely? E.g. what?).
   - How can we go on to pinpoint the reasons for structural uncertainty and resolve them? You allude to this in l. 654. Could you give an example of how you envision that?
   - As you state in ll. 654 - 656 resolving structural inconsistencies might make the ΔF_{aer} move or become wider again. You point to using your study's methodology in an iterative approach in model development. Here my thoughts return to the questions I posed first again: if your estimate of ΔF_{aer} is to be continuously changed and updated during model development, the relation to reality is questionable, right?

As said, the fact hat your study inspired me to so many questions and had my head spinning is a compliment and I'm looking forward to your thoughts and clarifications on these points.

2 Minor issues

1. To me the points in ll. 29 - 33 would make more sense the other way round: using all observations is impossible because they imply conflicting parameter ranges and thus do not narrow the uncertainty range. However, when you include only those that narrow uncertainty (and exclude the ones that point to structural inconsistencies), you can derive a tight estimate. After reading the paper multiple times, your order makes sense as this is how you present it in the following, but reading the abstract for the first time, this order was confusing to me. To me, my proposed order also takes away some of the optimism in the tight constraint (see above).

2. The abstract partly seems to oversell the scope of your results. E.g. in l. 36 I would put "which would **possibly** then narrow the uncertainty **of our model-based estimate** further".

3. l. 76: What do you mean by a PPE being a "substantial extension of normal model tuning"? I understand both are dealing with uncertain parameters, but one is aiming to find one combination, while the PPE aims to explore all combinations. Thus, they seem like utterly different approaches to me.

4. l. 79: "all important sources of **parametric** uncertainty"

5. ll. 80 - 81: Since there is no approach for a full sampling of structural uncertainties (l. 86) and multi-model ranges certainly don't reflect the full range, I find that your comparison of the size of parametric and structural uncertainties being similarly important needs a disclaimer here.
6. l. 89: "This is because" doesn't make sense to me here, because the second sentence is not the reason for the first statement but rather an illustrative example of an effect.
7. Fig. 1 is a great idea and it really helps to understand the Methods section after! However, there are a few small points you might think about improving here:
   - How do you get from the 1st stage to the 2nd stage PPE? I know you explain it in the text, but it is confusing to not have a small explanation here.
   - Similarly, between the subset of 225 and the optimal constraint, I think an explanation might be nice, something like "retaining only variables that narrow the estimate" or ignoring inconsistent ones.
   - I know you also structured your main text like this, but to me the order of the optimal constraint and the constraint of model parameters needs to be switched. After all, it's the constrained parameters that give you a constrained estimate of ΔF_{aer}, right?
8. Similar to your comment on l. 149, it would help to clarify that emissions are also perturbed where you explain them in the paragraphs before.
9. Sec. 2.1.1 is missing a description or at least mention of the 1-moment CMP scheme (mentioned as a caveat in l. 491).
10. l. 161: Why have you implemented these modifications? Have they been shown to improve performance or do they reflect an updated understanding?
11. l. 167: "Peace et al., (2020)"
12. Sec. 2.1.2: How did you derive the initial parameter ranges?
13. l. 178: What do you mean by "parameters associated with structural model developments"? Do these have any relation to sampling structural uncertainty?
14. l. 200: You're picking the most central member, adding 220 ones in addition to the old ones and get 221 members in total? So are you not using the old ones anymore?
15. I am missing some discussion or Figure of the skill of the emulation.
16. l 222: Why are only ocean boxes included?
17. l. 264: Maybe I misunderstood, but I thought you have an emulator for each of the variables as well, no? So why are you using the 221 PPE members and not the 1 million as for ΔF_{aer}?
18. l. 327: remove the first "the"
19. l. 348: "observationally implausible" vs. "after optimal constraint" (l. 351). As I understand the plot it shows the final constrained estimate, i.e. after ignoring the variables that show structural inconsistencies in the model. Observationally implausible variants would have been removed in a previous step where certain portions of the parameters space were excluded from the rest of the study (e.g. Sec. 2.4.3).
20. In l. 439 you discuss redundant variables, but later you only make the distinction between consistent and inconsistent variables. Is that because redundant variables will either tighten or loosen the constraint and will thus be classified as one or the other? Only a variable that does not change the ΔF_{aer} at all would be redundant in your classification and none is shown to do that? Could you clarify that point here or later in the discussion of results?
21. In Fig. 4 and several of your supplement Figures you use red and green to distinguish two different scatter variables. With regard to accessibility, this seems an avoidable obstacle as not many other colors are included in this Figure (I used the Color Blindness Simulator Coblis (`https://www.color-blindness.com/coblis-color-blindness-simulator/`) that is recommended by ACP (`https://www.atmospheric-chemistry-and-physics.net/submission.html#figurestables` to check).

22. Fig. 5 is a great visualisation of the constraining process! However, I have some small suggestions to improve the understanding of the lower right panel:
    - The observed value in the legend is different from how it's displayed in the Figure (dashed line).
    - To understand both this and the left Figure, it might help to indicate in the matrix where the lower right panel comes from.
    - The point here is that the mean of the pink distribution is further from the observed value than the green one, right? Could you highlight that difference to the mean if that is the point?
23. In Fig. 6, it might help to indicate the "optimal constraint" clearly, and to point out in the legend (not just in the caption) that the blue and purple points are synthetic.
24. l. 544: add that it's the tightest constraint with these observations and this model version.
25. ll. 668 - 669: I appreciate that the grand implications are improved model skill in the future, but I don't think that link is clear enough to passingly use it as the concluding sentence. Your work highlights a thorough way to appreciate, quantify and reduce model uncertainties, and a "step change in model development at reducing model uncertainties" would already be an amazing point to work towards. The relation between uncertainty reduction and improved skill is vague to me and not spelled out argumentatively in the paper, so I would refrain from using it here.