Response to Anonymous Reviewer #1

We appreciate the well-considered comments from this reviewer. It is clear this reviewer appreciates the challenges we face in constraining aerosol forcing uncertainty using very large ensembles of climate model variants and diverse observational data sets. We provide some detailed discussion that answers all of the reviewer's questions and clarifies some misconceptions. Suggestions on how we might improve our article have almost all been adopted.

Reviewer comments are shaded blue and are followed by our responses. Suggested changes to the text are bold.

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).

There are good reasons to expect a degree of variation between climate models and generations of the same base model. Each model in the CMIP6 (Coupled Model Intercomparison Project Phase 6) ensemble inhabits part of the uncertain historical aerosol radiative forcing (ΔF_{aer}) range. However, climate modelling centres usually only submit the well-configured best choice of their (tuned) model to model intercomparisons, so the ΔF_{aer} range they inhabit is reduced to a single point. Some models have relatively high ΔF_{aer} whilst others have relatively low values. Thus, the hidden distribution of ΔF_{aer} values will be centred on high values for some models and low values for others. What is needed, to more robustly constrain aerosol-forcing uncertainty, is a multi-model perturbed parameter ensemble (PPE) that samples both single-model parametric uncertainty (as we do here) and the structural uncertainty caused by model centre specific code choices (as we indicate on lines 660 to 670). The effect of observational constraint on the as-yet-unquantified uncertainty in these collections of models would almost certainly produce a constrained ΔF_{aer} range that is wider than our single-model constraint, since the unconstrained uncertainty would include the effects of model structural differences. Constraint of a multi-model PPE using the same set of observations would produce new methodological challenges, though the resulting constrained ΔF_{aer} range would be far more robust than intercomparisons between a few dozen models.

From a single-model perspective, the degree of generalization in our results between model versions will depend entirely on the structural changes implemented and how they affect the processes that cause ΔF_{aer} uncertainty. For example uncertainty in droplet activation, which is a significant cause of ΔF_{aer} uncertainty in our current model (and probably other current generation climate models), may be reduced by implementing a more complex parametrization scheme that better accounts for dependencies on spatial and temporal variations in environmental conditions. Such a change could conceivably a) increase/decrease the relative importance of this parameter as a source of model uncertainty, b) shift the central tendency of observationally constrained ΔF_{aer}, c) affect the range of constrained ΔF_{aer} and/or d) the degree of internal consistency with other model variables. We have constrained the associated parameter (Sig_w) to a range consistent with observations, so it is possible any new parametrization scheme would produce ΔF_{aer} values that span a similar constrained range. However, a replacement scheme with greater complexity may introduce new sources of parametric uncertainty and if new parameters introduce compensating errors, any constrained range may be larger than previous constraints.

The constraint of ΔF_{aer} we present here is not intended to be final. This article is the current end-point of a series of research projects where we have quantified, evaluated and constrained model uncertainty and inspired model structural developments. We have narrowed the ΔF_{aer} range with each step in this process, as our model evolved, our constraint methodology improved and we challenged the model with new observations related to previously unconstrained model process. Agreement with energy-balance constraints gives credibility to the state of our current model and our constraint

methodology, despite the potential inconsistencies identified. Although we have achieved an optimal constraint for our possibly structurally inconsistent model and our chosen set of observations, there is excellent potential to further constrain ΔF_{aer}. We have yet to reach the maximum feasible reduction in ΔF_{aer}.

- 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).

Our optimal constraint on ΔF_{aer} has a 90% credible interval spanning -1.3 to -0.1 W m^{-2}. We state this range in the abstract (line 30), in analysis of the optimal constraint (line 581) and again in table 1 (line 619). For additional clarity, we will move the bracketed values in the abstract to follow immediately from our constraint, rather than the energy balance constraint:

"constraining it to a range **(around -1.3 to -0.1 W m^{-2})** in close agreement with energy-balance constraints."

On line 108 where we say "This constraint does not make use of all available observations, therefore our central estimate of forcing may not be the final best value.", we're attempting to be transparent about the fact the ΔF_{aer} distribution may shift with further model developments which open the door to utilizing additional observations. We're certainly not aiming to achieve a single 'best value' of aerosol forcing, since our methodology is based on an understanding that some of the parametric uncertainty may be inherently irreducible (due to compensating errors) and an awareness that other structurally different climate models may be constrained to a different "best estimate". We therefore add to this sentence "**which would ultimately be achieved in a model with no remaining structural deficiencies.**"

- 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 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.

Agreement between our optimal single-model constraint (that neglects observations associated with possible internal structural inconsistencies) and constraints based on energy-balance arguments suggests increased confidence in the shared ΔF_{aer} constraint. We

openly concede that the distribution of ΔF_{aer} values may change between model versions and between models (lines 654 to 656) and call on our community to collaborate in the creation of multi-model PPEs so that more-robust observational constraints of ΔF_{aer} that account for both structural and parametric sources of uncertainty can be attained (lines 662 to 665). We think these statements make it clear that our constraint is not the final answer to this long-standing problem of reducing ΔF_{aer} uncertainty. For clarity, we will add the following sentence after line 665:

"Constraint of perturbed parameter uncertainty across multiple models will help close the gap between constrained model values of forcing and the real-world value."

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.

This is the tightest constraint achieved so far through our own research over the past decade and we are not aware of any other research that quantifies and constrains single-model ΔF_{aer} uncertainty to as narrow a range as we have achieved here. We also make our definition of optimal clear in the abstract (L33) and on lines 575 to 578: "We define the "optimal constraint" to be the greatest reduction in ΔF_{aci} achievable using our specified set of observations and structurally imperfect model."

We cannot see how to achieve a tighter constraint on aerosol forcing without either a) first realising structural model developments that address the possible inconsistencies we have identified so that more observations can be used for constraint, or b) identifying new observations related to processes that cause the remaining uncertainty.

- 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?

In the introduction, we are pointing out the incorrect assumption that constraint to historical observations automatically leads to a trustworthy value of ΔF_{aer} (lines 62 and 63). To make this a bit clearer we now write:

"It is assumed that good agreement of a model **simulation** with observations ensures that the model is able to make trustworthy estimates … Yet, "

In fact we do not fully trust our current estimate of …, which we state in the discussion on line 654 "In practice, the magnitude and distribution of observationally constrained ΔF_{aer} values in a structurally improved model may differ from the original model values…"

Additionally, we will reinforce this point in the abstract by adding:

"Structural model developments targeted at the identified inconsistencies would enable a larger set of observations to be used for constraint, which would then **very likely** narrow the uncertainty further **and possibly alter the central estimate.**"

- 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.

Our aim was to determine how tight the constraint could be in a model without structural deficiencies (or by excluding observations that revealed them). Our assumption is that the constraint would be tighter if we could include additional consistent observations, rather than excluding them because of possible structural inconsistencies. We also present the constrained forcing when we do include all 225 Nd-pairwise consistent variables (row 3 of table 1; a 0.2 W m^{-2} difference from our optimal constraint). So it's not cherry picking, but an attempt to find the starting point for further reductions in uncertainty when structural errors are fixed.

The effects of structural inconsistencies are conflated with the effects of redundancy in information provided by constraint variables that share causes of uncertainty (and relationships with ΔF_{aer}) that are already included in the optimal constraint. It is not yet clear how we might separate these two aspects, since we have not characterized the nature of shared dependencies of ΔF_{aer} and constraint variables on our 37 uncertain model parameters.

- 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.

There are some important subtleties in our method that this suggestion overlooks. We retain the 5000 model variants with lowest normalised root mean squared error (NRMSE) across all N_d-pairwise consistent constraint variables included in the constraint. Thus, any other set of randomly chosen constraint variables will likely have a larger average NRMSE than our optimal constraint, even if the 90% credible range aerosol-cloud interaction values

is reduced by the same amount. Thus, these contradictory sets of constraint variables would not be 'equally likely'. We cannot evaluate all possible constraint variable combinations, but have tested our assumptions and methods using a variety of alternatives (as described on lines 327 to 338).

The constraint described in the introduction on line 107 is our optimal constraint. We have by this stage, removed possible structurally inconsistent constraint variables (Fig. 4 and 5) then progressively added complementary constraint variables to achieve the tightest constraint on ΔF_{aer}. We think it is highly likely that these 13 constraint variables are internally consistent and collectively provide an optimal constraint.

- 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.

We would prefer to say that we have a positive attitude to a future way forward that brings together structural and parametric uncertainty in a single methodology. We are positive about the potential end point – a narrower parametric uncertainty range after structural deficiencies have been identified and fixed. Before this work, we didn't see that opportunity.

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.

We have moved the reviewer's sketch to this part of our response.

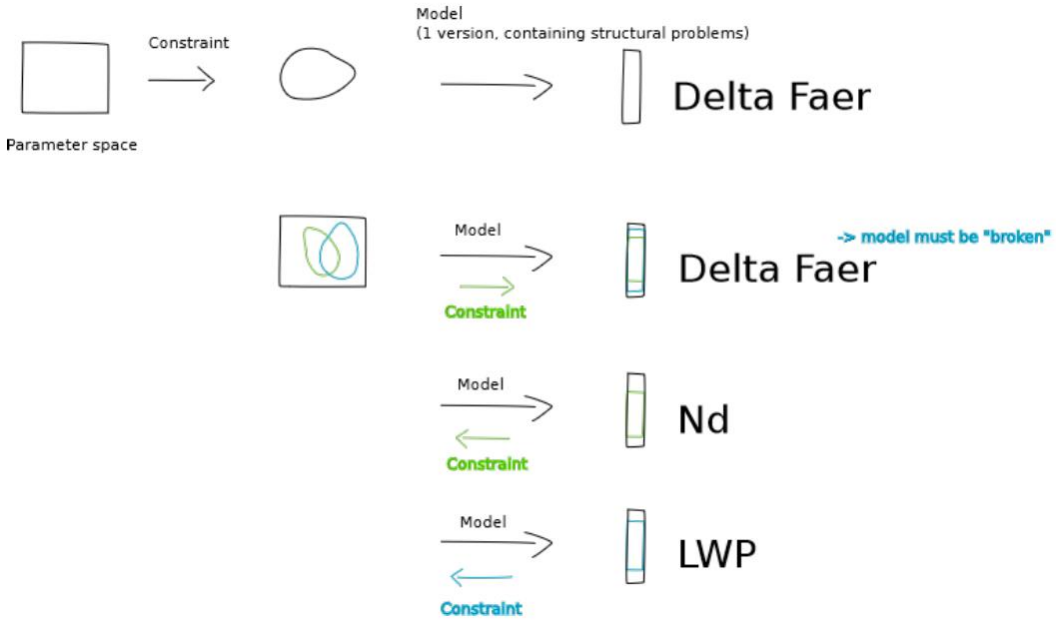


Figure 1: Sketch to clarify my understanding of your work's approach. Observations are used to constrain the parameter space and thus reduce the range of ΔF_{aer}. If they fail to reduce that range, the observations are seen to point to structural inconsistencies in the model and not taken into account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

This sketch addresses one key aspect of the methodology (summarised in Fig. 1 of our article), the identification of an optimal set of constraint variables. It is not entirely clear what the sketch is intended to summarise (an incomplete mixture of key aspects of our method), though the caption is factually correct. After, we evaluate the N_d-pairwise consistency of constraint variables and remove inconsistent variables, we then optimally constrain the model using a subset of available constraint variables that we consider structurally consistent. Additionally, some constraint variables are not included in the optimal constraint because they share causes of uncertainty with one or more of the optimal constraint variables, so provide no additional information about plausible parameter combinations.

- 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?

This is an interesting question and something we will explore in future. It would be interesting to know whether changing the order can provide additional information about structural deficiencies, even if the magnitude of constraint is similar (lines 327-334).

- 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%)?

The 3% of model variants retained span a diverse set of observation types (lines 556 to 566 and 578) which affect specific parametric causes of ΔF_{aer} uncertainty (section 3.3.3). This is a clear example of quality over quantity of observations. These 13 constraint variables (the 3%) provide the greatest constraint on uncertain parameters and ΔF_{aer} dependencies on these parameters, without introducing redundant nor conflicting information.

- 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?

We think it is likely we could find another combination of 13 constraint variables that achieve a very similar degree of constraint and similar constrained ΔF_{aer} values. The diversity of constraint variables in our optimal set (lines 556 to 566 and 578) and their effects on model parameters (section 3.3.3) suggest we have identified a genuine constraint. Within a randomly selected set of 13 constraint variables, any overlap in the dependencies and/or causes of uncertainty (redundancies) would weaken the constraint achieved, which would be sub-optimal. Yet, a more in-depth systematic search of constraint variable combinations may yield similarly strong constraints. In future, we will be able to investigate what such combinations tell us about model structural errors.

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?

Large perturbed parameter ensembles like ours provide multiple opportunities to identify potential structural inconsistencies and/or errors. We have prioritised detecting errors where observations are outside of the model's parametric uncertainty range (or 95% credible interval) and where model skill is inconsistent between variables. However, we hypothesise other detection strategies may further inform our understanding of model structural errors and will describe these more fully on line 594 (below). We hope we have interpreted the phrase 'look like' correctly.

"When applied in combination with the set of the 13 optimal constraint variables, any additional variables weaken the constraint (Fig. 6). This is because the additional variables are inconsistent with those already used. We retain at least 5 thousand model variants for each combined constraint (Sect. 2.4.1), so the result of adding further observations is a compromise in the sense that the existing constraints of ΔF_{aci} dependencies on key parameters need to be relaxed to accommodate conflicting information introduced by inconsistent variables. **We hypothesise the nature of this conflicting information could be revealed by exploring spatially and/or temporally coherent patterns of pairwise inconsistency.** Practically, this **compromise** means that some of the model variants with low NRMSE values are no longer retained…."

- 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?).

We have perturbed all parameters that experts considered may be important for aerosol forcing in our model. Our unconstrained ΔF_{aer} range is so wide because we perturbed a large number of parameters (37) as we describe on lines 341 to 348. The inconsistencies we have identified suggest structural model deficiencies, that may be addressed by either a) revising model parametrisations to better represent processes with a similar degree of complexity, or b) improve process representations by increasing the complexity (adding new free parameters, or increasing resolution). The inconsistencies we have identified do not necessarily imply a need for more-complex parametrisations with additional tuneable model parameters.

- 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?

We describe an ambitious model evaluation and development cycle that a) accounts for parametric uncertainty, b) strategically identifies potential structural inconsistencies (as we have done here), and c) identifies ways to fix them. We suggest prioritising uncertainty reduction and identification of structural inconsistencies in model evaluation and development, over increasing model complexity (through implementation of more complex parametrisations, or increased resolution). We think the causes of potential structural deficiencies could be revealed by exploring inconsistent constraints, as we have done here, in combination with more in-depth exploration of spatial and/or temporal patterns of pairwise inconsistency and analyses of the high-dimensional relationships between parametric causes of uncertainty and constraint variables. Of course, this approach would be far more informative if we could use multiple climate models in this evaluation.

- 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?

The agreement between our process-based constraint and constraints based on energy balance arguments suggests our ΔF_{aer} constraint agrees with reality. However, we agree the likely range of ΔF_{aer} values may change as the model is developed and new constraints are applied. We hypothesise that if we target model developments that reduce structural inconsistencies, then in principle we could converge on an estimate that better reflects reality.

As said, the fact that 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).

We will adapt the abstract text as suggested to say:

"**O**ur 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, **of an order that may be evaluated during model tuning efforts**. Incorporating observations associated with these inconsistencies weakens **any** forcing constraint because they require a wider range of parameter values to accommodate conflicting information. We show that **by neglecting variables associated with these inconsistencies** it is possible to reduce the parametric uncertainty in global mean aerosol forcing by more than 50%, constraining it to a range **(around -1.3 to -0.1 W m^{-2})** in close agreement with energy-balance constraints."

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".

We think a narrowing of the credible range would be almost certain if our model were developed to overcome the potential structural inconsistencies we have identified, allowing us to incorporate additional observations (with shared as-yet-unconstrained causes of ΔF_{aer} uncertainty) into our constraint. We will change this phrase to:
"would then **very likely** narrow"

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.

We are stating that adjusting parameter values is standard practice in model tuning, and that a PPE is a substantial extension because multiple parameters are perturbed simultaneously in an organised manner. PPEs are used to inform model tuning efforts, so they are certainly related endeavours.

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

We are happy to accommodate this suggestion and will move "**parametric**" from earlier in this sentence.

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.

This is a valid point. We compare the magnitudes of these sources of uncertainty to emphasise both are important sources of uncertainty. We will change line 78 to:

"The resulting unconstrained uncertainty in ΔF_{aer}, from sampling all important sources of **parametric** uncertainty in our model, is larger than the range based on energy balance constraints and approximately as wide as the multi-model range **(which conflates structural and parametric uncertainties without fully sampling either)**, suggesting that parametric uncertainties in ΔF_{aer} are as important as structural model differences."

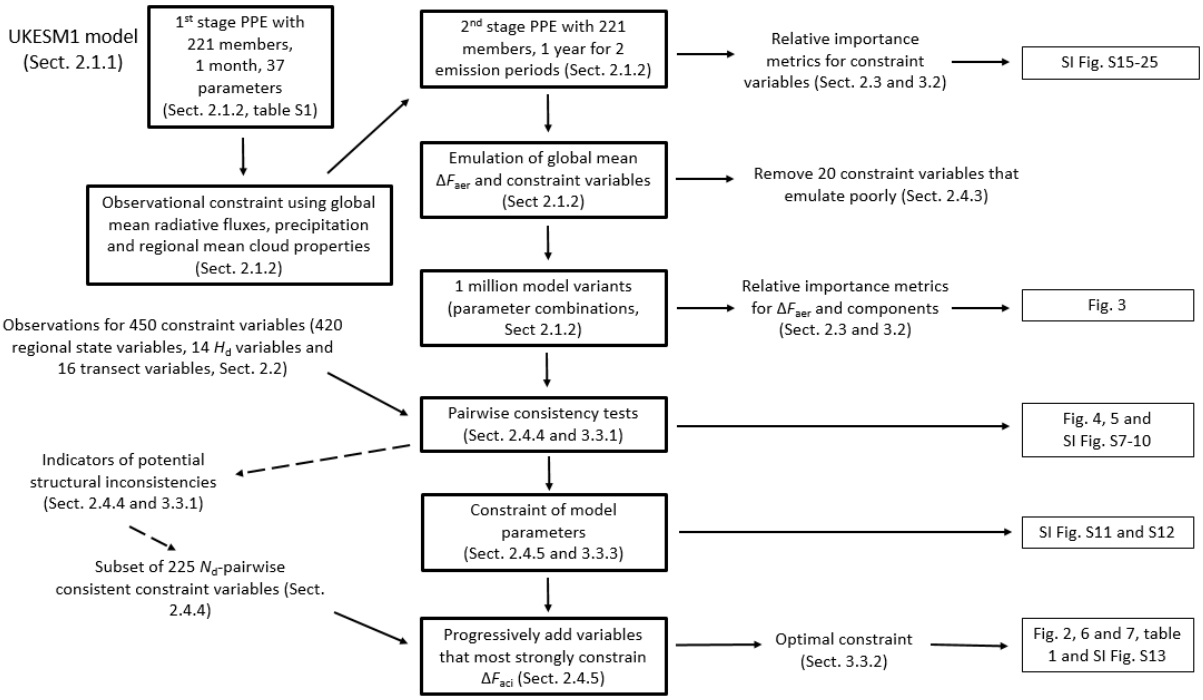
6) l. 89: "This is because" doesn't make sense to me here, because the second sentence is not the reason for the first statement but rather an illustrative example of an effect.

We will change this to "**For example,**"

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?

These are excellent suggestions for further clarifying our method and we have adopted them all in a revised version of the schematic (below).



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.

We think interjecting with descriptions of which parameters and emission fluxes were perturbed would interrupt the flow of what is hopefully a concise and accessible model description. We do give a full description of the perturbed parameters (Table S1) and our justification for including them (section 2.1.2). In the main article, we intentionally only mention parameters that cause model uncertainty and are then constrained, to restrict the information the reader needs to digest and make sense of our key results. With this in mind, we will remove "**and anthropogenic SO2 emission fluxes (anth_so2_r)**" which was included in error on line 391, since this parameter only affects the aerosol-radiation interaction component of forcing and this sentence describes causes of ΔF_{aer} uncertainty.

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).

We will revise lines 157 to 159 to be:
"The activation of aerosols into cloud droplets is calculated using distributions of sub-grid vertical velocities based on available turbulent kinetic energy (West et al., 2014) and the removal of cloud water by autoconversion to rain is calculated by the host model **using a single-moment cloud microphysics scheme**."

10) l. 161: Why have you implemented these modifications? Have they been shown to improve performance or do they reflect an updated understanding?

None of these modifications are new. We have included them all in previous PPEs cited in this article and include them again here as they are potentially important causes of ΔF_{aer} uncertainty. Model evaluation under uncertainty and model development strategies have not yet been fully aligned, so these modifications are not yet in the standard version of our model. We will include an additional sentence on line 162:

"**Including these structural changes adds complexity to our model that we consider worthwhile given their potential to interact with other processes and affect ΔF_{aer}.**"

11) l. 167: "Peace et al., (2020)"

Thank-you

12) Sec. 2.1.2: How did you derive the initial parameter ranges?

This is an important omission. We will include additional text on line 178:
"**Following Regayre et al. (2015), Yoshioka et al. (2019) and Sexton et al. (2021), uncertain parameter ranges were determined by formal expert elicitation using the approach described in Gosling (2018).**"

Gosling, 2018: https://doi.org/10.1007/978-3-319-65052-4_4

13) l. 178: What do you mean by "parameters associated with structural model developments"? Do these have any relation to sampling structural uncertainty?

This refers to a) primary marine organic carbon emission flux (prim_moc), b) scavenging efficiency of Aitken mode aerosol in convective clouds (conv_plume_scav), and c) cloud droplet spectral dispersion via effective radius shape (bparam). These structural developments and their effects are described fully in the referenced articles. We will add "**Parameters associated with recent model structural developments are highlighted in bold.**" to the caption of table S1, but will not elaborate in the main article as we prefer to only describe uncertain parameters that affect our key results.

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?

This is correct. In the paragraph above (lines 185 to 197), we describe sampling 1 million model variants from emulators of the constraint variables, and ruling out implausible model variants in this 1st stage. We are careful throughout to refer to PPE parameter combinations as 'members' and parameter combinations from emulators as 'model variants'. We will include additional text at the start of line 201:
"**Thus, second stage PPE members correspond to a diverse set of parameter combinations from the not-ruled-out-yet set of stage 1 model variants.**"

15) I am missing some discussion or Figure of the skill of the emulation.

Because the impact of emulator uncertainty relates directly to our constraint methodology, we describe this aspect of emulation in section 2.4.3. The focus of section 2.1.2 is the PPEs themselves. We validated all emulators and they perform surprisingly well with variance across the 37-dimensional parameter space typically being significantly larger than the quantified emulator uncertainty. We will include a new figure in the SI (copied below) that shows emulator skill for all constraint variables in our North Atlantic region, used to produce Fig. 5. For all constraint variables shown here, emulator skill is sufficient to rely comfortably on emulator values as representative of model output.

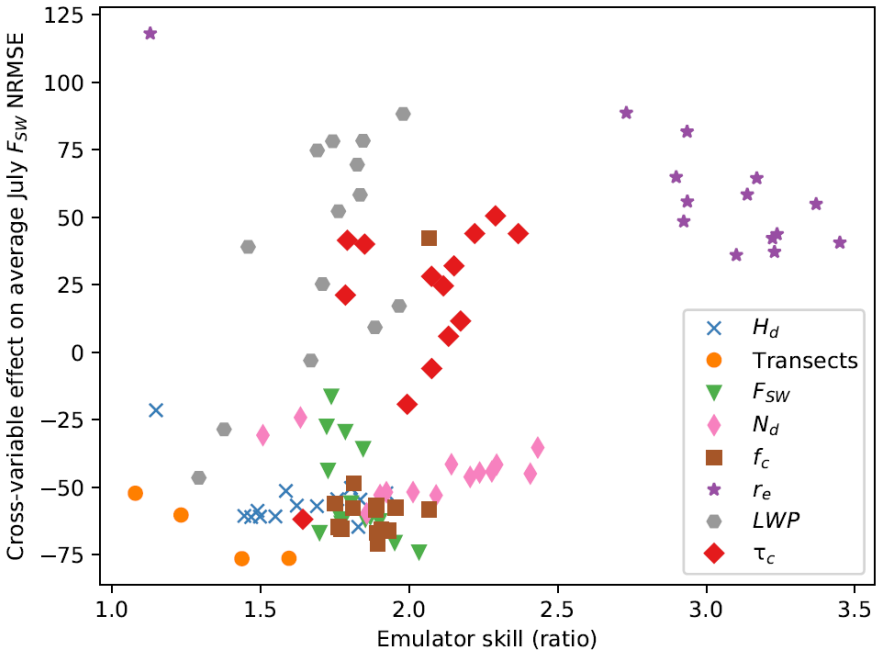


Fig. S_X. Emulator skill for North Atlantic constraint variables and H_d and cross-variable effects on average July F_{SW} NRMSE, which correspond to the shading in Fig. 5 of the main article (where F_{SW} is the constrained constraint variable in the subfigure). Emulator skill is quantified as the ratio of the standard deviation in emulator output across our 1 million model variants, to the mean emulator standard deviation corresponding to each variant (parameter combination).

16) l 222: Why are only ocean boxes included?

We think this follows from the earlier content in this paragraph. Namely, we restrict our attention to stratocumulus-dominated regions in an attempt to prioritise constraint of the ΔF_{aci} component of ΔF_{aer}.

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}?

We do have emulators for each constraint variable. We first calculated these relative importance metrics using the 221 members and later, for ΔF_{aer} and its components, we recalculated them using the 1 million model variants. The two methods produced

remarkably similar results. Since we only use the relative importance metrics to guide our understanding, but not our methodological decisions, we elected to make use of the original set of relative importance metrics and avoid lengthy recalculations of the 1 million values over 450 constraint variables.

18) l. 327: remove the first "the"

Thank-you

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).

This is a misunderstanding. The reviewer merges our approach to identifying structural inadequacies with our method for finding observationally implausible model variants (removed from our optimal constraint). For clarity, we will change the sentence at the end of line 348 to:

"As shown below, the associated model variants are amongst those ruled out as observationally implausible **after optimal constraint**."

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?

This is a good point. We've neglected the discussion of redundancy between similar constraint variables with shared causes of uncertainty and how this impacts the extent to which additional constraint variables are rejected from inclusion in our optimal constraint. The effects of redundancy and potential structural inconsistency are somewhat conflated in our approach. We have not yet developed a method for separating these effects, which would require a far deeper regional and seasonal analysis of cross-variable constraint effects and dependencies on our 37 parameters.

We will change line 586 to:

"This is because additional variables are **either redundant (no additional benefit in reducing aerosol forcing uncertainty range because key parameter dependencies are already constrained)**, inconsistent with those already used **(expand the parameter space and widen the uncertainty range)**, **or some combination of these**.". Additionally, we will change "is a compromise" on line 587 to "**can force a** compromise".

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).

This is an oversight on our part, which we will rectify.

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?

These are good suggestions. We will adapt the "observed value" legend to match the sub-figure. We previously indicated the constrained variable using '*' and will add similar features to the variables used for constraint. Additionally, we will clarify this in the figure caption. The 3rd suggestion here is nearly correct. We evaluate pairwise consistency based on the median NRMSE, not the mean which is obvious from the peak of each pdf. We will add visual indications of the change in median NRMSE after constraint as suggested.

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.

Another good suggestion. We will highlight the optimal constraint using a contrasting colour and will add "(synthetic)" to the first 2 legend items as suggested.

24) l. 544: add that it's the tightest constraint with these observations and this model version.

It is important the reader appreciate this point, though we cannot see how this relates to line 544. However, we have made it explicitly clear to the reader that our optimal constraint is caveated by our choice of observations and use of a potentially structurally imperfect model in the abstract (lines 33 and 34), on lines 574 to 577 where we define "optimal" and elsewhere in the text (lines 318 and 593). We decided not to include this phrasing elsewhere (e.g. line 643 and 644) where we think it has potential to obscure the key point being made.

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.

There is a well-established link between ΔF_{aer} uncertainty (which suffers from the degenerative nature of aerosol and greenhouse gas forcings) and equilibrium climate sensitivity. Models that we consider 'equally plausible' according to their ability to simulate historical observations and trends, diverge rapidly when simulating near-term climate (e.g. Peace et al., 2020). So, small shifts in the credible bounds of ΔF_{aer} will likely feed through to improved model skill (and user confidence) at making climate projections. Furthermore, constraining ΔF_{aer} uncertainty is perceived as having huge potential economic benefit through its impact on climate projections (e.g. Hope et al., 2020, https://doi.org/10.1098/rsta.2014.0429). We feel it is an important and appropriate point to include in our discussion section.

We will adapt the paragraph on lines 59 to 66 to make the importance of this work for future projections clearer much earlier in the article:

"It is assumed that good agreement of a model **simulation** with observations ensures that the model is able to make trustworthy **estimates of historical ΔF_{aer} and reliable projections of future** ΔF_{aer}, which cannot **themselves** be observed."

Response to Anonymous Reviewer #2

We appreciate the detailed comments from reviewer 2. The reviewer expresses enthusiasm for multiple aspects of the work and makes some good suggestions for how to improve other aspects. We describe here how we plan to adopt the majority of these suggestions.

Reviewer comments are shaded blue and are followed by our responses. Suggested changes to the text are bold.

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.

There is significant variation in emulator skill across our set of constraint variables. However, emulator skill is primarily determined by constraint variable type. We found the concept of using emulator skill as an early indicator of possible constraint efficacy, based on the degree of variation across constrainable parameters, intriguing. As such, we have created a new figure, which we will include in our supplementary information file (copied below).

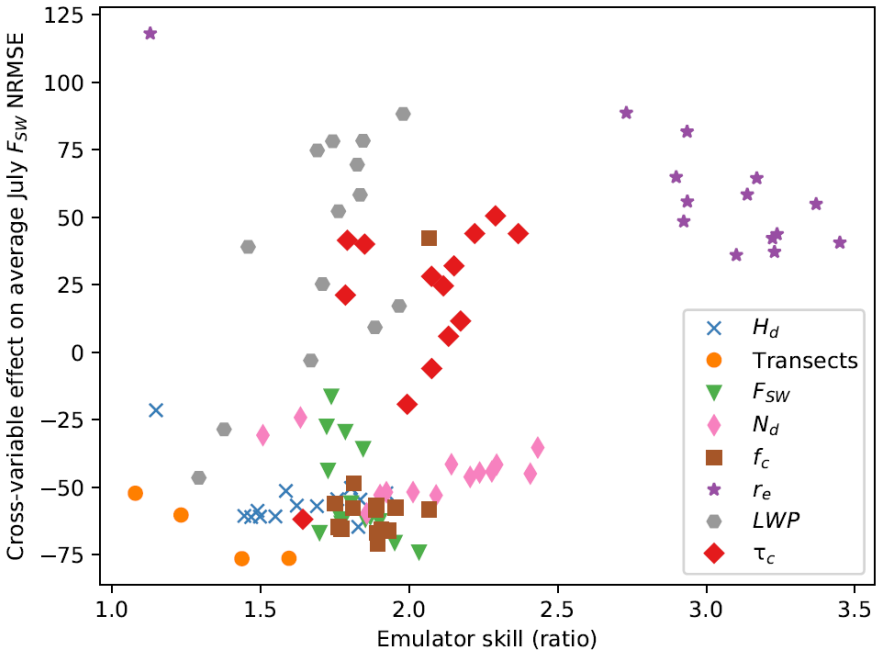


Fig. S_X. Emulator skill for North Atlantic constraint variables and H_d and cross-variable effects on average July F_{SW} NRMSE, which correspond to the shading in Fig. 5 of the main article (where F_{SW} is the constrained constraint variable in the subfigure). Emulator skill is quantified as the ratio of the standard deviation in emulator output across our 1 million model variants, to the mean emulator standard deviation corresponding to each variant (parameter combination).

This figure shows that emulator skill does not guarantee constraint efficacy. For all constraint variables shown here, emulator skill is sufficient to rely comfortably on emulator values as representative of model output. Constraint variables with the lowest emulator skill (e.g. Transect variables and H_d) are more complex as they combine multiple regional effects into a single variable. Seasonal amplitude constraint variables for f_c, r_e, τ_c and H_d are distinct from monthly and annual mean constraint variables in terms of emulator skill and cross-variable constraint effects. The additional complexity of seasonal amplitude constraint variables (requiring accurate simulation of seasonal differences) typically reduces the emulator skill and/or reduces beneficial cross-variable constraint effect on July F_{SW} NRMSE (increases average July F_{SW} NRMSE for r_e).

For some observation types, for example r_e, higher emulator skill does correspond to a modest reduction (or weaker degradation) of average July F_{SW} NRMSE in the North Atlantic. High variation in model output implies these constraint variables can be readily constrained, which will also potentially constrain the associated model parameters that cause changes in emulated values, as suggested by reviewer 2. Yet, despite the high emulator skill, r_e constraint variables are not pairwise consistent (with F_{SW} in this example). Emulators for τ_c and LWP also have relatively high skill, comparable to those for N_d, F_{SW} and f_c, but constraining these variables increases average NRMSEs for other key variables. Thus, these results further support our hypothesis that we have identified a structural model inconsistency.

We will add the following sentence to line 524:
"The degree of cross-variable consistency in not dependent on emulator skill (SI Fig. XX). We have identified two distinct sets of model variables that can be constrained independently, but not in a consistent manner."

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.

We calculate errors as the difference between emulated and observed regional mean values. Following this suggestion we will change line 217 to:

"All satellite-derived measurements were degraded to match the model resolution**, then averaged over time and space for each region**."

To avoid confusion, we will also implement the suggestion to refer to our metric as **'normalised absolute differences'** between model and observation values.

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?

We have not fully evaluated the not-ruled-out-yet (NROY) parameter space from the first wave of history matching. The ruled out space in the marginal pdfs we present in SI Fig. S11 and S12 are 1D representations of the effect of our (second wave) constraint on the 37-dimensional space. Some parameter combinations on the edges of our parameter space are likely ruled out as implausible at the first stage and there may also be gaps within the parameter space that produce observationally implausible output. Variance-based sensitivity analyses require the specification of prior probability distributions for each parameter. We cannot specify such priors without a much more thorough analysis of the first wave results. Any prior distributions we specified would potentially require sampling from emulators where we have no training data (due to ruled out space from the first wave of history matching) which may bias sensitivity analysis results.

Our approach to identifying the causes of uncertainty does not affect the results of this paper, so we don't plan to expand the discussion of these features in the text. We will moderate line 252 to better reflect our lack of in-depth analysis of the first wave of history matching:

"However, the multi-stage design of the present PPE **(Section 2.1.2) potentially** leaves gaps in the parameter space that **may** limit the interpretability of variance-based methods."

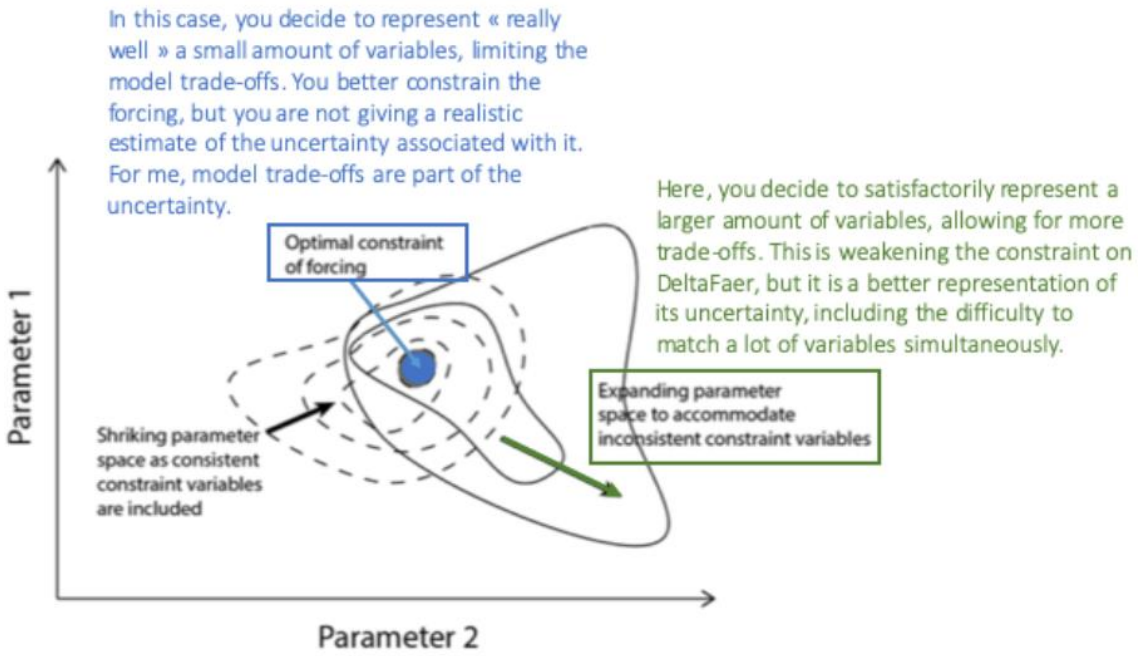
4. My last point is more of an open discussion. "Our estimated aerosol forcing range is the maximum feasible constraint using our structurally imperfect model and the chosen observations" → Aren't you afraid of over-constraining the range ? I am not sure I would call this the "optimal constraint", because 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 and their impact on forcing, feedback and climate sensitivity values.

We agree with this train of thought. There is a degree of subjectivity with any constraint that accounts for multiple sources of model-observation comparison uncertainty. We have achieved a constraint that utilizes a diverse set of observational data that constrain different parametric sources of uncertainty. We retain 5000 model variants that represent a broad range of model behaviour, as indicated by the lack of constraint on many parameter ranges in the marginal pdfs (SI Fig. S12 and S13). Our constraint on ΔF_{aer} is very tight considering the number of model variants we have retained. Relaxing our constraint criteria to include all of the 'trade-offs' to increase the number of constraint variables from 13 to 225 only increases our 90% CI range by around 0.2 W m^{-2} (Table 1). Alternatively, retaining twenty thousand model variants barely changes the constraint (SI Table S4).

We refer to our constraint as optimal with caveats that make it clear the term only applies to our structurally imperfect model and depends on our choice of observations used for constraint. We define 'optimal' in the abstract (L33) and on lines 575 to 578 and clarify our definition elsewhere in the text (lines 318 and 593). In our case, the compromises made in our constraint to take it from 'optimal' (good agreement with 13 constraint variables) to 'sub-optimal' (modest agreement with all 225 N_d-pairwise variables) only increase the observationally plausible ΔF_{aer} range by around 0.2 W m^{-2} (table 1). Our optimal constraint should be viewed as starting point for exploring structural deficiencies that currently prevent more observations being included in an optimal constraint.

As to whether our constraint actually represents real-world ΔF_{aer}, we're unable to say for sure. If we could confirm the real-world connection, our work in this field would be complete. However, our constrained ΔF_{aer} values agree with energy-balance constraints, so presumably give a reasonable indication of real-world 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.

Here, we are referring to the sorts of ensembles (of order 10 model variants) that might be created during model tuning practices, not our 221 PPE members. It is possible, perhaps even likely, that the same inconsistencies are present in our original set of 221 PPE members. But, we retain 5000 out of 1 million model variants (0.5%). We cannot comparatively test our constraint approach on only 221 members. Also, our 221 members disproportionately represent extreme parameter combinations, by design, so we can use them to create statistical emulators. Results from constraining such a set of PPE members will not be a reliable as the constraint we have achieved using 1 million model variants. Once a PPE has been created, it is highly efficient to create statistical emulators and produce a much larger sample of model variants for robust analyses, as we have done here.

We will clarify the magnitude of ensemble we are referring to in the abstract on line 30:

"**O**ur 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, **of an order that may be evaluated during model tuning efforts**."

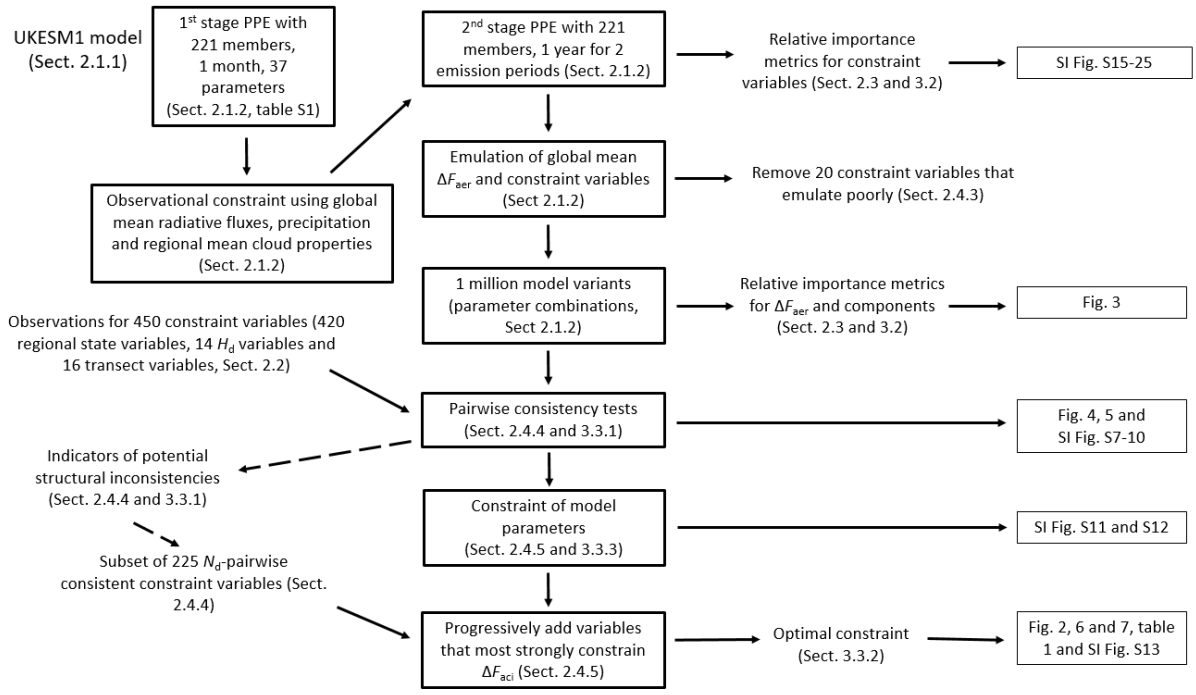
line 80 - "suggesting that parametric uncertainties in DeltaFaer are as important as structural model differences" → this sentence assumes that considering a multi-model ensemble allows the quantification of structural model differences. I would argue this is not true : the multi-model ensemble shows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

We agree that creating a multi-model ensemble with a perturbed parameter component may much better inform our understanding of ΔF_{aer} uncertainty, including what causes the uncertainty and how we might further constrain it. We dedicate the final paragraph of our article (line 660) to this point. We also agree that model inter-comparisons currently sample an unquantified mixture of structural and parametric uncertainties, so will change the sentence on line 78 to:

"The resulting unconstrained parametric uncertainty in ΔF_{aer}, from sampling all important sources of **parametric** uncertainty in our model, is larger than the range based on energy balance constraints and approximately as wide as the multi-model range **(which conflates structural and parametric uncertainties without fully sampling either)**, suggesting that parametric uncertainties in ΔF_{aer} are as important as structural model differences."

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.

Thanks for sharing your enthusiasm for this flowchart. We were initially reluctant to include an additional figure, so are grateful to hear you find it useful. We have updated the figure to include this in an additional box for the 1st wave of history matching and some additional clarifications (copied below).



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?

Yes, nudging is intended to reduce the effects of internal meteorological variability. Each of our PPE members will share large-scale meteorological features, which match conditions where cloud observations were made. Nudging all PPE members to match observed wind fields makes it possible to reliably compare model output to observations without having to average a very large number of free-running simulations. However, some potentially important sources of physical atmosphere model uncertainty are neglected (e.g. Sexton et al. 2021, doi:10.1007/s00382-021-05709-9) as their effects would be suppressed by nudging. The only alternative to neglecting these parameters would be to introduce internal meteorological variability, which would introduce additional uncertainty and significantly increase the number of PPE members required to robustly represent model output.

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.

We will revise line 137 to more clearly define these components:

"**We separately calculated components of ΔF_{aer} (Forster et al., 2021) caused by aerosol-cloud interactions (ΔF_{aci}) and aerosol-radiation interactions (ΔF_{ari}). The separation of these components** account**s** for above-cloud aerosol **radiative effects** (Ghan et al., 2016) and **multiple cloud adjustments** (Grosvenor and Carslaw, 2020)."

We will then remove the definitions from line 263.

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?

Latin hypercube designs are intended to span the entire parameter space defined by individual parameter uncertainty ranges. Points are spread as evenly as possible across the hypercube. However, the NROY space from the first wave of history matching is not guaranteed to be a neat hypercube. Parts of the parameter space (parameter combinations) may have been categorized as observationally implausible. A new hypercube designed according to constrained parameter ranges would potentially include design points within the ruled out parameter space (defined by parameter combinations not individual ranges). Our design is preferabl because it draws only from the NROY parameter space and replicates desirable space-filling properties of the latin hypercube design.

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.

Agreed. We will describe these components more fully on line 137.

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".

We also find these aspects of the method interesting, but do not think there are sufficient benefits to lengthening this section. Although emulators are an essential research tool, quantification of emulator uncertainty is not a key feature of our paper. We agree the title of this section could be more descriptive, so will revise the title to:

"**Identifying viable constraint variables**"

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

Thanks for spotting this mistake.

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

Again. Thanks

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.

We considered the difference between retaining 1000 and 20000 model variants might be much larger than we show in Table S4. The 90% CI bounds in table S4 are identical to 1 decimal place. We present all ΔF_{aer} values to 1 decimal place in the main text and so describe the bounds as insensitive to number of variants retained. We show constrained bounds to 2 decimal places in table S4 to reveal how small these changes are in practice.

We will clarify the meaning of line 337:

"However, the strength of constraint and the bounds of constrained ΔF_{aer} **(to 1 decimal place)** are insensitive to the number of model variants retained (SI Fig. S13 and table S4)."

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.

We agree there is no clear relationship and did not intend to suggest there was. We will change line 334 to:

"The **number of model variants retained affects the** number of constraint variables needed to optimally constrain ΔF_{aer}**, but not in a consistent manner** (SI Fig. S13 and table S4), **since changing** the efficacy of individual and combined constraint variables affects the potential for additional observations to further reduce the ΔF_{aci} uncertainty."

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?

Correct, some of the perturbed parameters should depend on other parameter values rather than being tuned independently. These dependencies will vary between model variables.

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.

Awesome. Thank-you.

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.

We had not considered how the arrows may mislead the reader by pointing to x-axis values. We will extend the x-axis, add a dashed vertical line at around 125 constraint variables and will remove ‘140’ (and any later values) from the x-axis. Additionally, we will add “synthetic” to the legend items related to the blue and purple lines. We describe these lines as hypothetical and explain their meaning on line 634 and 637.

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

We use optimal to describe our constraint because including any additional constraint variables leads to a sub-optimal constraint of ΔF_{aer}. We make sure the reader is aware throughout that this is only an optimal constraint for our structurally imperfect model and choice of observations (Lines 33, 318, 575 to 578 and 593). Our optimal constraint is a starting point for future constraint in a model free of model structural inconsistencies and is by no means the maximum feasible reduction in ΔF_{aer}.

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.

We agree with this interpretation. If it were possible to quantify the structural discrepancy term describing the difference between model and real world, we could calculate model variant implausibility more accurately and rule out model variants using impartial statistical methods. However, this is not possible, so in practice there is a balance to be found between the good constraint of a small set of variables (our optimal constraint) and a satisfactory constraint of a larger set (somewhere between 13 and 225 constraint variables in our case). We define our chosen constraint as optimal because including additional constraint variables is a sub-optimal compromise (or trade-off). In practice, the set of what are considered observationally plausible model variants could be much larger, depending on the degree of compromise the constraint is designed to accommodate. We provide the reader with options in table 1 and table S4. The agreement of our optimal constraint with energy balance constraints supports our choice of constraint.

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.

We will **remove ‘optimal’** here as we have already adequately defined the nature of our constraint. Some ambiguity in our second sentence here needs to be addressed. We will change this to:

“If we could make these variables consistent with **other model variables already used for constraint**, for example by altering the structure of the model, then they would instead **potentially** strengthen the constraint by further defining parameter relationships that were not constrained by the 3%.”