Table 1: Road-map for various audiences

Audience	Sections to read
Paleo modeller interested in implementation	all
Mathematically competent paleo researcher not interested in implementation	all but sections 4.2-4.8 and 5.5-5.6
Mathematically limited paleo researcher	"" and also skip the technical parts of section 2.4 and 2.7-2.8
Any paleo researcher with a working understanding of Bayes Rule	can skip all of section 2 except 2.5, 2.6, and 2.9
modellers in other areas of geophysical models	all of high level conceptual parts and selective reading of implementation section 4

Note to editor

It is unfortunate that no paleo modeller (especially one interested in implementing meaningful uncertainty assessment) was a reviewer. Most of section 4 (implementation) was targetted for this specific audience. As detailed below, in addressing reviewer comments, we do not see how the paper can be made shorter (whatever shortening we've implemented was offset by added text to increase accessibility). If we do not provide an intro to Bayesian inference, we've lost most of the paleo community. If we remove the implementation section, we lose the albeit limited level of detail needed to convey an example of what is actually required and some of the experiential lessons the authors have learned over their decades of work on the topic. However, as explained below, we believe our reader specific roadmap addresses the size issue for many parts of our wide (paleo) target audience.

Response to the reviewers

We thank reviewer 1 for her honest communication of difficulties in reading our submission. This has challenged us to make the document more accessible. The reviewer has subsequently also kindly read parts of some revised versions to assess progress (and critique) efforts into making this more accessible.

We thank the reviewer 2 for their critical assessment of our work though a bit more condensed version would have been appreciated.

In the following we address their concerns point by point.

Reviewer 1

Reviewer Point P 1.1 — I have had no formal mathematics training since the age of 16, focusing my studies on geography and biological sciences. I have 15 years' experience of field and lab-based Geoscience since the start of my PhD, focusing on reconstruction of past RSL change. I have worked with models, model outputs and the modelling community, but I would not call myself a modeller.

Reply: The original draft was run by a paleo data expert (Fabrice Lambert), who raised no concerns about accessibility, but the reviewer has made clear that there is a significant part of the paleo community

who have no University level math and for which significant parts of the initial submission would be inaccessible.

Reviewer Point P 1.2 — From my own personal perspective, as a data gatherer, I am afraid I am unable to realize the authors stated goal. The paper is likely of use to those with mathematical training, for example postgraduate palaeo modellers, but for those of us without this background, much of the paper is inaccessible as it heavily utilises mathematical jargon and equations. For example, in outlining what Bayesian inference is, the authors by page 5 include lines such as "To be consistent under various sets of natural axioms, the conditional probability P(A—B) can be defined as the quotient of the joint probability of A and B, P(A,B), and the probability of B, P(B)" followed by equation 2. I am going to have to admit (in open review!) I had to ask a postgraduate modeller what the "—" notation means (though the authors do state it in line 115, it is not clear to a non-expert that conditional probability refers the whole of the equation or the — notation) and then search for several of the terms in the text; by equation 3 I was incredibly confused.

Reply: We have significantly rewritten parts of the paper paper taking into account the above specific examples as well as some useful follow-up with the reviewer as well as a large set of comments from a member of LT's group who also very limited university math exposure.

Reviewer Point P 1.3 — I have read the whole manuscript, but have been left very lost, rather than informed. This is not a reflection on the authors work, but rather how it is written relative to my experience and training (which is very different to the background of the authors). However, if the authors really do want to achieve their desired goal of assisting the palaeo community in better understanding and quantification of uncertainties, the paper needs to be written in a far more accessible way; or the pitch needs to be altered as a paper for targeted to modellers rather than the wider palaeo community.

Reply: The breadth of the target community and required goals of the paper are a major challenge. This is partly evident by the strong contrast between the two reviews (with the second reviewer largely critiquing the lack of depth, partly arising from the effort to minimize the presentation of equations). From discussions with a glacial geologist group member of LT's (April Dalton), a major improvement that we've added is an explicit reader road-map, specific for different readers. For paleo data experts, most of the implementation section is un-necessary and may be inaccessible. The road-map makes clear that most of that section (4) is not intended for them. This group could also skip much of the mathematical parts of section 2 and still take-away key relevant points and some improved understanding of what Bayesian inference entails. Again, this is now made clear in the road-map.

Reviewer 2

Reviewer Point P 2.1 — The paper is an effort to overview the concept of uncertainty quantification for past ice and climate evolution, that sets out the stall for what the authors think a full (and indeed a minimal) uncertainty analysis should look like and how they think you should get there. It is not a review of current attempts at these uncertainty assessments, and makes scant

reference to them. The path they advocate for is Bayesian probabilistic assessment, where, in principle, all key sources of uncertainty, and in particular model discrepancy, are carefully quantified, models in the analysis have a fast and a slow setting to enable better emulation, data sources have community led uncertainty assessment, with potential forcing data sets being available online, and history matching is used to quantify parametric uncertainty.

Reply: Correct, this is not meant to be a review, but instead an intro and overview, raising the issue of how to meaningfully make model-based inferences of past earth system evolution. One correction, our example framework does not require slow/fast model configurations (but that is a natural and feasible option for at least ice sheet modelling).

Reviewer Point P 2.2 — though I hope my comments on disagreements can be treated by the authors as discussion points rather than corrections

Reply: Unlike the trend in statistical journals, reviews in the earth science community are unfortunately not usually taken as discussion points. But we are happy to take this approach and note this here for the editor's and reader's attention.

Reviewer Point P 2.3 — My main issues with the paper are its length; its lack of engagement with the literature, particularly the work of the bigger groups who have made recent genuine attempts at uncertainty quantification for land ice contribution to future sea level rise and who therefore should be the key target audience for this work; and the lack of examples/case studies/proof of concept work to indicate that the way forward (which is highly specific and comes across as the "only way") given by the authors is genuinely feasible or demonstrably the right thing to do.

Reply: The reviewer has not taken into account the limited (or non-existent) level of understanding within much if not most of the paleo community of what Bayesian inference is. What does the reviewer suggest for the first reviewer (Natasha Barlow), who was not able to get past the basic sections on Bayes Theorem in our original submission?

The core target audience is very clearly stated in the text and this does not include the future ice sheet modelling community. They are not dealing with a state-space estimation problem and are working within a much more bounded phase-space. They are using given climate projections, dealing with ice sheets strongly constrained by marine margins, and considering evolution over a few hundred years. They generally work at much higher ice sheet model resolutions than feasible for glacial cycle modelling contexts. Yes, we believe there is plenty in this paper that could be of use to them (and joint efforts on *e.g.,* internal discrepancy assessment are a no-brainer), and now make explicitly clear that the paper offers a paleo-centric example that readers from other communities will need to adapt to their own contexts. We have also now added a few references in the context of future ice sheet projections.

The breadth of the core target audience is a challenge and we do not see a way to shorten the paper without excluding a good part of our target audience. To address this issue, as indicated in the response to reviewer 1, we've now clearly delineated what parts should be read by which audience members. To add a case study would severely lengthen the paper and is easily worth a full paper on its own (which is in the process of being written).

Given the reader specific road-map, we are more comfortable with slightly lengthening the paper to critically review the few paleo context studies that have worked on this specific problem (to repeat state-space estimation of ice sheet and/or climate evolution (ie with full spatial resolution).

As to the reviewer's comment about the "only way", we are not clear what they are referring to. One of our main take-aways is that structural discrepancy "must be" addressed. Otherwise, we endeavour to provide options and examples for *e.g.,* how to address this and other aspects. Our intent is to sketch out an example (and we now further emphasize "example" in the relevant subsection title) minimal framework of what is needed to credibly address uncertainties in paleo modelling.

Reviewer Point P 2.4 — Not meeting the bar on either of these elements of their road-map in places means that even engaged audiences simply won't be able to follow the mapped out analyses and may never get them working if they try. Whilst I appreciate that the authors don't set out to actually perform the analyses they think ought to be done in this work, there is a middle ground between a full case study showing how to do everything and this piece which I think would be required to justify the length and would be needed in order to actually gain traction with the community.

It might be worrying to those who would honestly attempt to quantify uncertainty in these critical-to-humanity problems that the acknowledged 10-year collaboration between the authors has led only to ideas, rather than implementation (even of pieces of the puzzle). Furthermore, some elements of the so-called "minimal" uncertainty assessment are so hard that solutions are still coming from the statistical community (and have taken years so far and are far from perfect in the case of my own work).

How many years are we from the minimal analysis actually being feasible?

I would like to see the authors tackle the conflict between what they consider to be the bare minimum effort/components required, and the urgency of the decisions that must be made (for example in response to future mean sea level assessment). How many years are we from the minimal analysis actually being feasible?

This said (and to be expanded upon point by point in the corrections/comments I would ask the authors to address), a version of this paper with these particular voices giving the messages on uncertainty they have to give is important for the community and really needs to be published. My comments are designed both to expand the discussion by questioning some of the author's suggested approach (which at times comes across a little dogmatic, particularly for ideas that are unproven), and to try to ensure that key groups in the ice sheet modelling community who are constantly trying to improve their uncertainty assessments have the best chance to engage with the material.

As such, I am going to recommend major corrections, including a substantial cutting of the length of the paper and the addition (or expansion) of some form of worked or running example (the conflict between these recommendations is not lost on me) OR a reworking of the "minimal analysis" to bring some other ideas in if examples for the suggested method cannot be produced.

Reply: No where do we state this was a 10 year collaboration, only that it was a discussion that started 10 years ago (for the course of two months while LT was visiting Durham University and was restarted only after the onset of the Covid-19 pandemic).

The first author has worked on various aspects of implementation for the paleo ice sheet context (and some initial exploration of climate model calibration,Hauser et al. (2011)) for close to two decades. Yes it's not easy, as we well know, but if we are actually interested in meaningful/interpretable inference about past (paleo time scales to be specific) whole ice sheet and climate evolution, these are **example** steps of what is required.

The framework is doable for paleo ice sheets (and more easily for near (100 year) future ice sheet/sea level projections). LT has already implemented most of it for the context of all 4 last glacial cycle ice sheets, most of what remains is for the dual resolution test (which was already described as optional within the example framework) and for emulation of internal discrepancy. This work is expected to be submitted by early winter/23.

On a computational cost level, the framework (with the explicitly stated reduction in suggested quantities of simulations) is also doable for many of the EMIC (Earth System models of intermediate complexity) that make up the large majority of transient paleoclimate modelling to date (CLIMBER, LoveClim, Plasim, UVIC ESM) as well as some lower resolution full ocean/atmosphere CMIP3 era GCM models such as COSMOS (T31 ECHAM5/MPIOM/JSBACH ESM) which LT's group is currently using.

The reviewer raises "the urgency of the decisions that must be made". As we've seen with the last two years of Covid modelling, cherry-picking of science, and policy, poor modelling is a poor solution to "urgent problems". Policy needs interpretable risk assessment which in turn needs confident uncertainty assessment. If the relationship between model predictions and the physical system of interest is not meaningfully specified, the modelling has zero utility for policy. Quick modelling that significantly underestimates uncertainties can easily hinder the policy process. Quick but honest modelling will require very large uncertainty estimates. Yes, good uncertainty assessment takes time, so the sooner more effort in the community is put towards this the better.

Reviewer Point P 2.5 — 1. Paper length. This paper is 54 pages long and that is too long for almost any paper. Furthermore, there is no analysis to describe, there are no results to report and there are, therefore, very few figures. There is no novel mathematics and no new geophysical model to describe via equations. Perhaps if this were a review paper, taking a deep look at the key contributions or attempts at uncertainty quantification in ice sheet modelling, methodology and exemplar implementations from other fields, that might explain the length. But most of this is also lacking. What accounts for the unusual length is both attempting to place the contents of undergraduate texts in probability and Bayesian statistics into a journal article (and spending 12ish pages doing so) despite these concepts already being well established within the climate and paleo-climate literature. Most of the first 9 pages and all of the discussion of MCMC (and the confusing transponder analogy and references to it) are simply not required. MCMC is well known in paleo-modelling studies (e.g. see the important body of work by Andrew Parnell), Bayesian statistics is widely used in climate and paleo climate. The authors do cite Rougier (2007) and Sexton (2011) and their related works, but work by e.g. James Annan, Mark Berliner, Neil Edwards, Tamsin Edwards and many more have all sought, in the paelo climate modelling space to introduce Bayesian approaches and explain the approach to probabilistic reasoning. This paper reads as though past earth modelling is an endeavor entirely ignorant of Bayes and opens like a textbook. The place to do that is an actual textbook. In this article, almost all of that material can be replaced by a single page of text to set up the UQ discussion they want to have and appropriate referencing that both points to important statistical literature (as the authors do) and the use of it in the target audience communities. Referring ahead slightly to correction 2, the groups that have capacity and willingness to actually do the UQ suggested here, already know all of this and are attempting to be Bayesian.

Reply: The reviewer is ignoring the stated target audience which includes the first reviewer, namely community members who want to understand what meaningful UQ is for paleo ice and climates contexts (and therefore be able to critically judge relevant papers). Where would the reviewer point the first

reviewer (who is a well-established and well-published paleo researcher) to go? The first reviewer's difficulties have underlined why the intro to probability and Bayesian inference is needed. None of the authors cited by the reviewer provide an intro to Bayesian inference accessible to the first reviewer. What proportion of paleo modellers does the reviewer think would already understand how to meaningfully specify a prior or likelihood or what MCMC is?

More critically, how can one fully appreciate the importance of addressing structural discrepancy if one doesn't understand how the likelihood can be meaningfully specified? How can one critically evaluate a self-proclaimed Bayesian inference without this understanding? Like any discipline, there is a required pyramid of understanding that is missing in much of our target audience. Relegating all this to a textbook won't solve the issue and severely limit the readership.

Reviewer Point P 2.6 — Another cause for the length of the paper is the authors' tendency to write paragraphs or even pages of potential solutions to potential problems that haven't been encountered because nobody (authors included) has attempted the analysis they suggest. I found these passages particularly annoying. Even if (and a big if) the authors were correct in that their solutions solved the imagined problems, it's not established literature (and hence paper-worthy) until you can demonstrate it or demonstrate the technical or physical basis for it working. It seems as though the authors are trying to think of every possible technical difficulty one might encounter if undertaking the analysis they suggest and offering a solution. But is this valuable? Both authors, as experienced applied researchers, know that as soon as they attempt the analysis for real, the problems they imagined are either not real or trivial to solve, and 10 things they hadn't thought of suck all their time and funding. They also know, as most readers of the article will, that the first idea you have for solving a problem doesnt usually work for some reason you can't yet imagine. That fact doesnt invalidate the problems the authors discuss or the need to find approaches to solve them, but it does mean that a lot of the speculation (that takes a lot of space) is of limited value. I would suggest the authors try to limit these passages to explaining the problem that would have to be overcome and, if they must make a suggestion, keep it to a single sentence. It is also worth pointing out that a lot of the potential problems they do pontificate on have been worked on and there are papers. So rather than pontificating they should point to those papers and maybe say how the approaches that exist to meet the challenges they discuss might need refining (or don't meet the bar).

Reply: The problems raised are problems LT has faced over the past 20 years (as now made clear in the text). And we are not trying to create a blueprint, but instead provide starting points and examples. We strongly believe our now added reader road-map approach significantly addresses the length issue. We've also indicated above why we do not see a useful means to length reduction.

Reviewer Point P 2.7 — Lack of engagement with the most relevant literature. Lines 38-40 state that uncertainty assessment is not really happening in the ice/climate/earth system communities but says there have been recent calls for it (one 8 years ago one 7 years ago). Actual recent attempts at it are simply uncited (in fact no other citations to uncertainty assessment in this space are offered in the introduction and the next time I can find reference to it is Line 1234-1235 ("there has been an increasing rate of publication based on computer simulations of past earth and climate system evolution. Yet very few offer any clear uncertainty assessment" no citation). This type of sweeping statement is not enough and offers perhaps the best distillation of this paper's lack of engagement with the relevant literature and the actual work of those groups with capacity to carry

out the UQ the authors are promoting. My comment here amounts to far more than "the authors didn't cite my papers" (actually I was reasonably well cited). The papers that are important are those from the communities providing the assessment as it is they alone who could take up the gauntlet that the authors are throwing down. Easily the standout omission here is the Nature paper Edwards et al. (2021), representing a spectacular attempt at uncertainty quantification for the land-ice contribution to sea level rise. Whilst the analysis those authors produced was not done in exactly the way the authors suggest and, given that it is an academic paper, there is plenty of room for criticism and improvement, that paper represents the most detailed and transparent attempt at uncertainty quantification seen to date in this literature (in my opinion). Perhaps the other key research group in this space contains Robert DeConto and David Pollard who produced the assessment in DeConto and Pollard (2016) and, more recently Gilford et al. (2020) (with papers in between and in submission since), all based on evolving UQ efforts for land ice contributions to sea level rise. In order for this paper and the criticisms of existing efforts of the community it preaches to be valid, they should address at least the major attempts at UQ by the relevant community and state exactly why they fall short (of even a minimal analysis), what they do well and how to bridge the gap. By not even acknowledging the most relevant literature and, instead, simply claiming all attempts fall short, the authors leave their paper just as dismissable as they have treated the major efforts so far. If the authors want to argue that land ice projections and paleo ice assessment are entirely different (despite the obvious links, the importance of experiments with the latter to inform the former (the Gilford paper), and that it is precisely this link that makes UQ really important), that can be argued, but I would still expect the UQ approaches in paleo modelling of Neil Edwards' group (e.g. (to pick 2 from many) Holden et al. 2019, Tran et al. 2016), or Michel Crucifix's group (e.g. (to pick just 2 from many) Van Breedam et al. 2021, Carson et al. 2019). Whilst omitting the work of these groups is perhaps the major problem (because a lot of the UQ in this space as well as the policy relevant analyses has been done by them), a lot of UQ in the broader climate space is missing and that's only a problem because several of the methods or ideas the authors suggest have been developed or attempted for climate-type models and should be referenced. I have highlighted some of these in the minor corrections section as these only amount to citation to cover a point (usually), rather than being fundamental to the messaging and target audience of this paper.

Reply: We've described above how the future ice sheet modelling context is not a core target audience and is quite different than the paleo context. We have now added a brief evaluation of a representative sample of spatially extended (not time-series) paleo studies to date, in the form of a table to minimize further lengthening. But to address a few of your points. Your cited DeConto and Pollard (2016) above is an ensemble-based study with 4 ensemble parameters and no structural uncertainty accounting. Their "Geologically constrained Large Ensemble analysis of future ice-sheet retreat" only has 64 ensemble members. The cited Gilford et al, 2020 uses only 2 ice sheet modelling ensemble parameters, only one data constraint, and also no structural uncertainty assessment. Both are therefore examples of the gap between current practice and what is needed to fully address uncertainties in paleo modelling. Furthermore, two relatively recent overviews on key issues for past ice and climate modelling (Capron et al., 2019; Bas de Boer et al., 2019), make no mention of rigorously addressing simulator uncertainty. These examples further illustrate the context that we are trying to address.

Reviewer Point P 2.8 — Line 376. This and the general discussion of how one approximates internal discrepancy is extremely hand-wavey, and given that these ideas are far from established

in more general UQ, they need a more careful treatment with an actual example, or some watering down.

Reply: We disagree on watering down. Nor can a detailed example be covered without greatly extending the paper length and likely reducing the accessibility for parts of the target audiences. The section includes an example for a single noise source in the context of paleo ice sheet modelling as well as a cited reference that provides a relevant example (even if the reference is not explicitly assessing internal discrepancy). The reviewer is again reminded that he is not an intended audience for this section.

Reviewer Point P 2.9 — On the line in question: First the pilot exploration to find runs that don't give obviously bad fits means what? and how hard is this task actually? and how would you do it without quantifying discrepancy in the first place or at least having some order of magnitude idea as to some kind of tolerance to it? Salter et al. (2019) looked at ways to estimate a bound on such an initial discrepancy through history matching some early waves with a tolerance to error bound (and similar arguments were made in Couvreux et al. 2021, Hourdin et al. 2021), but such analysis takes a lot of setting up (months) and computational power. Most routine pilot explorations I have done with climate models throw up exactly 0 runs that give not 'obviously bad' fits to the data (and if this is not the authors' experience when working with climate models, that should be demonstrated).

Reply: Much of the issues raised about internal discrepancy assessment are addressed the example framework. But for a start, anyone using a complex geophysical model will already have some chosen base configuration. If they have only one, there already is a large problem in interpreting what the results mean. As stated in the example framework, start with just a couple base configurations and do a tentative internal discrepancy assessment. Then one will have to rely on an expanded external discrepancy assessment to account for the limited choice of test parameter vectors. For the ice sheet modelling context, such assessment is doable within two months (specifically the noise generation specification/coding and running of the noise ensemble) based on LT's own experience, and the specific example given in the text mirrors some of what LT and his students have actually carried out. As for "obviously bad" fits to the data, we applied that phrase in a modeller's context of judged very likely to have no inferential value (beyond being an example of a chronology clearly inconsistent with available constraints within interim uncertainties). Hopefully modellers are not wasting their time with model configurations that they think have no inferential value. However, to avoid confusion, we have replaced the sentence with"

"After a pilot exploration to find some ice sheet simulations that are not rejectable based on interim assessed uncertainties (as explained in subsections 2.6.2, 3.1, and 4.1), one could extract a small, high-variance, and low collinearity (*i.e.* parameter vector directions well-separated) subset of approximately 10 parameter vectors."

Reviewer Point P 2.10 — But supposing you could invest the time and expertise to find some, why approximately 10 parameter vectors? Is it possible to make them approximately orthogonal (or are all 10 in the same place)? What does "high variance" mean in this context? And then, even if there are answers to these questions and there is a method for selecting how many runs of what characteristics from a subset of pilot runs with some properties not quite specified but that could potentially be formalized, all that has to be done is re-run every member of this mini-ensemble by varying all of the boundary conditions systematically (the example here is the bed

topography). How one even captures the uncertainty in, let alone propagates the uncertainty of boundary conditions in climate models is a whole unsolved research problem by itself.

Reply: We have now added a footnote description/refinement of what "high variance" means:

By "high-variance", we mean the subset of parameter vectors that have the highest (or near highest) sum of vector component variances after component-wise normalization across the whole set of simulations. Even better would be to maximize the average of the sum of component variances with the minimum of component variances.

As for why approximately 10, one is not going to get a half confident variance estimate with say only 4, and as the reviewer already has indicated, computational costs need to be minimized, so choosing 20 or more basis vectors is we judged as likely not worth the extra cost (though there could well be exceptions depending on the simulators response surface within the NROY space and this is now raised in the text).

If the reviewer's main concern is that we are imposing "rules", that is definitely not the intent so we have added the following sentence to the end of the preamble for the example framework subsection: "Note, these numbers are based on the authors' experience and judgements; and are therefore tentative."

The reviewer is also ignoring that the example we describe in the internal discrepancy subsection is explicitly for ice sheet modelling, for which 10 basis parameter vectors is quite computationally feasible.

We are not sure what the reviewer means by "systematically" but do remind the reviewer that we state "For inferential contexts, one need only assess the combined internal discrepancy for all considered sources at the same time. This would entail imposing all internal discrepancy noise sources simultaneously for each internal discrepancy simulation. A concrete example of internal discrepancy assessment is provided in Goldstein et al. (2013)."

Reviewer Point P 2.11 — Liu and Guillas (2017) developed a method for quantifying the uncertainty in the ocean topography for a tsunami model, Salter et al. (2022) used existing runs from FAMOUS and paleo data to parameterize the spatio-temporal boundary condition for forcing North American Ice sheet deglaciation so that history matching could be used to constrain it, and Astfalk et al. (2022) developed a co-exchangeable process model for SST and Sea Ice using combing PMIP runs with proxies to obtain boundary conditions for a coupled atmosphere/ice sheet model. Whilst some of this work has only recently appeared, the fact that it requires bespoke statistical methodology to obtain the type of distribution on a boundary condition you could sample to get to internal discrepancy (and presumably different methods would be needed for different types of boundary condition), to brush over some ideas as if achieving them is just a matter of running a few experiments with 10 easy to obtain parameter choices and adding a bit of noise, does the community a disservice. The community should want to perform these analyses, discrepancy is important, and running experiments such as those mentioned to assess a component of it could therefore be very important (and I know the authors would say that it is). If it really is as trivial as the authors make it sound, they should be able to demonstrate it quite easily using models that the first author has worked with in the past. Such a demonstration would be valuable to the community and would lead to greater uptake. Some of the challenges I mentioned above that I have come across through my own collaborations in this field suggest that perhaps a lot of community effort is required in establishing a formal methodology for this kind of assessment that actually can work in practice. If that is the truth of it, then this section could be written more as a challenge to the community to work with UQ practitioners to develop these methods where the potential problems and existing attempts are outlined. As it stands it reads as "one could do this and one

could then do that" as if "this" and "that" are simple to do and totally sensible (as in the right thing to do). They may be but it takes more than saying it to establish it.

Reply: Nowhere do we claim the analysis is simple or trivial, but neither is good modelling of complex geophysical phenomena. The reviewer's critique in the above also in part contradicts his earlier statement about how all the "urgency" effectively requires solutions that can be implemented now and that what we outline is way too much to reasonably carry out. Yes, for a climate model, internal discrepancy assessment will take time, but so does climate model development. Model development is an ongoing process, and so is internal discrepancy assessment.

And as indicated above, for ice sheet modelling contexts, an initial assessment is quite doable, though as we mention there are refinements (such as choice and amplitude of noise) that will need attention and could be very well be research projects in their own right.

Reviewer Point P 2.12 — Line 390 (and around). The assumption that individual experiments with single boundary conditions can be done to give an internal discrepancy remind me of the one at a time approach to sensitivity analysis (and we know why that doesnt work). It seems to me to amount to an assumption that each contribution to structural error is independent, but of course it won't be like that at all. The simulator will impose physical constraints meaning that a one at a time assessment surely risks greatly overestimating discrepancy and, therefore, presumably rendering the subsequent UQ not very good at all. It seems to me that the actual internal discrepancy task is far far more complicated and intricate than the authors hypothesize. If it really is as simple as stated, I think it ought to be demonstrated. If not, a more careful discussion that outlines the actual challenges and calls for a pilot might be a good idea. My own view is that a concrete internal discrepancy assessment of a climate model would be years of work and cost hundreds of thousands of pounds in staff and supercomputer time, particularly given the spatio-temporal complexity of the models, vastness of the parameter spaces, and sheer number of boundary condition/forcing files. If that is true, there is genuinely a decision problem to solve as to whether it should be done at all or if the money is better spent attacking other uncertainties in the problem. The value of the proposed analysis intrinsically depends of course on how much it could change the current assessments (e.g. Edwards et al. 2021).

Reply: LT has and still is carrying out such estimation for paleo ice sheets modelling (for which ten thousand glacial cycle simulations is feasible within a project context). Yes the climate modelling context is much more challenging given the computational costs. But that doesn't preclude some internal discrepancy assessment, especially if the community were to coordinate resources. There are a lot of groups running CESM for instance, and it would be sensible for them to share the load of internal discrepancy assessment. And as Jonty Rougier already similarly raised in his 2007 overview paper, dedicating even ten percent of the person and computational resources in the climate modelling community towards rigorous calibration and uncertainty assessment would likely enable attainment of something comparable to our minimal framework example. How to tractably do this for CMIP 6 (latest climate model intercomparison project) climate models is well beyond this paper and will need appropriate methodological development for that context by the relevant community (a point now made the revised text).

Contrary to what the reviewer implies, we do not advocate single noise forcing experiments for internal discrepancy assessment. Instead, we state "For inferential contexts, one need only assess the combined internal discrepancy for all considered sources at the same time. This would entail imposing all internal discrepancy noise sources simultaneously for each internal discrepancy simulation." To avoid

a possible source of confusion (perhaps that mislead the reviewer?), we have now changed the first quoted sentence above to "For inferential contexts, one would only need the internal discrepancy due to the combined effects of all sources at the same time." We do state that individual internal discrepancy source assessment is a subsequent option to identify sources that are large enough to warrant possible reduction by insertion of relevant ensemble parameters.

Reviewer Point P 2.13 — External discrepancy Section 2.6.2 I found this section to be just as hypothetical and even less tangible than the internal discrepancy discussion. The 2 page block could be broken down into subsections for the different potential approaches and a more concrete example given. The general discussion of/reference to "the modeller" throughout here suggests that external discrepancy is somehow a property of their judgement. But philosophically I only really follow that argument if there is one (or a small group of) modeller(s) that understand every part of their model and how it captures the physical processes in the system (and more specifically how that is an approximation to the truth). Such modellers (or small groups) just do not exist in climate modelling. It could be that there are certain uncoupled ice sheet models where such a person/(group) does exist (do say so in the text if so), but the notion of "the modeller" does not exist as far as I am aware for these massive coupled systems. Further, surely a serious sea level assessment would at least involve a coupled ice sheet/atmosphere, so that my argument holds anyway. A climate model is at least 2 models: the "dynamical core" solves Navier stokes in some way and the "physics package" parameterises all of those processes that are not resolved at the relevant grid-scale. Each iteration of the core calls the physics, and the physics itself combines the work of many groups on many different parts of the physics that have evolved, potentially over decades or whole careers. The physics is not just a model but several models (maybe hundreds), each developed to capture something, and only a fraction of which will be revisited by the people wanting to do the current analysis and who might be considered by the authors to be "the modeller". The development of model physics by independent groups of course poses massive challenges for coupling and UQ when the model comes together (or even if a group wants to change just one process, given that all of the others are delicately tuned to the existing one). In these cases, how can we think about external discrepancy (and who should think about it)? If we water it down to "someone or a small group familiar with the model" as mentioned during the minimal analysis section, given that they didn't develop it, that familiarity must be based on having run it (and knowing the current best run) and hence can any external discrepancy from them be independent of the model?

Reply: So the reviewer is effectively saying that external discrepancy assessment is too hard to do. Good science is often hard. If one takes the usual route of just ignoring external discrepancy, then one is giving up on meaningful/interpretable simulator-based inference of paleo earth system evolution.

Some of the options are much more easier to do for ice sheet modelling contexts (for which model development generally involves a very small team with usually no more than two primary developers). But more to the point we explicitly state :

"The judgements may be as simple as "given ice sheet model intercomparisons, dynamical sensitivities, a typical 20 km model grid resolution, and my own extensive experience as an ice sheet modeller, the fit to the observed grounding line positions for a model of the Antarctic ice sheet is likely to have a root mean squared error of at least 30 km. Errors in grounding line position will in turn propagate into correlated errors in both past and present-day ice shelf areas." Each of these example considerations can then be backed up by appropriate referencing. The explicit specification of structural uncertainty

enables reasoned evaluation in the review process and more broadly in the general community."

Though far from ideal, it gives a starting point, and allows the reader to more meaningfully interpret and evaluate the results.

We already provide an example reference for each of the other options of how this has been done except for the one comparing high quality and low quality simulators.

The reviewer is also being disingenuous by the statement :
If we water it down to "someone or a small group familiar with the model" as mentioned during the minimal analysis section,
as that statement is never made. The actual statement in our submission is "This will require informed judgement by one or more experts familiar with the model". We have now inserted "appropriately" in front of "familiar" to ensure the reader doesn't take "familiar" in a weak sense.

The reviewer's example of current generation Earth system models provides an example context where external discrepancy assessment of a simulator component/process could feed into the internal discrepancy assessment of the whole simulator and we've now added a brief discussion of this point to the text. We also now make clearer that collective role of : internal discrepancy assessment, addition of ensemble parameters, and improvement of the simulator in reducing the challenge of external discrepancy assessment.

Reviewer Point P 2.14 — The discussion of model discrepancy neglects the more recent UQ papers for estimating it, even if the authors disagree with those approaches (E.g. Plumlee 2017)

Reply: We do not see how approaches such as that of Plumlee 2017 are applicable to the data-poor high dimensional paleo context.

We get the sense that the reviewer is reading this submission as if he was part of the intended audience which should obviously not be the case (except for the research agenda, cf the added reader road-map). Yes there are subtleties that are being ignored and we'll make that clearer in the revised submission.

Reviewer Point P 2.15 — The "minimal framework" in section 4.1 is not minimal, arbitrary in many places, too specific in others and fundamentally not backed up by sufficient evidence that it is at all feasible or sensible. Furthermore, given that if it were feasible, it would take many years to solve some of the challenges, the position that such a framework is "minimal" is insensitive to the pressing nature of some of the problems (such as feeding into policy support for sea level rise). The playing it down as "simplified" and "an example approximate approach" (line 865), actually serves to undermine the science, as presenting what looks to me to be a multi-year research program that is nearly impossible to follow in places, and to still claim that it is merely a dumbing down of the real problem that only approximates what would be needed, puts "proper" UQ on an unreachable pedestal. The goal should be to engage the community with ideas and to offer ways into the different problems (and lots of the paper does this). I would be strongly in favour of this whole "here is what you have to do" section being cut altogether. However, the authors may view it as a centre piece to their message. In that case, I would say that the authors should take one of 3 routes: a) Reduce the minimal framework to something more general and high-level with examples to avoid/answer some of the criticisms below. E.g. Get a model you can emulate rather than specify that it has to have fast/slow components (but mentioning that this could be beneficial for some quantities of interest), consider experiments for assessing discrepancy components, develop external discrepancy strategy, choose favoured calibration method etc. Some of the specifics you have given can then be

sold as untried ideas that might/might not work. b) Rebrand the minimal framework as not the absolute bare minimum you would have to do, but as the minimal amount the authors themselves would try to do before reporting their uncertainty to policy makers, and to make it clear that some of these ideas might require methodological development with unknown timescales. c) Stick to the minimal framework, water down the adhoc specifics (described below), and demonstrate, by citation and by experiments with some of the simplified ice models that the authors have access to, that the minimal framework is actually feasible to provide a genuinely followable analysis for the community. My specific quibbles with the framework as sold are (in addition to anything to do with discrepancy which has been discussed in the previous major comment):

Line 873. Selecting or developing a simulator with fast/slow components. Only in a very limited process context can you develop a new simulator for climate/paleo reconstruction and the more limited your analysis to a single process/handful of processes the more reliant you are on accurate forcings/boundary conditions from all the processes left out (and accounting for that uncertainty is a research project by itself). In most cases, these simulators are evolved over years by multiple groups or modelling centres themselves. So if we boil it down to selecting simulators, it's not entirely clear that a fast/slow representation is optimal and therefore should be part of the "minimal" framework.

Of course, in certain applications and for certain types of quantity of interest, fast/slow combinations are incredibly useful and they have appeared in climate studies many times for this reason. However, with climate models, often the existence of fast/slow versions of the same model is an illusion based on the names of the model and its parameters. Fast/slow typically relates to the resolution of the solver, but resolution determines the physical processes that can be resolved through the dynamics and which need to be captured by subgrid scale parameterisations (there is a grey area for processes that are permitted but not fully resolved where the nature of the required parameterisations change). That means that often higher resolution = a fundamentally different model even if the same dynamical core is there and some parameterisations are shared. Fast/slow might still be useful, just as two completely different models of the same thing might have strong correlations between parameter spaces that mean a joint analysis might be useful. But there is no guarantee, and there are many processes for which it is guaranteed not to work (processes which are not resolved on the fast model). That's enough doubt for me to mean that a requirement for fast/slow in a "minimal" framework seems too restrictive, especially when it could be a major barrier.

Reply: We respectfully and strongly disagree. Removing a concrete example framework would make the paper much more vague and thereby of much less value in our mind. The framework addition was prompted from discussions about earlier drafts within LTs research group, with members needing to see a concrete list of steps to do meaningful inference for the stated context. Furthermore, the framework was not branded as "the minimal" framework, which the reviewer already acknowledged by quoting "an example approximate approach (line 865)". But to make this even clearer, "example" has been inserted into the subsection title (it was already present in the first sentence of the subsection).

It is also frustrating that the reviewer is misrepresenting the contents of the framework, especially with respect to their main "quibble". For instance, in the framework section bullet 2, we state "If the fast simulator has adequate accuracy for the given context, a slow version is not needed and the implementation framework can be appropriately simplified." LT is also confident that hi/lo resolution (within the 50 to 10 km grid resolution range given current typical current computational resources) emulation should be quite attainable for glacial cycle ice sheet modelling contexts given his own experience in

comparing ice sheet model response at different resolutions. Furthermore, if papers such Tran et al. (2016) can build an adequate emulator to go from a very simple energy balance climate model to a low resolution general circulation climate model (Plasim) with thus a very different level of resolved physical processes, it is likely that useful multi-resolution emulators for present-day climate models can also be built for at least certain smooth fields such as air temperature (this point has been added to the text).

The revised text will make clear that this framework is computationally accessible for the stated example (paleo ice sheet models) and further emphasize that this is not meant to be a blueprint. It will also make clear that for much more computationally expensive models such as state-of-the-art climate models, much more development needs to be done to figure out a computationally accessible route to meaningful UQ.

Reviewer Point P 2.16 — Furthermore, current thinking in this area, at least at some of the major modelling centres, is that process-based UQ is perhaps the way to go. I.e., doing the majority of the UQ (or what they more likely identify as "tuning") at process level (e.g. the single column analyses in Hourdin et al. 2021) where the models are necessarily fast and data comes in the form of very high resolution resolved processes and expert judgment (which can exist because the concept of a single modeller or small group of modellers does apply to individual parameterisations). History matching individual processes (Couvreaux et al. 2021), means the "slow model" which might now be taken to be the coupled GCM has a much reduced parameter space to consider.

Reply: Process-based UQ is an obvious continuation of the dominant GCM tuning paradigm: tune individual parameterizations against process data where possible, then couple, then retune a few key parameters as needed to meet chosen constraints. But the effectiveness of the traditional approach can be broken by non-linearities in the coupled system. One could argue that the absence of whole model calibration is one causal interpretation of the likely excessive equilibrium sensitivity in the CMIP6 results (Zhu et al., 2020; Wang et al., 2021).

Process-based history-matching on the other hand, if done well, should eliminate the risk of over-tuning, and as argued in Hourdin et al. (2021), could make good UQ and simulator calibration much more accessible for computationally expensive and complex geophysical simulators. However, such an approach would still need whole simulator history-matching to address nonlinearities arising inter-process feedbacks. And it would still need whole simulator structural discrepancy assessment that would hopefully be able to make efficient use of component/processed-based structural discrepancy assessment (and which we've added to our list of potential research topics).

We have added a paragraph on the above to the revised version citing the above paper.

Reviewer Point P 2.17 — Line 888 part (e) here sounds like a multi-year research project (as do many of these). In fact our attempts to do it required new methodology that has taken years to develop. If the goal is that groups take up the mantle of doing this minimal analysis, having small sub-parts that require multiyear research projects just puts the minimal analysis out of reach (for many years/decades). Surely a minimal analysis can be gotten off of the ground in time to assist with making important decisions soon?

Reply: LT has spent two decades on that problem (we're presuming reference to "An adequate number of degrees of freedom in the forcing components"). If that issue is not addressed, what is one supposed to infer about the results of some ensemble of paleo ice sheet simulations? Much of the reviewer comments seem to boil down to "this is too hard or too expensive". LT has been doing this in the paleo

ice sheet context. We also disagree with the reviewer's restriction to a 1-2 year time frame when model development (thinking of key ice sheet and climate models in the community) is a decade or longer process (that is ongoing). But this doesn't mean that significant process can't be made on a few year time scales.

It maybe that the reviewer is misreading the criteria as not involving trade-offs. To make it clearer that every criterion will involve trade-offs we've moved the existing sentence describing example trade-offs to the end of the paragraph and modified it to read: "All of the above criteria/choices likely involve trade-offs, with, for example, a poorer quality simulator presumably being faster and therefore enabling more simulations, some of which may be required to address more sources of internal discrepancy."

LT has also been informed by nearly 3 decades of seeing model-based claims in the paleo context that far from adequately addressed uncertainties and that have therefore been subsequently contradicted. Such claims have resulted in eg large wastage of computational resources by climate modelling groups using the divergent ICE4G and ICE5G ice loading "reconstructions" as boundary conditions for past PMIP paleo modelling intercomparisons. FYI, the former reconstruction had no Keewatin ice dome, the latter had an excessive Keewatin ice dome (for which a critical comparison against marine limit data would have likely invalidated the excessively high ice dome). This is the largest inferred last glacial maximum ice dome over North America and thus a very significant feature for a climate model.

We are considering whether some discussion of the above points should be added to the text to better convey why meaningfully addressing UQ is important (balancing off against further lengthening a long paper).

Reviewer Point P 2.18 — Line 906 and elsewhere. The arbitrary nature of the numbers used for the minimal analysis. This line is "order 200 member ensemble", yet there is no link to how many parameters you have, what variables you are trying to capture, how complex you expect the output to be as a function of the input etc. These specific numbers really bothered me and that might seem petty, but my experience is that as soon as someone pulls a rule of thumb number from nowhere (Loeppky's famous 10xparameters springs instantly to mind), those numbers are just used because the paper that pulled them from nowhere can be cited. So, why order 200? Is 200 enough to find that key data cannot be bounded point-wise (line 920) and how would you know given that you could only cover all the corners of up to a 7 dimensional parameter space with 200 points? Similarly, the 50 member ensemble for internal discrepancy on line 933, the "100 or more parameter vectors" from line 954, the "8 more runs" from line 966, the 50-100 slow runs on line 973

Reply: The beginning of that subsection clearly states "The example numbers of model runs given below are for ice sheet modelling contexts. For more expensive climate modelling contexts, run numbers can be reduced, in most steps, by an order of 10, with more attention put on efficient emulator development". One omission that has now been corrected is the insertion of "paleo" in front of the above modelling contexts, in line with are our target context. To avoid different interpretations of "order 10", we also now state "by a factor 4 to 10".

We agree with the danger of others blindly citing numbers here to justify choices and therefore add the following caveat to the end of the example framework preamble:

"Note, these numbers are based on the authors' experience and judgements; and are therefore tentative."

Minor

Reviewer Point P 2.19 — Typo in line xy.

Reply: Fixed.

Reviewer 3

Reviewer Point P 3.1 — Line 23. 'about it evolution', here and next sentence could do with a rewrite

Reply: Agreed for the first sentence. Changed to: "For those studying the past evolution of the earth/climate system, a primary goal is to improve our understanding of the past on both phenomenological and process levels." We see no problem with the subsequent sentence.

Reviewer Point P 3.2 — Line 90 - approx 176 covering around 3 pages as an introduction available in many textbooks or one that could be written in a short book by the authors if they deemed it necessary for the community to revisit the basics of (subjective) probability theory. It is not clear to me that the community needs this, they already go well beyond the basics. A 54 page paper is going to be read thoroughly by almost no-one, and so it is very much worth cutting whatever can be cut. If it is absolutely necessary to redefine probability for this paper, and a journal article is still judged to be the right medium, then the authors should justify their revisiting of this topic with reference to actual articles where probability is misused by the community they are writing for.

Reply: Where should the first reviewer go, given their difficulty in working through even basic probabilistic notation? The reviewer betrays a very limited sense of the paleo "community" to make the implied claim that most have an adequate (for this context) understanding of what probability means.

Reviewer Point P 3.3 — Lines 245-249, it should be noted somewhere that this relationship itself is a model

Reply: As stated, these lines are really just definitions of what simulator and observational error are. How is this a model at this stage? The relationship becomes a model when an "error model" is specified and the terminology already makes this clear.

Reviewer Point P 3.4 — Section 2.5, the discrepancy example. Suppose the discrepancy model was a+bx with some very weak prior on (a, b). Would the number of observations here not trivially lead to the finding of relatively tight posteriors for a and b that centred on (20, 0)? I just don't see how line 340 is true in this simple example. Sure, if you must stick to the simulator, then you have a problem as we know. But if there are this many data, even weak priors will work. The answer to "can't you just estimate discrepancy"from model and data experiments is not a blanket "no". True parameters of a model and the discrepancy are not identifiable (and yet I don't really see this as the point of the example), but even a weak prior can be enough given sufficient data. The problem is what sufficient means, how weak you would have to be, and doing this in massive

spatio-temporal land in such a way as to get a reasonable discrepancy estimate for the future (where a GP discrepancy would have quickly reverted to the mean). Of course, the authors are attempting to illustrate an important problem/concept here, but I think they need to reconsider the example.

Reply: You are choosing a discrepancy model that has the full complexity of our "reality". Is this realistic? As you state: "if you must stick to the simulator, then you have a problem as we know", That is the point of the analogy. To make this clearer, we've changed "However, for realistic contexts, we generally lack the omniscient perspective" to "However, for realistic contexts, complete knowledge of reality ($R(x)$) is inaccessible."

The reviewer seems to be envisioning himself as the audience here. The challenge is to explain the importance of assessing structural discrepancy to a wide paleo audience, including those with no University level math and who might be struggling to get even this far in the paper. LT has used this example in numerous classes with graduate students, and they have generally found this example enlightening. We have made the choice of the simplistic example as is possible to try to make the widest conceptual accessibility.

We've also made the presentation clearer for a non-mathematical audience what exactly is the structural error in the diagram.

Reviewer Point P 3.5 — Lines 303-304 and anywhere Gau appears rather than N please change. The paleo modelling papers and climate papers I have read and reviewed over the years are perfectly happy with the standard notation for Normal distributions.

Reply: We've hesitantly made the change given our diverse intended readership (*i.e.* many with very limited mathematical or statistical competency).

Reviewer Point P 3.6 — Line 355. Is the reason that the min residual is 0?

Reply: We're not following what you are referring to. Linear regression "minimizes the root mean square of the data-model residuals" by design, so it can't be that statement.

Reviewer Point P 3.7 — Line 460. Whilst it might appear that MMEs are ensembles of opportunity, in fact MIPS (and I guess PMIP is the relevant one here) are designed by the target audience together. It is much more realistic that MIP design can be improved to help with discrepancy assessment than a hypothetical "modeller" being able to assess it themselves, and hence perhaps the authors might consider what could be done in this space.

Reply: We agree that designing MIPs to facilitate external structural discrepancy assessment would be useful and have added that point to our list in section 5.6, specifically : "Intelligent experimental design of future model intercomparison projects would also aid external discrepancy assessment."

Reviewer Point P 3.8 — Section 2.7, I suggest is cut.

Reply: We disagree given the experience LT has had explaining MCMC to paleo community members and graduate students. A glacial geologist colleague who reviewed/annotated a recent draft identified this example as very helpful.

Reviewer Point P 3.9 — Line 535, if this section is not cut, note that thinning is not really required or done in MCMC anymore, and adjustments to the effective sample size of a converged sample are made instead.

Reply: We've dropped "thinning" and have added a footnote about need to account for sequential correlations of MCMC samples.

Reviewer Point P 3.10 — Line 600, please give examples. It is not enough to simply claim these Bayesian studies exist and are wrong. There are studies that attempt to be Bayesian that do include them within the paleo climate literature.

Reply: We have changed "geophysical" to "ice sheet and climate models" as in hindsight the term was too broad. Cited examples are now listed in a summary table.

Reviewer Point P 3.11 — Line 623. Ensemble weighting is a hot topic of course and even though I'm not strictly in favour, ignoring the last 12 years of literature (covering attempts with CMIP5 and CMIP6) when attempting to say that explicit efforts in this space have failed to date doesn't seem fair.

Reply: We do not state that all "explicit efforts" have "failed". We do state "the topic is clouded by a preponderance of ad-hoc approaches and reasoning". We do agree that progress is being made, and so have changed the first sentence in the paragraph to "However, this topic has been clouded by a preponderance of ad-hoc approaches and reasoning." which we stand by (and as backed up by the Eyring et al, 2019 paper that we cite, which states for instance: "However, only a few studies have specifically focused on the likelihood of weighted results providing benefits for the intended application").

Reviewer Point P 3.12 — Line 634. I don't know what "Bayesian Averaging" is. Maybe you mean "Bayesian Model Averaging" which would only ignore structural uncertainty if all the Bayesian models that the MME members were embedded in ignored it? Clearly references are needed to establish that it really is in fact very common to ignore structural error when combining models.

Reply: Yes we meant "Bayesian Model Averaging" and we will provide a couple example references.

Reviewer Point P 3.13 — Line 699. It seems strange to cite Sexton's 2011 work (UKCP09) as the most detailed attempt "to date", when a much more detailed effort was published in 2018 by the same authors (UKCP18).

Reply: We respectfully disagree but with clarification. Yes UKCP18 uses more advanced models, but the associated papers lack the detailed description of diagnostic checks and structural uncertainty assessment that are described in the two Sexton et al 2011 papers. To make this clear, we've inserted "descriptively" before "detailed".

Reviewer Point P 3.14 — Line 706. History matching does not provide a set of uncertainty bounds. It provides a subset of parameter space that can be accessed by a membership function (you don't get the bounds).

Reply: Fair enough that the choice of "provide" was inaccurate. The membership function enables the identification of joint parameter vector bounds. Through emulation, it also enables identification of joint uncertainty bounds on predictions. These bounds are conservative, but still represent a major step forward compared to the lack of any credible bounds. As such the text now states "Instead of determining

a likely non-robust posterior probability distribution, they collectively enable the identification of an initial set of uncertainty bounds and a sample of simulations within those bounds."

Reviewer Point P 3.15 — Line 712. "History matching is a Bayesian approach". Is it? What specifically makes it Bayesian? As a history matcher myself, this is in no way meant as criticism of the approach, but there is no likelihood and definitely no prior (because to admit a prior would be akin to saying the best input model was true). Evaluating implausibility amounts to a statistical test (can we reject this value at some level determined by our cutoff) and seems much more frequentist on that basis. The emulators used (if they are used) in the literature can be Bayesian/Bayes linear etc, but that makes emulation a Bayesian method not history matching.

Reply: Good point and text has been corrected to "History matching is a stepping stone.."

Reviewer Point P 3.16 — Line 739 and elsewhere, could cm be changed to c_m or something else?

Reply: The latex source is subscripted, it just doesn't show up well.

Reviewer Point P 3.17 — Line 720. Salter et al. 2022 (but it's been on Arxiv since 2018), have applied it to deglaciation of an ice sheet model in a far less limited context (reference at the end).

Reply: Thanks for bringing this paper to our attention. It is now cited in the paper though in a different spot.

Reviewer Point P 3.18 — Line 765 (and this is a problem for everyone) the authors glaze over how one can history match without a formal discrepancy assessment up front. It seems like the idea is HM then internal discrepancy assessment that will refine the implausibility metric. But that requires starting with a large enough wrong discrepancy so that refining it makes sense (and the internal discrepancy can't be larger). The authors mention later (line 784) that this should be the strategy, but then state that this means structural discrepancy can be learned (785). The authors had already said this was impossible in the case of external discrepancy, so what exactly are they saying here?

Reply: This point is addressed in the example framework, sic

"8. Carry out initial internal discrepancy assessment. Select two to three runs that best fit data constraints while being substantially distinct (sufficiently different parameters and features in the output)..."

"9. Provisionally assess external discrepancies.."

"13. Carry out an initial history matching wave with the emulators. The error model for history matching 960 should include observational, internal and external discrepancy, and emulator uncertainty components. Uncertainty components should err on the side of over-estimation of possible uncertainty"

"15. More carefully assess internal discrepancy. From the set of model runs, select 8 more runs with low correlation between key metrics and between associated parameter vectors that otherwise have the closest fits to the constraint data within observational and structural uncertainties. Use the cumulative 10 parameter vectors (including the two from the original assessment) to reassess internal discrepancy..."

And yes, internal discrepancy assessment is a means "to learn" about a component of structural discrepancy. We would argue the steps above are all a learning process for discrepancy. Furthermore,

to make the framework more logically self-consistent, we've moved the initial provisional external discrepancy assessment to before the first internal discrepancy assessment.

Reviewer Point P 3.19 — Line 769, 770. Couvreaux at al. (2021) and Hourdin et al. (2021) both give very nice examples and description of this process for calibrating convection parameterisations and history matching a 3D GCM.

Reply: We will add a reference illustrating this process.

Reviewer Point P 3.20 — Line 787. Is this actually a methodological error? Up to complete repetition of subsequent waves being OK, I would say that re-evaluating discrepancy is actually part of the methodology. As discussed in Couvreaux et al. (2021), particularly when developing a model and using history matching as a tool to explore the capabilities of a model or falsify a particular development (showing, for example, that a new parameterization doesn't actually improve the model or get close enough to reality to be useful), ruling out all of parameter space (hence finding out that your initial discrepancy was too small) is a powerful feature of the method.

Reply: Fair enough. We've replaced the above with "A methodological inefficiency to avoid (provided you can find a reasonable upper bound).. "

Reviewer Point P 3.21 — Line 845. Are these actually the only two studies. As stated, there are more recent serious UQ attempts from large groups doing these estimates and since they are not cited anywhere, I'm worried about such sweeping statements here.

Reply: We've rechecked the recent litt, and the statement is correct as is. Again, this is a reminder to the reviewer of the context for which this paper is being written.

Reviewer Point P 3.22 — Line 927. What is a "Regression Stochastic Process emulator"? There are so many papers that use emulators in the climate and paleo climate literatures now, and yet this description is new (as far as I can tell and from my own experience) and unexplained and uncited.

Reply: We use the term to emphasize the regression component of the emulator. The subsequently referenced section (subsection 4.6) explains the term as well as points the reader to an entry point paper on the subject (Vernon et al, 2018) . We've now added a reference to that subsection to the line 927 sentence as well as a recently developed R package for this type of emulator (https://cran.r-project.org/web/packages/hmer/vignettes/emulationhandbook.html).

Reviewer Point P 3.23 — Line 945. What is "climate forcing noise"?

Reply: changed to "noise added to the climate forcing".

Reviewer Point P 3.24 — Line 1030. Is there a reference to "GCMs [. . .] generally have more than one hundred poorly constrained explicit parameters"? It seems that the published literature in this area has never really gone much beyond 20.

Reply: The UKCP18 papers the reviewer mentioned Sexton et al. (2021) use 47 parameters and this is just for the atmospheric and terrestrial components (not including vegetation components relevant for

paleo modelling). Depending on process representations chosen, there are more than 40 real parameters that can be set in namelist definition file for POP2 (ocean component of current CESM). There are also numerous discrete-valued flags. The current CICE sea ice model in CESM has at least 35 tunable parameters (google cice-consortium-cice-readthedocs-io-en-master.pdf for the CICE Documentation file (dated Sept 27/21).

Reviewer Point P 3.25 — Lines 1039-1054. This method for parameter selection is an idea, not really in line with current methodology in UQ, and would only work at all if the simulator outputs were quadratic/linear or approximately so. There is an entire literature on emulation/surrogate modelling and sensitivity analysis that could just be cited instead of giving this particular community an uncited idea

Reply: The relative appropriateness/utility of this approach will depend on the choice/structure of the emulator and context. If an RSPE is used for which the global part captures most of the variance, then this ensemble parameter selection approach will naturally fall out of the selection of the structure of the global part of the emulator. And if the main characteristics of the response surface can't be captured by a low order regression model (perhaps in a transformed space), how is any emulator going to have a chance of working for high dimensional contexts where the number of simulator runs is severely limited and emulator predictions are effectively extrapolations? We have now added discussion of these points as well as citations to 2 examples (Cumming and Goldstein, 2009; Andrianakis et al., 2017).

Reviewer Point P 3.26 — Line 1084 Salter et al. (2019) showed that principal components can lead to spurious model falsification when history matching and demonstrated the idea with the Canadian AGCM.

Reply: We'll add that reference to the list of potential limitations of using principal components that is provided in that paragraph.

Reviewer Point P 3.27 — Line 1163 "emulators introduce a probability distribution for simulator output". Is this accurate in a History Matching context?

Reply: Given that the term has no single clear definition in this regard (and that the offending statement doesn't hold for *e.g.,* Bayes Linear approaches), we've removed the offending sentence.

Reviewer Point P 3.28 — Line 1169. Salter and Williamson (2016) demonstrated why early waves should not use pure linear regression when the intention is to use Gaussian Processes for later waves (not only do you cut less space for the price of the model runs, but you often can't recover the losses over waves).

Reply: This depends on how much of variance is captured by regression/global part of the emulator versus the GP part. If most of variation is in global term, one can speed implementation up a lot by just using regression, especially for early waves. So this is an efficiency question. This is now made clear in the text.

Reviewer Point P 3.29 — Line 1180. I think this is false. The residual stochastic process is not a PDF and it does not necessarily have a pdf unless you give it one (many history matching papers don't).

Reply: Fair enough in general. As this section is meant for mathematically competent readers, we have now replaced the description with a more precise definition of a second order stationary stochastic process.

Reviewer Point P 3.30 — Line 1183. I suggest that the word discrepancy is reserved for structural error and shouldn't be used in this context.

Reply: Agreed but no ideal replacement. Replaced with "predictive error".

Reviewer Point P 3.31 — Line 1224. This paragraph seems to have been said multiple times in the internal discrepancy conversation and so might be cut (you might also reference the papers suggested above that have actually attempted this for paleo models).

Reply: The point of the first two sentences "Emulators are not just used to represent deterministic simulators. Emulators can also emulate models that have stochastic components" is not clearly stated before (only implicitly). Furthermore, the previous mentions of using emulators to represent aspects of internal discrepancy are only in passing with reference to this subsection. So we've chosen to retain (especially given that most intended readers of this section will be new to emulators).

Reviewer Point P 3.32 — Lines 1311 - 1316. I think the major problem with this line of reasoning is that the NROY sets for the non-coupled are not necessarily the right sets for the coupled model

Reply: Our stated reasoning doesn't make the above assumption. It only asks "how can the existing emulators and NROY sets be most efficiently used to facilitate history matching of the coupled simulator?". Or does the reviewer believe there is never any useful information to glean from the uncoupled emulators and NROY sets? If there is no useful information, the referee's previous mention of an increasing focus on processed-based history matching for GCMs would be hard to justify. However, we do not mean to imply there will always be a way to do so, so to avoid that interpretation we now state: "can the existing emulators and NROY sets be used to facilitate .."

Reviewer Point P 3.33 — Lines 1358-1364. There is already an extremely active emulator development community and comparison of emulators has also received a lot of treatment. However, though the community considers linear models, Gaussian processes and comparison of kernels, polynomial chaos and more recently deep Gaussian processes, the community does not view Bayesian Neural Networks as a competitor. If the authors are particularly interested in "hybrid NN GP emulators", they might start with the existing GP representations of Bayesian Neural Networks (A NN is a finite approximation to a GP and you can write the kernel of that GP down), or look at the Neural Network kernel for GPs.

Reply: The most recent SIAM conference on UQ had a number of Bayesian Artificial Neural Network (BANN) talks. LT has used BANNs for nearly two decades as emulators (*e.g.,* Hauser et al., 2011). And artificial neural networks are still being used by others for emulation (*e.g.,* Rosier et al., 2022) in the Earth system's modelling field. So it remains unclear to us whether BANNs have no useful role in the field. However, dropping explicit mention of BANNs from this paragraph doesn't detract from our intent, so we have now done so.

Reviewer Point P 3.34 — Line 1374-1377. Would also be good to see some actual theory for or examples of the quantification of internal discrepancy through designed experiments and some case studies of it working on a GCM class model.

Reply: Agree. LT is writing up internal discrepancy results for Antarctic and Greenland paleo ice sheet models but we aren't aware of any yet for the climate modelling context. Which again is one of the motivations for this paper.

References

Andrianakis, I., McCreesh, N., Vernon, I., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G.: Efficient History Matching of a High Dimensional Individual-Based HIV Transmission Model, SIAM/ASA Journal on Uncertainty Quantification, 5, 694 719, https://doi.org/10.1137/16m1093008, 2017.

Bas de Boer, F., Golledge, N., and DeConto, R.: Paleo ice-sheet modeling to constrain past sea level, PAGES Magazine, 27, 20, https://doi.org/doi.org/10.22498/pages.27.1.20, 2019.

Capron, E., Rovere, A., Austermann, J., Axford, Y., Barlow, N. L., Carlson, A. E., de Vernal, A., Dutton, A., Kopp, R. E., McManus, J. F., et al.: Challenges and research priorities to understand interactions between climate, ice sheets and global mean sea level during past interglacials, Quaternary Science Reviews, 219, 308–311, 2019.

Cumming, J. A. and Goldstein, M.: Small Sample Bayesian Designs for Complex High-Dimensional Models Based on Information Gained Using Fast Approximations, Technometrics, 51, 377 388, https://doi.org/10.1198/tech.2009.08015, 2009.

Goldstein, M., Seheult, A., and Vernon, I.: Assessing Model Adequacy, in: Environmental Modelling: Finding Simplicity in Complexity, edited by J. Wainwright and M. Mulligan, pp. 435–449, John Wiley & Sons, 2013.

Hauser, T., Keats, A., and Tarasov, L.: Artificial neural network assisted Bayesian calibration of climate models, Clim. Dyn., pp. 1–18, 10.1007/s00382-011-1168-0, 2011.

Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., Musat, I., Fairhead, L., Diallo, F. B., and Volodina, V.: Process-based climate model development harnessing machine learning: II. Model calibration from single column to global, Journal of Advances in Modeling Earth Systems, 13, e2020MS002 225, 2021.

Rosier, S. H. R., Bull, C., and Gudmundsson, G. H.: Predicting ocean-induced ice-shelf melt rates using a machine learning image segmentation approach, The Cryosphere Discussions, pp. 1–26, 2022.

Sexton, D. M., McSweeney, C. F., Rostron, J. W., Yamazaki, K., Booth, B. B., Murphy, J. M., Regayre, L., Johnson, J. S., and Karmalkar, A. V.: A perturbed parameter ensemble of HadGEM3-GC3. 05 coupled model projections: Part 1: Selecting the parameter combinations, Climate Dynamics, 56, 3395–3436, 2021.

Tran, G. T., Oliver, K. I., Sóbester, A., Toal, D. J., Holden, P. B., Marsh, R., Challenor, P., and Edwards, N. R.: Building a traceable climate model hierarchy with multi-level emulators, Advances in Statistical Climatology, Meteorology and Oceanography, 2, 17–37, 2016.

Wang, C., Soden, B. J., Yang, W., and Vecchi, G. A.: Compensation between cloud feedback and aerosol-cloud interaction in CMIP6 models, Geophysical research letters, 48, e2020GL091 024, 2021.

Zhu, J., Poulsen, C. J., and Otto-Bliesner, B. L.: High climate sensitivity in CMIP6 model not supported by paleoclimate, Nature Climate Change, 10, 378–379, 2020.