Note to editor

It is unfortunate that reviewers 1 and 3 did not adequately take into account that they are not part of the target audience (nor fully take into account the reader road-map table 3) when writing their reviews. The only reviewer with a paleo background (Evan Gowan, reviewer 2), placed more attention in their review on their own work and previous papers by LT's group and has largely missed the key points of this submission. Reviewer 3 is the most favourable though appears to want this for a broader audience given their statement that the choice of examples "seems like an attempt to conform to the scope of Climate of the Past".

The core issue we address is what is needed to meaningfully relate model results to the actual real world and thereby make inferences about past ice and climate evolution that the reader can have confidence in. This is fundamental to a good part of the paleo modelling enterprise. It is non-trivial and given the breadth of relevant technical understanding within the paleo community, this is not something that can be accessibly and adequately addressed in ten or twenty pages. The more than 400 downloads of our original submission to Climates of the Past attests to interest in the community in what we present.

The length issue could be addressed by making our submission into a monograph, but the importance of the above issue warrants a more accessible forum. LT thought Climates of the Past would be more appropriate.

General response to the reviewers

What struck the authors is that none of the reviewers addressed a core and logically self-evident message of our submission: As a modeller interested in making inference about actual past ice sheet and/or climate evolution, you need to specify what the relationship is between your model results and the real world and why the reader should have confidence in your assessment of this relationship.

Reviewers 1 and 3 raise the main issue of length, but present somewhat opposing solutions. Neither reviewer seems to consider that this paper is not meant for either of them (except for subsection 4.7 : a community methodology research and development agenda). The abstract clearly states: "This overview is intended for all interested in making and/or evaluating inferences about the past evolution of the Earth system (or any of its components), with a nominal focus on past ice sheet and climate evolution during the Quaternary". LT has run drafts of this paper by a number of grad students (both in and outside of LT's research group), as well three colleagues in the paleo field. The length is in part due to addressing suggestions and difficulties in comprehension raised by those students and by a reviewer of a previous version.

The reviewers also fail to take into account that for many in the target audience, the length going by the reader road-map (table 3) is only a fraction of the whole paper.

The reviewers and editor need to consider, not whether this is a good read for statisticians, but whether this will provide a useful reference for at least some of those within the stated target audience. LT has in good part pursued this project as a document he would have liked to have had 20 or even 15 years ago. Reviewer 1 talks about the need for evidence-based presentation. Then as evidence to the utility of and interest in this document, consider the more than 400 recorded downloads of the original submission to Climates of the Past. This download count is surpassed by only one of the five 2023 highlighted (ie not just published) papers on the Climates of the Past home web-page.

Yes, we could boil down the underlying message of the paper to 1 paragraph: "For the contexts of making inferences about a physical system, one needs to appropriately specify the relationship between one's model(s) and the system and between one's observation or experimental data and the physical system. Without this specification, the inference, whether it is a hypothesis test, posterior (*i.e.* Bayesian) inference, or whatnot, is by definition uninterpretable. These relationships are respectively the model and data uncertainties." However for the majority of our target audience, the above will have little meaning. Furthermore, "Bayesian" is starting to become a buzzword in paleo-modelling, but many in the field do not have the conceptual basics to critically evaluate what the results of some self-professed Bayesian modelling study means or if it has any interpretable meaning for the stated intent.

In support of the above concerns/motivation, we point the editor and reviewers to a current The Cryosphere submission: "Quantifying the Uncertainty in the Eurasian Ice-Sheet Geometry at the Penultimate Glacial Maximum, Pollard et al." (https://tc.copernicus.org/preprints/tc-2023-5/) that does a far from complete uncertainty assessment contrary to the claim of the title. This is especially evident in the contrast between their adhoc model uncertainty estimate (simply a chosen fraction of ensemble variance) and their plotted misfit of a test "history matching" to an old glaciological model (one of the GLAC1-D chronologies) for last glacial maximum. Furthermore, the abstracts states "We perform Bayesian uncertainty quantification", even though Bayes rule is never invoked.

Instead of reviewing the submission, reviewer 2 (Evan Gowan) focused on presenting their own work as well as arguing against a number of points that are in fact opposite to what we make in our submission. Their review also make a number of erroneous claims concerning two papers from LT's group (one of which is over a decade old). Nevertheless some of their arguments offer a useful foil for expanding our key points. Gowan's whole review also provides a clear example of why we feel our submission is important for the paleo community.

In the following we address reviewer points.

Reviewer 1

Reviewer Point P 1.1 — Despite the want for this paper, I found it long, dense and difficult to follow, largely containing opinions rather than evidence-based guidance, and ignorant of the existing statistical literature.

Reply: After 4 months of not looking at the submission, LT has gone through a complete read and edit. He has cleaned up a few spots where reading stumbled, but did not find it "difficult to follow". The reviewer seems to expect this paper to be a review of relevant statistical litterature, it is not. It would also help if the reviewer provided some example evidence to back their claim of "largely containing opinions". The appendix offers a suggestion (not prescription) for how to address a number of the issues raised based on the authors' experience from working on these issues for decades. Perhaps one is free to call experience "opinion". We do not see anything in the main text that can be called an "opinion".

Reviewer Point P 1.2 — This paper has been previously submitted for publication (a public process in Climate of the Past), and previously reviewed. I have reservations that many of the previous reviewers' comments have not been sufficiently addressed, and find myself agreeing with the previous reviewers on many points.

Reply: The previous version of this paper that was submitted to Climate of the Past was heavily revised in the second stage review, but was rejected by the editor (given the paper length) without going back to the reviewers. Given our extensive response (24 pages mostly in response to the 10 page review by Danny Williams) to the original reviewers (https://cp.copernicus.org/preprints/cp-2021-145/cp-2021-145-AC2-supplement.pdf) along with the edits, it would help to have specific examples of which of the reviewer comments were not adequately addressed.

Reviewer Point P 1.3 — 22 pages are spent introducing facets of the Bayesian framework after which there were a short(er) 4 pages on some useful techniques and then 7 pages of, from what I can tell, opinions with no examples.

Reply: We are curious about what the reviewer see's as "opinions". Section 4.1 states "A key step to towards encouraging rigorous uncertainty assessment about earth system evolution would be for modellers, editors, and reviewers to ensure the following questions are answered: What model parameters are adjusted and to what extent is the criteria used for their selection explicit and appropriate? To what extent does the chosen set of constraint data capture what is available and relevant? To what extent have the associated uncertainties been explicitly accounted for?..."

Are the above all just opinions that any modeller can defensibly choose to ignore when making claims about inferring past ice sheet or climate evolution? Or taking the more clear "opinion" subsection 4.7 a community methodology research and development agenda. Yes there is judgement involved in the selection of agenda items. Is the reviewer equating opinion with judgement?

Reviewer Point P 1.4 — If we're just going to history match why do we need MCMC and posterior predictive distributions? In fact, why do we need Bayes rule at all?

Reply: Even if we are history matching, we still need a conceptual framework for understanding what history matching is and isn't and what by contrast a meaningful Bayesian inference entails. "Bayesian" is starting to become a buzzword in the paleo modelling community, but unfortunately usually as a stated claim not reflected in the actual analysis given the limited uncertainty assessment and often limited sampling.

History matching also requires emulators which currently are generally either Bayesian or Bayes Linear.

Reviewer Point P 1.5 — Writing and grammar. As a whole I find the paper to be quite sloppy. To name just some of my concerns, title cases are off, table formatting is inconsistent, there are erroneous parenthesis, the figure fonts are massive, and Tony O'Hagan needs an apostrophe. Some grammatical and editorial errors are inevitable, but I would expect a more thorough edit to be conducted before the submission of a journal article (especially before the second submission). I would also check your usage of colons throughout, the clause preceding the colon should be a complete sentence: see, for example, line 395. Also, the authors' colon spacing (with a space preceding the colon, and only sometimes) is foreign to me and I can't see a similar example in either US or UK style guides. Finally, the writing reads like a series of dot-points that ramble on rather than with any real narrative or structure. I keep finding myself lost and have to remind myself of what is happening and where am I going.

Reply: Figure fonts are "massive" on purpose. Contrary to the attention placed on colour-blind accessible colour maps, the usage of difficult to read small fonts seems to get no attention. For a description of writing and grammar being "quite sloppy", the listed issues seem relatively minor and are easy to fix (and have now been fixed).

Reviewer Point P 1.6 — Literature. There is a lot of literature in statistics on calibration of computer models, prior selection and elicitation, emulation, modelling simulator discrepancies, as well as accessible introductory texts on Bayesian analysis. Not much of this is cited. Some of the more statistical texts are potentially inaccessible to the uninitiated;

Reply: This submission is not meant to be a comprehensive review on the topic. We have carefully chosen a limited set of references to provide the interested reader with no Bayesian stats background a starting point into the litterature. This is an obvious difference in judgement (or "opinions") between the reviewer and the authors. LT, for instance, when trying to grasp a new methodology, would rather start with a targeted list of refs instead of a complete and overwhelming list of the current litterature. We would welcome suggestions of more appropriate choices of citations with the above in mind.

Reviewer Point P 1.7 — I also feel the paper would benefit by acknowledging, in text, what applied work is being conducted, what is it doing well, and what is it missing (with appropriate citations). A dense table of citations is hard for the human brain to process, and despite my best efforts I still don't have an appreciation for what the authors are trying to say in the tables.

Reply: For eg table 1, we feel this is clearly spelled out in the text: "To provide some motivation for this overview, and the chosen focus on a paleo context, it is worth considering glaciological model reconstructions of past ice sheet evolution (Table 1). To date, none have adequately addressed relevant uncertainties. Furthermore, the number of ensemble model parameters varies widely (from 2 to 39), as does the range and quantity of data constraints. No published studies after 2014 have used more than 5 ensemble model parameters, raising concerns about methodological progress in assessing uncertainties within the paleo ice sheet community".

Reviewer Point P 1.8 — 4. Lack of examples. This paper is quite dogmatic about the need to do a seemingly arbitrary sub-set of modelling stages but does not provide an example of it actually being done. Surely the authors have a more relevant toy simulator (that is not a linear model) that can be used to discuss the statistical concepts and also provide a helping hand to the struggling reader. In reference to point 2, adding an example that traverses the paper will add narrative and continuity.

Reply: The reviewer's assessment contrasts with that of reviewer 3 who states "throughout the manuscript the authors incorporate numerous examples". The reviewer is forgetting the technical breadth of the audience. Reviewer 1 for our original submission, though an internationally accomplished glacial geologist, never had a University level math course. Going beyond a linear model would limit accessibility to a part of our target audience, and for what gain given the conceptual intent?

As to the suggestion of an example that traverses the paper, going beyond the conceptually very accessible linear regression example would entail a major step up in details, the simplest contextually relevant model would probably be a shallow ice approximation flow-line model. But that will already involve a number of details (mass-balance forcing, basal drag, basal topography, temperature forcing) that will result in a longer paper, with potentially reduced accessibility for part of the audience. For the part of the audience interested in detailed implementation, such a paper is currently being written up by LT for the context of approximate ice sheet history matching.

Reviewer Point P 1.9 — Feasibility. I have a fundamental problem with the feasibility of some of the methods recommended. In my experience climate models are not computationally cheap to run, although I admit that the authors have more experience than me here and so could perhaps provide run times of certain simulations that are valuable to the community. In the internal discrepancy section the authors recommend that the parameter space is effectively explored and for each of these points the boundary conditions are sufficiently explored to accurately calculate potentially quite large variance covariance matrices.

Reply: The comment reflects a limited understanding of the range of climate and earth system models as well as the computational cost of ice sheet models. LT's 3D glacial systems model (GSM) can run a full glacial cycle for Eurasia at roughly 50 km grid resolution on a single commodity compute core in about 5 hours (depending on ensemble parameter values using a cluster that is over 7 years old). LT's group is also running the fully coupled LCice (LoveClim EMIC and GSM) that can run about 1500 years in 24 hours on a single core. As already cited in the submission, LT's group explored efficient approximate Bayesian calibration of reduced complexity general circulation climate models back 2011. The core issue is something that is already raised in the paper. There are design and resource choices being made to focus all efforts and resources on the latest most complex and highest resolution climate models that can be run over the intervals of interest. There is much less effort put towards using/building somewhat simpler climate models with low enough resource requirements to enable more than trivial uncertainty assessment.

Reviewer Point P 1.10 — The process of (1) exploring the parameter space in the order of millions of times, (2) exploring the spatio-temporal boundary conditions at each of these millions of locations, and (3) predicting and calibrating from these models seems computationally demanding in the extreme.

Reply: Nowhere do we state a step (2) above for each location in the parameter vector sampling. Either the parameter vector includes components to specify the boundary conditions and/or the rest of the impact is addressed via internal discrepancy assessment. And yes, this is computationally demanding, thus our suggestion that history matching would be a first major step (and overall a more appropriate tool for this context) for which the uncertainty assessment is much simpler.

Reviewer Point P 1.11 — Further, the boundary conditions are generated from "adding appropriately correlated noise" – in my personal experience this process is not simple and to cast it as such is wrong. Accurately representing, modelling and quantifying uncertainty for many spatio-temporal processes is exacting on even the most seasoned statistician and I think at least a nod to this should be included

Reply: The reviewer is mis-representing the sentence by not quoting the critical prior "could be created by" in the sentence, making clear this is not prescriptive (though something like this would be needed for full Bayesian posterior inference). We will add some simpler options if developing more limited internal discrepancy error models for history matching.

Reviewer Point P 1.12 — Later in the paper, emulators are thrown into the mix with no explanation or introduction. Do your methodologies need emulators, and if so where and why?

Reply: "Later" is a bit disingenuous, as emulators are first introduced right after the scale of the sampling problem is indicated: "In the first author's own experience with ice sheet model calibration of approximately 40 ensemble parameters, at least order ten million point sampling is still required (as compared to the astronomical 10^{40} for a simple grid search over deciles). As this is still beyond computational tractability for ice sheet and climate models, one other component is required. This component, a set of emulators, consists of very fast approximate statistical models that predict statistical characteristics of simulator output of interest as a function of an input parameter vector."

Reviewer Point P 1.13 — Point of the paper. I feel the paper tries to do too many things, and so falls short on each of them. It attempts to provide textbook level mathematics on fundamental statistical principles, ground said mathematics in the application, and then provide recommendations and guidance. I know that the authors have tried to write a document that is accessible to a non-mathematical audience, but by taking the middle ground and then trying to teach mathematics it distracts from the rest of the paper. I would recommend removing as many equations as possible and instead focusing on what the point of the mathematics is. The equations and the rigour can instead be deferred to a supplement for people to read once they've decided that it is worth the time and burden to learn and implement probability theory, MCMC, history matching, and the other number of techniques mentioned.

Reply: We have already kept the math to the minimum and unlike as implicitly suggested by the reviewer, no equations are provided on MCMC. One intent was for the mathematical competent reader to understand how a Likelihood can be specified and we do not see how this could be done without the minimal number of equations. Again the reviewer is pointed to the Table 3 road-map. The choice of target audience and submission goals should be that of the authors, not the reviewer.

Reviewer Point P 1.14 — Unfortunately, I do not think that the current version of this manuscript does a good enough job. I recommend a significant re-write that prioritises the point of the mathematics, and not equations for equations' sake.

Reply: Instead of broad swipes, it would be more helpful if the reviewer gave concrete examples of where he/she thinks we are using "equations for equations' sake".

Reviewer Point P 1.15 — I recommend ... that has narrative and verve rather than 40+ pages of dense writing; and that leads by example with some demonstration of feasibility so that we are motivated to follow.

Reply: We may strongly differ on what narrative and verve mean and where it's appropriate.

Reviewer 2

As detailed below, reviewer 2 (Evan Gowan) is using their "review" to espouse their own methodology, make numerous erroneous claims, and has fundamentally missed the point of the submission. Our submission does not advocate full Bayesian inference for the stated context in at least the near term. The submission presents Bayesian inference as a conceptual framework for making inferences that we argue is the probabilistic underpinning of the classical scientific method. This presentation is in part so that the reader can understand that Bayesian inference without robust uncertainty assessment has limited to uninterpretable meaning.

Reviewer Point P 2.1 — Tarasov and Goldstein propose that paleo (specifically ice sheet) modellers and data practitioners should be making use of Bayesian analysis in order to ensure that everything has an error uncertainty attached to it.

Reply: The reviewer has seemed to miss a key point of the paper. We are not advocating determination of Bayesian posteriors anytime soon. We do state, and it follows by the definition we provide, that a scientific inference with no meaningful uncertainty assessment has no usefully interpretable meaning about the actual physical system. That assessment might be trivial depending on the context but this is generally not the case for paleoclimate and paleo ice sheet modelling contexts.

Reviewer Point P 2.2 — They suggest that most ice sheet modelling exercises do not provide what they think is an adequate assessment of uncertainty, and that the only way forward is to run hundreds or thousands (or even millions, line 586) of model simulations to create an uncertainty range.

Reply: The reviewer ignores the subsequent relevant lines starting at 590: "As this is still beyond computational tractability for ice sheet and climate models, one other component is required. This component, a set of emulators, consists of very fast approximate statistical models that predict statistical characteristics of simulator output of interest as a function of an input parameter vector."

Reviewer Point P 2.3 — In the first part of the paper, they go over what Bayesian inference is (using a strange and poorly explained example (line 141),

Reply: We are not clear what is "strange" about the conditional probability example (which is not a Bayesian inference example as the reviewer erroneously states). Nor do see what needs to be explained, given that this is just an example of the precisely stated definition of conditional probability in the previous sentence, sic: "The expression $P(A = a|B = b)$ denotes the conditional probability of the variable A having some specific value "a" if the statement that the variable has some value "b" were true".

Reviewer Point P 2.4 — Dr. Tarasov has made it very clear to me (and probably many other paleoclimate scientists) that any study that does not involve a full Bayesian uncertainty assessment is not worth doing.

Reply: Dr. Gowan is misrepresenting what LT has communicated. Furthermore, bringing in allegations and/or confused past perceptions instead of the concrete task at hand of addressing what is in the submission, is very unprofessional.

Reviewer Point P 2.5 — It is thoroughly unapproachable to anyone who is unfamiliar with advanced statistics (and maybe even to those with such training, judging by the other reviewer's comments).

Reply: That claim doesn't match the response LT has received from a number of grad students and a couple of colleagues in the paleoclimate and paleo sealevel fields who've gone through earlier drafts.

Reviewer Point P 2.6 — I want to use this opportunity to provide an alternative view to what Tarasov and Goldstein are propos- ing, because I disagree with the idea that Bayesian uncertainty assessment is needed in every paleoclimate problem.

Reply: We are not clear what the reviewer means by Bayesian uncertainty assessment. If they mean Bayesian posterior inference, than as already been made clear in the general remarks to reviewers, this is clearly not what we are proposing.

If the reviewer means full accounting for uncertainties, then we would like to know where the reviewer disagrees with the following chain of logic presented in the first paragraph of the conclusions: "This review started from the premise that a defining feature of any aspect of science which is concerned about making statements about the real world is the rigorous quantification of uncertainties. Within the context of computational models, this claim can be easily supported if one recognizes that uncertainty assessment is simply the principled assessment of the relation of model results to the physical system. Without robust uncertainty estimation or, at the very least, a more limited mix of quantitative and qualitative assessment by the modeller, the reader has no basis to interpret the relevance of modelling results to the actual physical system. As our toy model illustration demonstrated above (c.f. Sect. 2.5), ignorance of structural uncertainties will generally result in model predictions that do not intersect the physical system within computed prediction limits, even if Bayes Rule is used for the inference."

Reviewer Point P 2.7 — The First point I will raise is that if you read this commentary, Tarasov and Goldstein make it sound like they are the only people who are using Bayesian analysis in paleoclimate applications.

Reply: The reviewer makes a derogatory and misguided allegation without any substantive evidence (eg example text) to back up the statement. First, we are not advocating Bayesian inference for at least near term paleo modelling work. Furthermore, if the reviewer were to look at Table, 1, only 1 of the two cited papers from LT's group are indicated as having any structural uncertainty assessment with the added qualification of "limited".

Reviewer Point P 2.8 — The code for the Glacial Systems Model that Dr. Tarasov uses for his ice sheet modelling and Bayesian analysis exercises have never been made publicly available, despite the fact that the primary paper describing the application of it was published over a decade ago (Tarasov et al., 2012). In my opinion, it is arrogant for Tarasov and Goldstein to say with regards to paleo ice sheet modelling efforts to date none have adequately addressed relevant uncertainties (Line 71)", when they have not contributed their own programs that, according to them, are the only way to evaluate model uncertainties.

Reply: It is also unprofessional for the reviewer to make claims that cannot be backed up by relevant text, in this case "their own programs that, according to them, are the only way to evaluate model uncertainties". Where do we make such claims in the text?

As for the Glacial Systems Model code, the model has heavily evolved over the last decade (changelog since 2015 has currently 21273 lines). Such ongoing changes would have made it a pain for anyone outside of LT's group to use. LT has been working on making the model easier to use and port, as well as writing an associated paper that will document the model. A bare-bones version of the GSM for idealized configurations is already available in a publicly accessible archive (cf assets for Hank et al. https://egusphere.copernicus.org/preprints/2023/egusphere-2023-81/). The paper with a full code archive for paleo ice sheet modelling is nearing completion for submission.

As for the statistical tools, we already cite a new freely available history matching code suite in the text.

Reviewer Point P 2.9 — As such, can a test truly be made to demonstrate whether a model is right or wrong?

Reply: The Popperian framework doesn't include testing to demonstrate a model (in the general sense) is right, it only includes falsifiability. And yes, models in the form of theories (which generally assume 0 uncertainties for a stated context), such as Newtonian Gravitation, can and have been falsified for specific contexts. If the reviewer is restricting their concept of "model" to computational models ("simulators" within the uncertainty quantification community), this does not change as both theories and computer models are often used as approximate representations of the physical world around us. The whole history matching approach (described in section 3 of the submission) is based on falsification ("ruling out") of trial models within a chosen statistical threshold (and this can be done in milliseconds not "centuries").

Reviewer Point P 2.10 — Tarasov and Goldstein argue that with the usage of Bayesian analysis that it is possible to :. go confidently well beyond storytelling but, as detailed below, only when all uncertainties are rigorously addressed and assessed" (line 37-38). But as highlighted above, this is impossible, because an ice sheet model is an approximation of the real world and will always have some level of uncertainty that will be impossible to statistically model.

Reply: Any model is an approximation of the real world. So, as we spell out in the paper, physical world = model + uncertainty. If you are advocating we can ignore uncertainty (or some component of it), then you are either saying uncertainty is negligible or that you are not concerned about relating the model to the physical world.

To simplify this discussion, let's focus on history matching (which we are advocating as to where efforts should be focused the next while). All history matching needs is uncertainty quantification that brackets the relationship between model and reality within a chosen confidence range (being loose on the term "confidence"). And that is doable.

Reviewer Point P 2.11 — The usefulness of ice sheet modelling therefore comes not from being able to predict some geologically constrained reality, but rather from their utility in storytelling (that is, to pose problems) to help explain our world.

Reply: We do not denigrate hypothesis creation (aka storytelling). We are challenging claims about actual past ice or climate system evolution based on models for which the uncertainty relationship between model and physical system is not meaningfully (within a scientific context) specified.

Reviewer Point P 2.12 — The consequence of this problem in terms of a Bayesian calibration procedure that Tarasov and Goldstein are proposing, is that the resulting error bars are going to be strongly biased to the climate model that they are using. This problem is likely so severe that I question the usefulness of this exercise.

Reply: Anyone looking at the divergence of climate sensitivity or the precipitation errors of current PMIP 4 models will already have a sense of the large uncertainties even in current state-of-the-art earth system models. So again, what value are the models in relation to inferences about past physical system evolution without some kind of uncertainty assessment specifying the relationship between the model and actual physical system. As such, the reviewer's point repeats some of the arguments we present for the need for robust uncertainty assessment.

As an aside, instead of blind attacks, the reviewer would benefit from actually reading some of LT's modelling papers in which it is made quite clear he does not use a single climate model and that the majority of order 40 ensemble parameters for each paleo ice sheet that his group has worked on are there to try to address the uncertainties in climate forcing. LT is curious if the reviewer has complained about the large majority of published paleo ice sheet modelling papers that only use a handful of ensemble parameters?

Reviewer Point P 2.13 — In this paper, they apply their Bayesian analysis scheme on a model that couples their Glacial Systems Model with a climate model of intermediate complexity.

Reply: Again the reviewer is mis-representing the paper. Even a simple text search of "Bayesian" in that paper will come up blank. No Bayesian analysis is attempted in that paper and such mis-representation is unprofessional.

Reviewer Point P 2.14 — In their current state, ice sheet models do not have the predictive power to precisely reconstruct ice sheet history.

Reply: Nor will they have such ability for precise reconstruction in the foreseeable future, thus the need for structural uncertainty assessment.

Reviewer Point P 2.15 — The story-line of this paper should have been that it demonstrated that it is possible to grow an ice sheet rapidly after inception started, and that the Eurasian and North American ice sheets have different sensitivities to external forcing. That, to me, would be a more useful application of this modelling exercise, as it tells us something interesting about the climate system without over interpreting the results in terms of how the ice sheets incepted.

Reply: The "should have been" story-line is part of the actual story-line as eg the conclusions of the cited paper state "The EA ice sheet is more sensitive to orbital forcing and ensemble parameter values".

Reviewer Point P 2.16 — The Bayesian analysis technique generally uses Latin hypercube sampling in order to select the values of the parameters that are varied in the modelling experiments (lines 1164-1168)

Reply: Incorrect as stated. Is the reviewer conflating MCMC sampling via emulators with the exploratory Latin Hypercube ensemble described in 1164-1168? Furthermore, the example history matching framework in A1 is not Bayesian, though it can be used as a "stepping stone towards a complete Bayesian inversion".

Reviewer Point P 2.17 — This presents what I consider the biggest weakness of large parameter studies that are used for Bayesian analysis. In a large ensemble with many parameters varied in tandem, it is not possible to clearly tell why one simulation fails while others succeed. This makes storytelling difficult, and reduces the utility of the model to tell us something about the behaviour of the ice sheet. It is much easier to tell a story with an ice sheet model by holding most variables as constant, and varying a small number of variables in a controlled way. In that way, we know exactly how sensitive the outcome of the experiment is to a parameter.

Reply: The reviewer is conflating different research goals into a single method. If the modeller is aiming to understand the role of specific parameters, than sensitivity analysis is an appropriate tool. If the modeller wants to understand process interactions, then sensitivity experiments are appropriate. But if the modeller is making a claim about last ice sheet or paleo climate evolution, then one needs to address uncertainties, and given the complexities of the paleo ice and climate system, a handful of model parameters makes that very hard to do.

Reviewer Point P 2.18 — 6 In modelling, there is always subjectivity

The main appeal of applying the Bayesian analysis approach for ice sheet modelling is that it gives the illusion of objectivity by assigning a probability value to every decision (section 2.1). Since the experiment allows the parameters for each ensemble member to be selected at random from the experimenter's probability range, it removes responsibility of the outcomes from the experimenter. For practical reasons, there is always going to be a certain limits to how wide of a range of values that can be used in a Bayesian analysis modelling study. I will use the North American deglacial study by Tarasov et al. (2012) as an example.

Reply: Again, we wonder if the reviewer has actually endeavoured to fully read our submission. Unlike conventional frequentist statistics, Bayesian inference is much more explicitly cognizant of the judgement aspect of statistical inference. In section 2.1 we state "No matter what interpretation of probability one chooses, the assignment of probabilities require judgements. To be testable and potentially falsifiable, these judgements must be made and treated in a rigorous and self-consistent way." In section 2.3 we also state "This judgement aspect has often been a target by critics of Bayesian approaches, with a usual focus on the specification of the prior. This focus has no clear justification as judgements are required for all aspects of the inferential process and not just the 225 initial specification of the prior. But this holds true for any statistical inference including those by frequentist approaches".

Reviewer Point P 2.19 — In Tarasov et al. (2012), they varied 39 different components of the ice sheet model, ...

Reply: At this point, the reviewer's obsession on a more than 10 year old (2012) paper is getting obnoxious, especially when some relevant points are being mis-represented. The GLAC-1D product of that paper is deprecated. And the stated limitations the reviewer raises in this section about GLAC-1D (limitations which were raised explicitly in that old 2012 paper) as well as a number of others are being addressed in ongoing work. LT's struggles with confident Bayesian inference are one reason the current submission recommends history matching or variants thereof.

Reviewer Point P 2.20 — 7 Using the right tool for the job

In my methodology, I use an ice sheet model with perfectly plastic ice and assumes the ice sheet is equilibrium (Gowan et al., 2016a). This model requires just three inputs: the ice margin at a specified time slice, a model of the basal shear stress, and GIA deformed topography.

Reply: The claim of "just 3 inputs" is highly misleading. Each of those inputs is a field (varying in space and time) that has associated uncertainties. Furthermore, the static perfectly plastic approximation is not at all appropriate for any region with ice streams (much of the southern North American ice sheet margin sector, Hudson Strait,...). The static equilibrium approach that the method entails raises questions of what predictive/retrodictive value it has especially for times prior to last glacial maximum when available paleo constraints are very sparse in time and space. How are the errors from the large approximations addressed?

Reviewer Point P 2.21 — When my reconstruction was tested with a lake filling algorithm, it successfully captured the geometry Lake Agassiz during the periods that had strandline data (Hinck et al., 2020). Only a couple of dozens of iterations were needed to tune the ice sheet reconstruction to fit the strandline data (in combination with many other GIA constraints).

Reply: Since you are making a claim of a "reconstruction", how are you specifying the relationship between your reconstruction and past ice sheet evolution? Or is this just a curve fitting exercise?

Reviewer Point P 2.22 — If you are interested in finding the geometry of Lake Agassiz to make estimates of its volume to say, figure out how much water could have potentially drained out to disrupt Atlantic circulation at the start of the Younger Drays (Broecker et al., 1989), a dynamic ice sheet model would not be an appropriate choice, no matter how many ensemble members are used. The climate forcing and/or calving in the dynamic ice sheet model would have to be manually manipulated to do it, defeating the purpose of the Bayesian framework that Tarasov and Goldstein are proposing.

Reply: So now the reviewer has gone from the topic of uncertainty assessment to the topic of the appropriate choice of ice sheet model. Again the reviewer is missing the point, that no matter what type of model is used, the relationship between the model and the physical system has to be specified. What is the reader supposed to make of the volume of water drained the reviewer generated in their modelling? Is the reviewer claiming it was the exact amount, within 50%, within 5000%? Without that specification and justification thereof, the results are meaningless to most except those who understand the intricacies of the physical system and model uncertainties. Furthermore, for the context of actually inferring the physical runoff, the reviewer's approach necessarily ignores (since it's not computed by their modelling approach), the significant contributions to discharge from ongoing surface melt and meltwater drainage and lacustrine ice calving.

The reviewer also chooses to not mention that Tarasov et al (2012) describe and use a nudging methodology to address the issue they raise in the text block cited by the reviewer "Given the partially lobate structure of the geologically inferred ice margin, as well as the high sensitivity of ice margin location ... it is unlikely that any glacial systems model will ever freely approach inferred margin chronologies to the degree required for accurate modeling of proglacial lakes".

Reviewer Point P 2.23 — It is better to use a model where the margin location is strictly defined.

Reply: Even when the margin location is not precisely known and may have been quite dynamic?

Reviewer Point P 2.24 — 8 The data are never perfect ... For these reasons, I tend to judge the fit of a the data using a simple "consistent or not" metric, because developing an framework for inclusion in a more complicated statistical model is not clear. .. The "consistent or not" assessment has served me well in evaluating my models. A more complex statistical model is not needed.

Reply: It sounds like the reviewer is asking the reader to blindly accept their judgement of "consistent or not" (as well as "served me well") instead of specifying what it actually means. If the relationship between a datum and the actual physical system, can't be quantitatively specified in some meaningful manner, than again what meaning/value does the datum have in the context of the actual physical system? How is another scientist supposed to assess their judgement of "consistent or not"?

Reviewer Point P 2.25 — So, instead, I think it is fine to assess the models in a less rigorous (and definitely less computationally intensive) way rather than attempting to create some probability distribution that is going to be hand-wavey.

Reply: The reviewer is has quite obviously missed one of the key points of our submission : "However implementation of standard Bayesian inference for complex simulators is a challenging and potentially non-robust endeavour....No matter how the sampling is carried out, the result can be highly sensitive to the exact specification of the error model and therefore the likelihood..As such, inferences for say a most likely ice sheet history will have limited meaning contingent on a large set of assumptions. ..For many contexts, a more limited product than a rigorous posterior inference may have adequate 735 utility and can be much more robust".

Reviewer Point P 2.26 — I am confident that the general history of the ice sheet is correct.

Reply: Again, what do you mean by this beyond empty words? How is a reader supposed to interpret what your results mean in the context of the actual last glacial cycle?

Reviewer Point P 2.27 — Although it is stated in this paper that GLAC-1D is an ice sheet reconstruction, that is not strictly true. The ensemble average will not fit any particular observation that the modelling exercise used to evaluate the individual simulations, and it is not glaciologically consistent.

Reply: Again the reviewer is making false claims. The GLAC-1D chronologies provided long ago to the community are from individual ice sheet model runs. The chronologies are identified by the actual run numbers (e.g. nn9927 for one of the North American chronologies) and it is even stated in Tarasov et al (2012): "The single run (nn9927, with detailed plots and tabulated summary characteristics in the tertiary supplement).

Reviewer Point P 2.28 — I'd argue that it would have been better to use this exercise to pick a few simulations that performed well, and make those available.

Reply: That is what was done. But given finite computational resources and depending on the exact purposes of the PMIP intercomparison it might have have made more sense to have some of those resources used for model runs with bounding ice sheet chronology boundary conditions instead of the best fit.

Reviewer Point P 2.29 — Tarasov and Goldstein complain about this (lines 62-69), but I think it would be more productive to be attentive to the realities and needs of other modelling groups rather than lecturing them about the need for ensemble studies to produce an uncertainty range.

Reply: More borderline ad hominem attacks based on false claims as detailed above.

Reviewer Point P 2.30 — 10 Telling a story
Tarasov and Goldstein are of the opinion that storytelling alone is not an adequate way to use ice sheet models (lines 38-48).

Reply: No. That is nowhere stated nor intended. On the contrary we state "Story-telling, or in more usual terminology, hypothesis creation/elaboration, is a central part of science'. Logically one can't test a hypothesis without first creating it. One of our complaints though is the mis-representation of hypothesis creation as hypothesis testing.

Reviewer Point P 2.31 — Most modellers fully understand the limits of what their models can predict.

Reply: The reviewer has already provided a counter example in their claim of inferring Lake Agassiz discharge. We do state in our introduction "It is also our own experience that modellers who know their models well are often the most skeptical about their model results". And no matter what the modeller understands, this doesn't necessarily mean the reader will understand it without clear meaningful communication of the uncertainties involved.

Reviewer Point P 2.32 — 11 Ethical modelling 11.1 Carbon Footprint
Due to the discovery of bugs, I had to run these simulations a few times. Each glacial cycle simulation took roughly one week to complete on 144 processors, while the idealized simulations too between one and two days... Imagine running thousands of model experiments and finding a bug that invalidates the results.

Reply: The logic is getting pretty desperate here. Why run such an expensive glaciological model thousands of times? LT uses glaciological models that span the range of 5 hours (low resolution) to 1 week (high resolution) using a single compute core. Or by the reviewer's logic, is a current generation earth systems model that takes hundreds of compute cores for a single simulation and that takes weeks to months of run time unethical especially given that the model likely includes bugs? Modellers face a trade-off between fewer runs with computational more expensive (and hopefully more accurate) high resolution complex models versus lower resolution and simpler models that enable larger ensembles for the same amount of compute resources. Currently few resources are being applied to the further development of computationally cheaper models and to the development of methods to efficiently synthesize the results of both cheaper models and expensive models.

Reviewer Point P 2.33 — Is it practical to develop a sophisticated statistical and modelling framework under these conditions? Do they think that a student or postdoctoral researcher can afford to wait months or years to get enough results to publish something? This is just not feasible under the current funding regime of science. We can only design modelling exercises that can be accomplished under short time-frames. It would be unfair to students and postdoctoral researchers to be forced into a narrow pathway to accomplish their research with little chance to explore.

Reply: Why does the reviewer assume that this work has to be done within one group as opposed to being a part of a multi-group collaboration or that every student or post-doc in paleo-modelling must endeavour to do say a full history matching for past ice sheet evolution or climate? If a student is interested in the topic, internal discrepancy assessment for some computationally accessible paleo model, for example, would already be a useful and achievable project. And how this issue of time required to carry out fits within the "ethical modelling" section title escapes the authors.

Reviewer Point P 2.34 — From my perspective, there is a still long way to go before dynamic models can reliably be used to precisely reconstruct past climate and ice sheets.

Reply: Again, reinforcing the need for meaningful uncertainty assessment.

Reviewer Point P 2.35 — Using a single, hand tuned reconstruction, I found that the far-field sea level observations could be matched with a smaller volume ice sheet configuration than previously assumed. ...

Reply: With a "single hand tuned reconstruction" and no uncertainty assessment, the reviewer's statement offers no meaningful inference about whether past LGM ice volume was less "than previously assumed". If, for instance, the error induced by the reviewer's modelling approach is 20 m of sealevel, then their finding is consistent with previous findings.

Reviewer 3

Reviewer Point P 3.1 — The authors have a clear aim which is to encourage the paleo modelling community to employ a full Bayesian uncertainty quantication framework in the analysis of their (coupled systems of) computer models, and most importantly making inferences for the real world physical system by jointly incorporating all sources of uncertainty, including structural model discrepancy

Reply: While the second point is correct, the first point is not. As stated at the very top of general response to all reviewers, we want the target audience to understand what Bayesian inference properly entails and means. But for the stated context, this is currently not an appropriate tool. Instead we are suggesting history matching as a tool that can provide robust meaningful inferences.

Reviewer Point P 3.2 — The content is there- fore of greatest use to those less familiar with such methods, although could also be used as a reference by those more experienced in statistical uncertainty quantification.

Reply: Correct, as we spell out in our Table 3 road-map for various audiences.

Reviewer Point P 3.3 — Firstly the length of the manuscript is excessive and would require a committed and motivated reader to finish the paper. In addition, the abstract, introduction and conclusion do not necessarily provide an adequate summary of the main methods and the consequences of (none) implementation.

Reply: We know it's long, but much of this length has come from ongoing efforts to reach the broad target audience. For many readers, our road-map (Table 3) would only entail reading about half of the text. As for the structure, this is obviously not the typical scientific paper, so the role of the abstract, intro, and conclusions are a bit different, especially given the challenge of keeping this paper from getting any longer. And we are not intending to be prescriptive with respects to solutions and approaches, just about issues that need to be addressed for meaningful scientific inference about the world around us. We do feel the series of questions in the concluding paragraph of the conclusions summarizes the latter. We are unclear of what specifically the reviewer believes is lacking from the introduction.

Reviewer Point P 3.4 — For any reader with even a moderate understanding or familiarity of Bayesian statistics, this treatment seems unnecessary. For example, sections 2.1 and 2.2 could be much briefer with the reader referred to an introduction to probability theory undergraduate textbook or online resource where necessary, rather than repeating the material within this manuscript.

Reply: Again we refer the reviewer to our Table 3 road-map. Most of our intended audience has essentially no understanding of Bayesian statistics. And for those that do, the road-map clearly identifies what sections can be skipped.

Reviewer Point P 3.5 — The discussion of MCMC in section 2.7 is not the main purpose of this work, hence it could also be condensed without detriment to this manuscript by referring the reader to other papers, textbooks and resources where desired.

Reply: One aim of this submission was to be as self-contained as possible for the reader not interested in implementation, especially those with limited or no statistical background. Forcing such a reader to jump around to other texts will likely mean that most such readers will stop reading. However, we do agree that MCMC is not a required component for the main text (beyond brief mention) and will either relegate it to appendix or drop it.

Reviewer Point P 3.6 — Throughout the manuscript the authors incorporate numerous examples, both toy models, and from across paleo climate and ice-sheet modelling, but also in relation to present and future modelling efforts. Whilst these provide clarity to the meaning of the statistical framework, in particular, relating this to the analysis of paleo computer models, it also seems like an attempt to conform to the scope of Climate of the Past.

Reply: The examples were chosen to impart understanding to the target audience (all interested in the inference of paleo ice and or climate evolution), which is basically much (if not most) of the Climate of the Past readership. A number of them were added based on efforts by LT to make the text more comprehensible to members of LT's research group (of paleo ice and/or climate modellers). They were not chosen to fit into any journal guideline.

Reviewer Point P 3.7 — In many instances these examples seem unnecessarily long and are often vague in their description without any actual results, for example, lines 399-429, in the discussion of internal model discrepancy.

Reply: We do not see what is vague about the indicated lines 399-429. This is fairly precisely what LT has done in his group in ongoing work that will soon be submitted for publication. Again we remind reviewer, that to date, no one in the paleo ice or climate modelling fields has published an internal discrepancy assessment, and LT (in the paleo modelling field for over 27 years) strongly suspects that most paleo modellers would not know how to do this or even what it is.

Reviewer Point P 3.8 — Greater clarity of exposition could be achieved by taking inspiration from Vernon et al. (2018) and restructuring the manuscript, only using short examples where absolutely necessary. Firstly, I would suggest combining sections 2 and 3 to describe the framework including: prior and model specification; emulation (see comments in section 2.3); uncertainty quantification; history matching and the need for simulations to bound reality; and uncertainty quantification in making predictions/retrodictions.

Reply: We respectfully disagree with this suggestion. Section 2 provides a Bayesian conceptual framework, that the paleo modelling community is far from meaningfully implementing. Section 3 introduces a tractable non-Bayesian framework that can be implemented in current modelling work.

Reviewer Point P 3.9 — The paper should finish by briefly highlighting what further steps are required by the paleo modelling community, rather than the extensive section 4 which seems to repeat many of the points found in sections 2 and 3.

Reply: As mentioned above, parts of section 4 (specifically 4.1 Ensuring uncertainty is addressed in model-based studies and 4.2 Addressing uncertainty in data-based studies) take on some of the more traditional role of the Conclusions section, thus the repetition. None of the subsections in section 4 are long, (with "4.7 a community methodology research 1030 and development agenda" at 1.5 pages being the longest.)

Reviewer Point P 3.10 — Given their importance within the presented Bayesian framework as a fast statistical approximation to the simulator's output(s) for as yet unevaluated paper settings along with a corresponding statement of the uncertainty, I believe that emulators warrant a more detailed exposition within the main body of paper, rather than leaving it to Appendix A6.

Reply: Given that the large majority of the intended audience won't know what a Gaussian process is, we respectfully disagree. Beyond our brief description in the main text, we do not see a useful analogue such as we presented for MCMC that is accessible to the broad target audience. The appendix is meant for those who are specifically interested in implementation, which is a relatively small subset of the target audience.

Reviewer Point P 3.11 — It is not guaranteed that the reader will know what is an emulator, or more specifically, a Regression Stochastic Process Emulator (RSPE) as it is termed in this manuscript. It is therefore necessary to expand on the discussion currently provided in Appendix A6 with further mathematical details of their formulation, as in Vernon et al. (2010), highlighting the range of possible choices to encapsulate the possible model output behaviour, for example, see Rasmussen et al. (2006) for an in-depth discussion of Gaussian Processes (GPs). Within this manuscript there is a focus on emulating outputs one-by-one and treating them independently which for many applications is a suitable and justifiable approach. I would recommend that the authors also reference that there exists numerous multi-variate emulation techniques such as: separable GP emulation (Conti et al. (2010)); the outer-product emulator which extends the separability assumption onto the regression components (Rougier (2008) and Rougier et al. (2009)); non-separable emulators (Fricker et al. (2013)); parallel partial GPs (Gu et al. (2016)); and through the use of basis representations of multivariate outputs (Higdon et al. (2008) and Salter et al. (2019, 2022)). It is unnecessary to provide further methodological details in this manuscript. A further minor point regarding the approach to emulation described in lines 500-502; it would help to provide the name for this technique as multilevel, multi-scale or multi-delity emulation, in order to aid the reader in identifying other literature or code implementa- tions.

Reply: We will add most or all of the suggested references to the revised submission.

Reviewer Point P 3.12 — Practical implementation of the described framework is important and the authors therefore signpost the reader to the recently released hmer R package. It would also be useful to mention the accompanying website, https://hmer-package.github.io/ website/index.html, which provides detailed tutorials, as well as links to published research using the package. For GP emulation, see the RobustGaSP R package, https: //cran.r-project.org/web/packages/RobustGaSP/index.html, and the associated papers (Gu et al. (2016, 2019)). In addition, many within the paleo modelling community use Python as their main programming language, hence it would be of use to suggest similar Python modules such as GPy for GP emulation, https://sheffieldml. github.io/GPy/, which also includes tutorials.

Reply: We thank the reviewer for all the above links and will add them to the text.

Reviewer Point P 3.13 — It would also add to the content of section 2.4 to comment on biases in observational errors and structural model discrepancy and how these should be accounted for within the error model

Reply: Good point, and will do.

Reviewer Point P 3.14 — A comment should also be included regarding any sources of uncertainty exhibiting a parameter dependency, with this being particularly relevant for (internal) structural model discrepancy. In equation 14, the authors should comment on the implications of whether the model M is deterministic or stochastic. This is not immediately clear when going from the first to the second line of this equation.

Reply: We already state "It may well be that the structure of the internal discrepancy is non-Gaussian and/or has significant dependence on ensemble parameters. For both of these cases, a more generalized statistical model is required to represent it (c.f. appendix A6 on emulation)." as well as "Emulators can also emulate models that have stochastic components. When internal discrepancy has spatial, temporal, and/or ensemble parametric dependence, an emulator can therefore offer an efficient representation. Such emulation also may enable the introduction of ensemble parameters (i.e. thereby 1475 subject to history matching and/or posterior inference) to set the structure or amplitude of the stochastic noise used to assess the internal discrepancy."

We will make clear that model M is deterministic in the revisions.

Reviewer Point P 3.15 — The authors include the following overly strong statement about structural model discrepancy: Within most (if not all) scientific disciplines, the tendency to date has been to effectively ignore this source of uncertainty" (lines 340-341). Whilst it is true that model discrepancy is often overlooked (or given minimal treatment) within the (paleo) climate and ice-sheet modelling literature, there exist examples across a range of applications where model discrepancy is explicitly assessed, for example: in cosmology (Vernon et al. (2010)); epidemiology (Andrianakis et al. (2015, 2017)); and in climate modelling (Edwards et al. (2019)).

Reply: Unless the reviewer can provide two scientific fields/disciplines in the natural sciences for which even a tenth of the modelling papers published in the last 6 years adequately address model discrepancy, we stand by the above statement. Model discrepancy is almost always overlooked in the paleo modelling community (beyond perhaps throw-away statements such as "our results are subject to model uncertainties"). For the paleo ice sheet modelling context (for which internal discrepancy assessment is much more computationally accessible), only two past studies to date (as listed in Table 1) have even addressed it to a limited extent, none have adequately.

We will see what useful and relevant examples of internal discrepancy assessment can be cited beyond those that already are (external discrepancy already has relevant cited examples) is useful and we will add a couple of citations.

Reviewer Point P 3.16 — In particular, it is necessary to greatly condense the description of the framework to the core elements, with additional information provided for those interested through supplementary material and via references to other resources that already provide an appropriate coverage of the facet.

Reply: As we describe in our general reply to reviewers, we do not see how the content can be further usefully condensed without losing much of the target audience or making the text even more "dense". We have already placed the material of value mostly for those interested in implementation in the appendix.

Reviewer Point P 3.17 — The exposition would be greatly helped by including a substantive example applying this framework to paleo models, along with appropriate code, which will also motivate and demonstrate to the reader how such an analysis can be practically implemented.

Reply: This is a whole other paper in itself. LT is working on a write-up of working that addresses much of this for glacial cycle ice sheet modelling contexts.