Review of Tarasov and Goldstein

06/03/2022

Overview

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.

As a Bayesian statistician who has been attempting UQ with climate models (including paleo ice sheet models more recently) for more than 10 years, this paper was largely preaching to the choir and I rather enjoyed the early tone. I am on board with many of the ideas but I do not agree with everything, though I hope my comments on disagreements can be treated by the authors as discussion points rather than corrections. 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. Not meeting the bar on either of these elements of their roadmap 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). 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. Higher level major corrections are first, then minor line by line comments.

Major corrections

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 literatures. 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.

   Another cause for the length of the paper is the authors' tendancy to write paragraphs or even pages of potential solutions to potential problems that havent 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).

2. 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.

3. The presentation of the quantification of internal and external discrepancy in section 2. I have several discussion points here that should be addressed.

- 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. 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). 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 formalised, *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 propegates the uncertainty of boundary conditions in climate models is a whole unsolved research problem by itself. 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 parametrise the spatio-temporal boundary condition for forcing North American Ice sheet deglatiation 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.

- 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).
- 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 gridscale. 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?

- 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)

4. 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. 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.

- 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 multi-year 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?
- 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 pointwise (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.

Minor comments by line

- Line 23. 'about it evolution', here and next sentence could do with a rewrite.

- 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 revist 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 thoroghly 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.

- Lines 245-249, it should be noted somewhere that this relationship itself is a model.

- 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.

- 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.

- Line 355. Is the reason that the min residual is 0?
- 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.
- Line 500 delete "must"
- Section 2.7, I suggest is cut.
- 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.
- Line 597 "best best"
- 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.
- 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.
- 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.
- 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).
- 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).
- 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.
- 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).
- Line 739 and elsewhere, could cm be changed to c_m or something else?
- 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?

- 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.

- 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 parameterisation 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.

- 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.

- 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.

- Line 945. What is "climate forcing noise"?

- 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.

- Line 1038 might cite Liu and Guillas/Salter et al mentioned above as serious attempts at this.

- 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.

- Line 1084 Salter et al. (2019) showed that principal components can lead to spurrious model falsification when history matching and demonstrated the idea with the Canadian AGM.

- Line 1163 "emulators introduce a probability distribution for simulator output". Is this acurate in a History Matching context?

- 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).

- 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).

- Line 1183. I suggest that the word discrepancy is reserved for strucutral error and shouldnt be used in this context.

- 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).

- 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.

- 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.

- 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.

Daniel Williamson

References

- Edwards, T.L., Nowicki, S., Marzeion, B. et al. Projected land ice contributions to twenty-first-century sea level rise. Nature 593, 74–82 (2021). https://doi.org/10.1038/s41586-021-03302-y
- DeConto, R. M. & Pollard, D. Contribution of Antarctica to past and future sea-level rise. Nature 531, 591–597 (2016).
- Gilford, D. M., Ashe, E. L., DeConto, R. M., Kopp, R. M., Pollard, D., Rovere, A. Could the Last Interglacial constrain projections of future Antarctic ice mass loss and sea-level rise? J. Geophys. Res. Earth Surf. 125, e2019JF005418 (2020).
- Holden, P. B., Edwards, N. R., Rangel, T. F., Pereira, E. B., Tran, G. T., Wilkinson, R. D., (2019). PALEO-PGEM v1.0: a statistical emulator of Pliocene–Pleistocene climate. Geoscientific Model Development, 12, 5137–5155, 10.5194/gmd-12-5137-2019.
- Giang T. Tran, Kevin I. C. Oliver, [. . .] and Neil R. Edwards Building a traceable climate model hierarchy with multi-level emulators. Adv. Stat. Clim. Meteorol. Oceanogr., 2, 17–37, 2016, doi: 10.5194/ascmo-2-17-2016.
- J Van Breedam, P Huybrechts, M Crucifix. (2021) .A Gaussian process emulator for simulating ice sheet–climate interactions on a multi-million-year timescale: CLISEMv1. 0 Geoscientific Model Development 14 (10), 6373-6401.
- J Carson, M Crucifix, SP Preston, RD Wilkinson (2019). Quantifying age and model uncertainties in palaeoclimate data and dynamical climate models with a joint inferential analysis. Proceedings of the Royal Society A 475.
- JM Salter, DB Williamson, J Scinocca, V Kharin (2019) Uncertainty quantification for computer models with spatial output using calibration-optimal bases. Journal of the American Statistical Association 114 (528), 1800-1814.
- F Couvreux, F Hourdin et al (2021) Process-based climate model development harnessing machine learning: I. a calibration tool for parameterization improvement. Journal of Advances in Modeling Earth Systems 13 (3).
- Hourdin, F. et al. (2021), Process-based climate model development harnessing machine learning: II. model calibration from single column to global, J. Adv. Model. Earth Sys., 13.
- Liu, X., Guillas, S., 2017. Dimension reduction for Gaussian process emulation: an ap- plication to the influence of bathymetry on tsunami heights. SIAM/ASA Journal on Uncertainty Quantification 5, 787–812.
- Salter, J.M., Williamson, D.B., Gregoire, L.J., Edwards, T.L., 2022. Quantifying spatio- temporal boundary condition uncertainty for the North American deglaciation. Accepted, SIAM JUQ. arXiv:1808.09322

- Astfalck, L., Williamson, D. Gandy, N., Gregoire, L., Ivanovic, R. (2021) Coexchangeable process modelling for uncertainty quantification in joint climate reconstruction. Submitted to Journal of the American Statistical Association. arXiv:2111.12283.

- M Plumlee (2017) Bayesian calibration of inexact computer models, Journal of the American Statistical Association 112 (519), 1274-1285.

- Sexton, David MH, et al. "A perturbed parameter ensemble of HadGEM3-GC3. 05 coupled model projections: Part 1: Selecting the parameter combinations." Climate Dynamics 56.11 (2021): 3395-3436.

- Salter, J.M. and Williamson D. (2016) A comparison of statistical emulation methodologies for multi-wave calibration of environmental models. Environmetrics 27 (8), 507-523.