Bayesian analysis and paleo ice sheet modelling: a commentary on the proposal by Tarasov and Goldstein

Evan J. Gowan

evangowan@gmail.com

1 Overview

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. This paper is primarily a proposal, since all they are saying is that they want to see a united framework for uncertainty assessment. They do not present any new models or code, or provide a detailed literature review to back up what they want. 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. In the first part of the paper, they go over what Bayesian inference is (using a strange and poorly explained example – line 141), and how it can be applied to create an assessment of uncertainty in models (and specifically paleo ice sheet modelling). The second part of the paper goes over the way that they would like to see a Bayesian framework applied to paleoclimate and ice sheet modelling. The third part of the paper lists off what the authors want to see in any paleoclimate data and modelling study.

Since 2005, I have been involved in the world of paleoclimate science. I have been involved in the collection of data for determining sea level proxies, compiling paleoclimate and sea level data into databases, ice sheet model development, ice sheet modelling, ice sheet reconstruction using glacial isostatic adjustment techniques, and paleoclimate modelling. I would say that I have experience in the entire gamut of paleoclimate sciences and therefore the target audience of the proposal at hand. Because I work on parallel topics to Dr. Tarasov, it should not be surprising that we have had a number of encounters. 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. This paper is an extension of his opinion on this matter. However, I doubt that this dense, jargon and equation filled paper that gives few workable examples will convert anyone. 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).

I want to use this opportunity to provide an alternative view to what Tarasov and Goldstein are proposing, because I disagree with the idea that Bayesian uncertainty assessment is needed in every paleoclimate problem. In fact, as I will demonstrate, using Bayesian uncertainty assessment is probably a wasteful exercise in most paleoclimate and ice sheet modelling experiments. Before I start this, I want to make it clear that I have no problem with Tarasov and Goldstein applying Bayesian methods to their modeling experiments. The joy of science is the freedom of experimentation and exploring curiosity. It is for this very reason that I disagree with the entire premise of the proposal of this paper, as I believe it amounts to stifling creativity.

2 Bayesian Analysis is already widely used in paleoclimate sciences

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. However, this is not true – tools that make use of Bayesian analysis are widely used in the paleoclimate community. Anyone who is calibrating radiocarbon dates or making age models for sediment cores use tools like OxCal (Bronk Ramsey, 2009) or MatCal (Lougheed and Obrochta, 2016). The reason these tools are widely used are because they are easily accessible, and can be run quickly on any computer. You rarely see any papers that do not make uses of tools like these for making age models. Therefore, the general premise of that Tarasov and Goldstein make, that paleoclimate scientists are not sufficiently aware of Bayesian analysis, is not true.

If Tarasov and Goldstein want the paleoclimate community to start using their proposed Bayesian uncertainty analysis technique, they could start by making their code and programs available. The Bayesian analysis tools used for radiocarbon dating and age modelling would not have caught on if they were restricted to the laboratories of the primary authors of those tools. I can speak from my own experience. I made my ice sheet reconstruction program, ICESHEET (Gowan et al., 2016a), available on Github, and several groups have used it with minimal to no involvement on my end. One group has even created a Bayesian analysis framework for it (Pollard et al., 2023)! 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.

3 The usefulness of modelling

Lewis (2017) presented an interesting treatise on the philosophy of climate modelling (the conclusions that also apply to ice sheet modelling). She asked the question "Is climate modelling really science?" A climate model represents an approximation of the real world with various simplifications and approximations to make something that is solvable on a computer. As such, can a test truly be made to demonstrate whether a model is right or wrong? In order to falsify an ice sheet model, it would be necessary to wait centuries to see if the predictions made by them are correct. Because of this, Lewis (2017) states that for practical reasons, a climate (or ice sheet) model fails Popper's test of falsifiability, and therefore is a form of "psudo-science".

Instead of restricting the definition of science in the restrictive bounds of falsifiability, Lewis (2017) instead suggests that we should define the science of modelling in terms of its usefulness, rather than whether or not it is strictly right or wrong (something that may not be possible to assess):

> Climate models are useful because of the change they *themselves* introduce into the world. Within the climate science discipline, climate models push us introspectively to the limits of our understanding of the physical system. Extrospectively, climate models are powerful, flexible tools that are useful to society because they provide us with the capacity to pose problems and, ultimately, to act.

In the introduction, 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. 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.

I can understand the appeal of the Bayesian ensemble modelling approach, as it is comforting to think that this will make up for the unknowns that are inherent with models. I will illustrate a couple of examples of why this is not an advisable approach for modelling paleo ice sheets.

4 Past climate - the black box

At the most fundamental level, the evolution of an ice sheet is dependent on the surface mass balance. If the amount of snow and ice that melts in the summer is less than the amount that accumulates in the winter, the ice sheet grows. Therefore, the climate forcing that is included in an ice sheet modelling experiment is the most important thing to get right.

In 2019, the ice sheet modelling group I was involved with published a study comparing the Paleoclimate Intercomparison Project 3 (PMIP3) Last Glacial Maximum (LGM) experiments (Niu et al., 2019). A number of climate modelling groups were involved with PMIP, and the generated precipitation and temperature fields for the LGM and the preindustrial period. Niu et al. (2019) interpolated these fields using an climate index from a Greenland temperature proxy record, in order to create a psudo-climate time series for the last glacial cycle (*i.e.* the last 120,000 years). This climate forcing was run in an ice sheet model covering the Northern Hemisphere. The resulting LGM ice sheets were then compared. The range of ice sheet geometries for the Northern Hemisphere ice sheets was vast. Some climate models produced massive ice sheets that far exceeded the LGM limit. Other models produced ice sheets that were very small compared to the geological evidence.

The topography and ice sheet boundary conditions of the LGM PMIP3 experiments were all the same. Yet the results were very divergent. Why? This exercise showed that the simplifications, parameterizations and resolution used in climate models are important factors in the outcome of these experiments. In the areas where the Northern Hemisphere ice sheets covered, there are no proxy records that can tell us which of the PMIP3 experiments is actually the closest to the truth.

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.

4.1 The glacial inception experiment

In order to demonstrate this problem, I will highlight the recent paper from Dr. Tarasov's group, a Northern Hemisphere glacial inception experiment (Bahadory et al., 2021). The glacial inception is geologically constrained to have happened sometime between 120 and 110 ka. 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. The results of their modelling ensemble shows a broad, continuous ice sheet linking Iceland, Greenland, the Canadian Arctic Archipelago, the Cordillera and Alaska, and variable amounts of ice in Europe.

There is a big problem with this though – many of the areas that their ensemble consistently shows to be ice covered are very well constrained to have not been ice covered during the glacial inception period. First, a large part of Alaska has never been glaciated (Kaufman and Manley, 2004). Second, the Laurentide and Cordillera Ice Sheet did not merge during the glacial inception period, as it is well constrained by geological observations that this only happened once – at the LGM (Duk-Rodkin et al., 1996; Trommelen and Levson, 2008; Jackson et al., 2011). Banks Island, located in the western Canadian Arctic Archipelago, likewise was only glaciated once, during the LGM (England et al., 2009).

The severity of this mismatch means that the true ice sheet volume during this period will fall far outside of the range of the uncertainty produced by the ensemble. The amount of ice produced in these simulations, even if those areas are not included, is likely to be excessive, as far-field sea level probably remained above present until 116 ka (Clark et al., 2020). This points to problems in the climate model – it is too cold in the summer and/or there is too much snow in the winter in high latitudes. At the end of the paper, they state:

> We especially hope that the field data community will use this archive to test, refute, and/or validate which, if any, of the model-derived LGI [last glacial inception] trajectories (and characteristics thereof) are consistent with the paleo record.

However, considering the major problems with the climate forcing, I can't see how these model results will be useful for the geologists investigating the glacial inception. In their current state, ice sheet models do not have the predictive power to precisely reconstruct ice sheet history. They cannot be used as a "treasure map" (line 999). The storyline 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 overinterpreting the results in terms of how the ice sheets incepted.

5 The Latin Hypercube Shotgun

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). Latin hypercube sampling selects random values for all of the given parameters in the modelling study, such that the parameter space is well sampled given the number of ensemble members that are in the experiment. In essence, it is like shooting a shotgun at a target from a distance – some of the values will undoubtedly hit the bullseye, but others will miss the target completely.

A few years ago, I reviewed a paper by Gandy et al. (2021), which explored the deglaciation of the North Sea sector of the Eurasian Ice Sheet using a dynamic ice sheet model. The story they told in their paper was compelling, and I felt it was easy to suggest publication. In their study, they ran an ensemble of 70 simulations using seven variables. The values of these variables were selected via Latin hypercube sampling. The main criticism I had was that in varying the parameters using Latin hypercube sampling, it was not possible to say what the relative influence of each parameter on the outcome of the simulation. As a result, they stated:

> There is no clear distinction between the parameter values of the NROY [not yet ruled out] simulations and those of the rest of the ensemble, indeed the range of values for each of the seven parameters is almost the same in the full ensemble as within the subset of NROY simulations. This is partly because all parameters are varied in tandem and the parameter effects can compensate each other.

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 behavior 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.

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.

In Tarasov et al. (2012), they varied 39 different components of the ice sheet model, with parameters related to climate forcing, bed conditions and floating ice calving. After calibration, they ran 50,000 simulations, which were assessed based on how well they were able to fit observations of past sea level change, proglacial lake strandline tilts, and present day uplift rates. The modelled versions of these parameters come from a glacial isostatic adjustment (GIA) model. This model approximates the Earth as a series of spherical shells of uniform rheology. The typical Earth model used in this kind of analysis is one with an outer shell with elastic rheology (i.e. the lithosphere), and two or more shells representing the mantle, which behaves as a viscoelastic material at the time scales that ice sheets operate on. The response of the Earth to the time varying ice load is calculated by transforming the load history into the spectral domain, as the solution to the response of a viscoelastic material to loading is much easier than in the time domain. The weakness of this model is that the Earth is not a series of uniform spherical shells, and the thickness of the lithosphere and mantle viscosity are spatially variable. A 3D Earth rheology model takes at least a couple of orders of magnitude more time to solve, meaning that this is not a practical way to run a large ensemble of ice sheet model simulations. The viscosity of the lower mantle, which is an important parameter for the uplift rate in the center of the Laurentide Ice Sheet, is also not well constrained.

Tarasov et al. (2012) settled on a single Earth model in all of their simulations. This is not necessarily a bad thing. As mentioned above, it can allow for a better understanding of how the variables that are actually varied impact the ice sheet trajectories. The end product of this paper was GLAC-1D, which is an average of 10,000 of the best performing ensemble members, and an uncertainty on ice thickness. However, since the scoring of how "good" the simulation did is strongly dependent on the choice of Earth model, the average and uncertainty range is biased to their subjective choice of a single Earth model. In the paper, there are hints that the Earth model they chose might not be the best choice:

> Especially disconcerting is the weak fit to the data-rich southeast Hudson Bay sites. ... We have been unable to create a model that can dynamically (*i.e.* without ad-hoc brute force reduction of ice) produce thin enough ice over the Hudson Bay region to fit the local RSL record

So in the end, the result of the study produces a mean and uncertainty range for the deglacial of the North American ice sheets, but there is no way to know how precise this mean is to reality because it strongly dependent on their subjective choice of Earth model. Tarasov et al. (2012) acknowledge this problem, but still insist that the uncertainty ranges they create are meaningful.

7 Using the right tool for the job

One of the best GIA constraints for the center of the Laurentide Ice Sheet are the strandlines of proglacial Lake Agassiz. As the Laurentide ice sheet retreated, it opened up a vast depressed basin that meltwater could collect in, forming the lake as the ice blocked the northward drainage routes. The preserved strandlines and beaches formed by Lake Agassiz stretch for hundreds of kilometers. By measuring their present day elevation, it is possible to determine the relative amount of GIA induced uplift along the length of the strandline since its formation, and use it to constrain the ice load history.

The importance of this constraint means that it is commonly used for evaluating ice sheet reconstructions, and has been used by both myself (Gowan et al., 2016b) and Dr. Tarasov (Tarasov et al., 2012). Our approach to ice sheet reconstruction is very different, though. A dynamic ice sheet model, though in theory should be more "realistic", can never hope to precisely model the geometry of Lake Agassiz, even with 50,000 model simulations (Tarasov et al., 2012):

> Given the partially lobate structure of the geologically inferred ice margin, as well as the high sensitivity of ice margin location to what will invariably be a poorly constrained climate forcing, 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 (required for strandline predictions) and surface drainage.

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. Though this is a very rudimentary model of an ice sheet (for instance, the ice sheet was likely never in equilibrium), it has the advantage in that the ice margin can be precisely defined so that the geometry of the lake can be reproduced to match the geological observations. 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).

When designing a modelling study, it is important to use the right tool to realize the goal of your modelling exercise. 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. It is better to use a model where the margin location is strictly defined.

8 The data are never perfect

Paleoclimate and ice sheet reconstruction generally requires some kind of geological data for validation. For paleo ice sheet reconstruction, this usually comes in the form of records such as past sea level change, dated morainal features or ice flow direction indicators.

GIA based ice sheet reconstructions generally are judged based on paleo sea level data. Most available paleo sea level data from the areas covered by Late Pleistocene ice sheets are what I would call legacy data. These data were collected over 30 years ago, mostly through reconnaissance style surveys that prioritized collecting data from a wide area for constraining the ice sheet retreat history rather than detailed stratigraphic studies for constraining sea level position. The constraints on age come from liquid scintillation measured (conventional) radiocarbon dates, which generally were a composite age of dozens of mollusc shells. This means the potential for contamination is high. In most legacy studies, the elevation is stated with no explanation on how it was measured or determined, and without any reported uncertainty. For places not close to the coast, the uncertainties on elevation could exceed 10 m. In most cases, the indicative meaning concept (Shennan, 1982) cannot be applied, so the data represent minimum limiting constraints on sea level. Ross et al. (2012) and Woodroffe et al. (2014) showed that the legacy data should be treated with caution. It is unlikely that governments are going to allocate millions of dollars to resurvey using AMS radiocarbon dating and differential GPS elevation measurements, so we have to accept this.

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 uncertainties of the age of the legacy dates are likely larger (potentially a lot larger) than the reported laboratory error, and the uncertainty range for a past sea level position cannot really be confidently determined when details are not provided. So, instead, I think it is fine to assess the models in a less rigourous (and definitely less computationally intensive) way rather than attempting to create some probability distribution that is going to be hand-wavey. It also removes the need for database cleaning that might remove data points that are more accurate (line 705-715). The "consistent or not" assessment has served me well in evaluating my models. A more complex statistical model is not needed.

This problem becomes worse if you look at periods older than the LGM. One could easily make the argument that any radiocarbon date that has an age over 30,000 years should be considered suspect, and other dating methods tend to have uncertainties of thousands or even tens of thousands of years. For ice sheet reconstruction, you also have to start using undatable features like flow direction indicators to reconstruct the ice sheet, and infer the history though logic alone in a conceptual forward model. This is what I did in my most recent reconstruction (Gowan et al., 2021). I fully acknowledge that such a reconstruction might be wildly off in terms of the timing, but based on what I know about ice sheet dynamics and the patchy framework of pre-LGM geological markers of the ice sheet, I am confident that the general history of the ice sheet is correct. Since the climate forcing used in ice sheet modelling exercises tend to cause the ice sheet to grow by accumulation over broad regions rather than through ice flow from an accumulation center (a concept that fits geological observations better), these kind of observations likely cannot be used for assessment in the framework that Tarasov and Goldstein propose at present time.

9 Are error bars of paleo ice sheet simulation ensembles useful?

Tarasov and Goldstein state (lines 35-38):

> The exponential growth of accessible computer power over the last few decades has permitted the ongoing development of a synthesis of the above two approaches: inference based on rigorously combining computational modelling with paleo observations. This potentially offers detailed pictures of ice sheet and climate system evolution that can go confidently well beyond storytelling but, as detailed below, only when all uncertainties are rigorously addressed

and assessed.

This sounds great, but the question I have is, are the error bars generated from these assessments useful for paleo scientists? Do people pay attention to them?

The main product of Dr. Tarasov's modelling that people use is GLAC-1D. This is an amalgamation of his lab's modelling exercises for several of the ice sheets, where they run thousands of ice sheet model simulations by varying a bunch of parameters after reducing the range through a calibration step. GLAC-1D is an average of a subset of those simulations that performed well against the chosen evaluation metrics, along with an ice thickness uncertainty range.

The primary usage of GLAC-1D is as a boundary condition for climate modelling intercomparison experiments (Kageyama et al., 2017). 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. There is no guarantee that the ice sheets looked anything like the ensemble average. Still, it has been demonstrated that GLAC-1D is useful, because the climate modelling community does not see this as a significant problem. The key point is that the error ranges presented in GLAC-1D are not being used. No one is performing climate modelling experiments using the minimum and maximum ranges. This is largely because running climate modelling experiments are expensive (a single experiment may take several months to run), and they cannot do large ensembles. So, if the error ranges are not being used, is it really necessary to go through thousands of simulations to create one? 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. That would give the climate modelling community glaciologically plausible ice sheet reconstructions. It would make it very explicit that there is a range of possibilities for ice sheet configuration and give them an idea of the strengths and weaknesses of each reconstruction.

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.

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). I would like to highlight one recent study that shows that this is not the case. Lofverstrom et al. (2022) performed two ice sheet/climate modelling experiments to simulate the conditions at the glacial inception period. In one experiment, they closed off the straits of the Canadian Arctic Archipelago, and in the other they forced them to remain open. The experiments showed that closing the straits was a sufficient condition to initiate glaciation in far away Scandinavia. I was one of the reviewers of this paper, and my main comment was that the current (though sparse) geological evidence points towards these straits becoming blocked at a later time, and not at the glacial inception. The story of this study, however, was very compelling and interesting, and I was happy to see it published as a great example of a well designed modelling experiment. Their apples to apples comparison tells us something about the behavior of the climate system, even if it does not necessarily match a specific geological constrained reality. It was not necessary to do more than two simulations to tell that story.

Tarasov and Goldstein state (line 1089-1091):

> Many if not most modellers are more interested in improving model process representation, understanding model sensitivities, and analyzing modelling results than working through the numerous statistical issues and algorithms required for such meaningful uncertainty assessment.

There is a good reason for that, and it is because it is much easier to understand the results of a model and tell an interesting story if the model is used in a controlled way. Most modellers fully understand the limits of what their models can predict. There are limits to how much an advanced statistical framework can produce new insights when there are still so many processes that we do not fully understand.

11 Ethical modelling

11.1 Carbon Footprint

In Gowan et al. (2023), I ran three glacial cycle ice sheet model simulations of the Laurentide Ice Sheet and a number of idealized simulations to test a basal conditions module that I programmed. 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. These simulations likely represented a sizable fraction of my annual carbon footprint, and I hesitated to consider doing more simulations, especially since the simulations I performed demonstrated what I wanted to show.

As the cost of computers has decreased over time, there has been a tendency to create more complex models that require more processors and longer computational times. Complex models are desirable because they are more likely to capture small scale processes that may govern the trajectory of a simulation. However, complex models that run for a long time on hundreds or thousands of processors also use a lot of energy. Considering that one of the motivations of a climate and ice sheet modeller is to warn of the dangers of anthropogenic climate change, we should consider the impact of our simulations on the environment. We should be thoughtful about the impact of our research and not to run more simulations than what are needed to tell our story. Imagine running thousands of model experiments and finding a bug that invalidates the results! This is not as severe of a problem if the experiments are chosen wisely.

11.2 The realities of modern science funding

At line 1127, Tarasov and Goldstein propose the first step of realizing their goal is to “assemble the team”. In their ideal world, a modelling group would include people who are knowledgeable about the data that is used to assess model results, people who develop the statistical models, and someone who runs and understands the ice sheet model.

The reality is that modelling research groups are typically led by one or two permanently employed principle investigators who handle the logistics for the group, and a team of tenuously employed students and postdoctoral researchers who do model the developement, run the experiments and interpret the results. The funding cycles for projects typically last between one and five years (three years is pretty common). 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 timeframes. 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.

Scientific careers are unfortunately tied to timely publication of results. Developing a model and running hundreds or thousands of ice sheet simulations for Bayesian analysis are not possible given the time frames for most tenuously employed researchers.

12 Final remarks

A couple of years ago (Gowan et al., 2021), I published an ice sheet reconstruction study where my goal was to investigate the "missing ice problem", the problem that the global sea level drop at the LGM did not match the evidence of ice sheet configuration (*i.e.* volume). 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. I remarked:

> Our reconstruction also demonstrates that there is no consensus on Late Pleistocene ice volume, and we anticipate future refinements, for instance with different Earth rheology assumptions and ice margin histories, will produce different configurations.

In this statement, I made it very clear that I do not consider the reconstruction I published to be the one true reality, but rather that the uncertainties on ice volume are likely larger than was previously assumed (*i.e.* the uncertainty on ice volume at the LGM may be as large as 20 m of sea level equivalent). The single reconstruction, however, was sufficient to tell the story I wanted to tell. Now, if you read what Tarasov and Goldstein have written, I should not have published this because I have not done the job of quantifying the uncertainty ranges explicitly. Would I have loved to perform the ice sheet reconstruction exercise with a greater range of Earth models and ice margin histories to produce an uncertainty range? Of course! The reality is that I (along with many other modellers) do not always have the luxury of time, computer resources and funding to do this. If this makes me a "mongrel" (line 88), so be it.

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. I believe that the uncertainties in ice sheet and climate models are more likely to be narrowed through laboratory experiments, field data measurements and idealized modelling experiments where a small number of parameters are varied in a controlled way.

In the end, I am not going to suggest that the methodology that Tarasov and Goldstein want everyone to use is necessarily wrong, even if I do not intend to follow it for the reasons I have stated above. I am not a person who will stifle the creativity and imagination of other scientists. Instead, I challenge Tarasov and Goldstein to create a model development study demonstrating their Bayesian ensemble analysis concept, and publish the programs and scripts that make it possible to perform. At an even more basic level, they need to show that what they are proposing is even possible (from my reading of section 4.7, it doesn't sound like they are sure). If people agree with their methodology, they would then have the option to incorporate it into their own modelling framework.

Best Regards,
Evan J. Gowan

References

Bahadory, T., Tarasov, L., Andres, H., 2021. Last glacial inception trajectories for the Northern Hemisphere from coupled ice and climate modelling. Climate of the Past 17, 397–418. doi:10.5194/cp-17-397-2021.

Broecker, W.S., Kennett, J., Flower, B., Teller, J., Trumbore, S., Bonani, G., Wolfli, W., 1989. Routing of meltwater from the Laurentide Ice Sheet during the Younger Dryas cold episode. Nature 341, 318–321. doi:10.1038/341318a0.

Bronk Ramsey, C., 2009. Bayesian analysis of radiocarbon dates. Radiocarbon 51, 337–360. doi:10.1017/S0033822200033865.

Clark, P.U., He, F., Golledge, N.R., Mitrovica, J.X., Dutton, A., Hoffman, J.S., Dendy, S., 2020. Oceanic forcing of penultimate deglacial and last interglacial sea-level rise. Nature 577, 660–664. doi:10.1038/s41586-020-1931-7.

Duk-Rodkin, A., Barendregt, R.W., Tarnocai, C., Phillips, F.M., 1996. Late Tertiary to late Quaternary record in the Mackenzie Mountains, Northwest Territories, Canada: stratigraphy, paleosols, paleomagnetism, and chlorine-36. Canadian Journal of Earth Sciences 33, 875–895. doi:10.1139/e96-066.

England, J.H., Furze, M.F., Doupé, J.P., 2009. Revision of the NW Laurentide Ice Sheet: implications for paleoclimate, the northeast extremity of Beringia, and Arctic Ocean sedimentation. Quaternary Science Reviews 28, 1573–1596. doi:10.1016/j.quascirev.2009.04.006.

Gandy, N., Gregoire, L.J., Ely, J.C., Cornford, S.L., Clark, C.D., Hodgson, D.M., 2021. Collapse of the Last Eurasian Ice Sheet in the North Sea Modulated by Combined Processes of Ice Flow, Surface Melt, and Marine Ice Sheet Instabilities. Journal of Geophysical Research: Earth Surface 126, e2020JF005755. doi:10.1029/2020JF005755.

Gowan, E.J., Hinck, S., Niu, L., Clason, C., Lohmann, G., 2023. The impact of spatially varying ice sheet basal conditions on sliding at glacial time scales. Journal of Glaciology , 1–15doi:10.1017/jog.2022.125.

Gowan, E.J., Tregoning, P., Purcell, A., Lea, J., Fransner, O.J., Noormets, R., Dowdeswell, J.A., 2016a. ICESHEET 1.0: a program to produce paleo-ice sheet reconstructions with minimal assumptions. Geoscience Model Development 9, 1673–1682. doi:10.5194/gmd-9-1673-2016. 2016.

Gowan, E.J., Tregoning, P., Purcell, A., Montillet, J.P., McClusky, S., 2016b. A model of the western Laurentide Ice Sheet, using observations of glacial isostatic adjustment. Quaternary Science Reviews 139, 1–16. doi:10.1016/j.quascirev.2016.03.003.

Gowan, E.J., Zhang, X., Khosravi, S., Rovere, A., Stocchi, P., Hughes, A.L.C., Gyllencreutz, R., Mangerud, J., Svendsen, J., Lohmann, G., 2021. A new global ice sheet reconstruction for the past 80 000 years. Nature Communications 12, 1199. doi:10.1038/s41467-021-21469-w.

Hinck, S., Gowan, E.J., Zhang, X., Lohmann, G., 2020. PISM-LakeCC: Implementing an adaptive proglacial lake boundary into an ice sheet model. The Cryosphere Discussions 2020, 1–36. doi:10.5194/tc-2020-353.

Jackson, L.E., Andriashek, L.D., Phillips, F.M., 2011. Limits of successive middle and late Pleistocene continental ice sheets, interior plains of southern and central Alberta and adjacent areas, in: Ehlers, J., Gibbard, P.L., Hughes, P.D. (Eds.), Quaternary Glaciations - Extent and Chronology A Closer Look. Elsevier. volume 15 of *Developments in Quaternary Sciences*. chapter 45, pp. 575–589. doi:10.1016/B978-0-444-53447-7.00045-3.

Kageyama, M., Albani, S., Braconnot, P., Harrison, S.P., Hopcroft, P.O., Ivanovic, R.F., Lambert, F., Marti, O., Peltier, W.R., Peterschmitt, J.Y., Roche, D.M., Tarasov, L., Zhang, X., Brady, E.C., Haywood, A.M., LeGrande, A.N., Lunt, D.J., Mahowald, N.M., Mikolajewicz, U., Nisancioglu, K.H., Otto-Bliesner, B.L., Renssen, H., Tomas, R.A., Zhang, Q., Abe-Ouchi, A., Bartlein, P.J., Cao, J., Li, Q., Lohmann, G., Ohgaito, R., Shi, X., Volodin, E., Yoshida, K., Zhang, X., Zheng, W., 2017. The pmip4 contribution to cmip6 – part 4: Scientific objectives and experimental design of the pmip4-cmip6 last glacial maximum experiments and pmip4 sensitivity experiments. Geoscientific Model Development 10, 4035–4055. doi:10.5194/gmd-10-4035-2017.

Kaufman, D.S., Manley, W.F., 2004. Pleistocene maximum and Late Wisconsinan glacier extents across Alaska, USA. Quaternary Glaciations-Extent and Chronology: Part II: North America 2, 9–27. doi:10.1016/S1571-0866(04)80182-9.

Lewis, S.C., 2017. A Changing Climate for Science. Palgrave Macmillan Cham. doi:10.1007/978-3-319-54265-2.

Lofverstrom, M., Thompson, D.M., Otto-Bliesner, B.L., Brady, E.C., 2022. The importance of Canadian Arctic Archipelago gateways for glacial expansion in Scandinavia. Nature Geoscience 15, 482–488. doi:10.1038/s41561-022-00956-9.

Lougheed, B., Obrochta, S., 2016. MatCal: Open source Bayesian 14 C age calibration in MatLab. Journal of Open Research Software 4.

Niu, L., Lohmann, G., Hinck, S., Gowan, E.J., Krebs-Kanzow, U., 2019. The sensitivity of Northern Hemisphere ice sheets to atmospheric forcing during the last glacial cycle using PMIP3 models. Journal of Glaciology 65, 645–661. doi:10.1017/jog.2019.42.

Pollard, O.G., Barlow, N.L.M., Gregoire, L., Gomez, N., Cartelle, V., Ely, J.C., Astfalck, L.C., 2023. Quantifying the uncertainty in the eurasian ice-sheet geometry at the penultimate glacial maximum (marine isotope stage 6). The Cryosphere Discussions 2023, 1–31. doi:10.5194/tc-2023-5.

Ross, M., Utting, D.J., Lajeunesse, P., Kosar, K.G.A., 2012. Early Holocene deglaciation of northern Hudson Bay and Foxe Channel constrained by new radiocarbon ages and marine reservoir correction. Quaternary Research 78, 82–94. doi:10.1016/j.yqres.2012.03.001.

Shennan, I., 1982. Interpretation of Flandrian sea-level data from the Fenland, England. Proceedings of the Geologists' Association 93, 53–63. doi:10.1016/S0016-7878(82)80032-1.

Tarasov, L., Dyke, A.S., Neal, R.M., Peltier, W., 2012. A data-calibrated distribution of deglacial chronologies for the North American ice complex from glaciological modeling. Earth and Planetary Science Letters 315–316, 30–40. doi:10.1016/j.epsl.2011.09.010.

Trommelen, M., Levson, V., 2008. Quaternary stratigraphy of the Prophet River, northeastern British Columbia. Canadian Journal of Earth Sciences 45, 565–575.

Woodroffe, S.A., Long, A.J., Lecavalier, B.S., Milne, G.A., Bryant, C.L., 2014. Using relative sea-level data to constrain the deglacial and Holocene history of southern Greenland. Quaternary Science Reviews 92, 345–356. doi:10.1016/j.quascirev.2013.09.008. aPEX II: Arctic Palaeoclimate and its Extremes.