Review of ”Quantifying the Role of Parametric Uncertainty in Projections of Large-Scale Glacier Change”, by Megan C. James et al.

EGUsphere

Jordi Bolibar

1 General comments

James et al. present a study where they assess the role of parametric uncertainty in simulations of a large-scale glacier model. The premise of the study is that currently in most large-scale glacier modelling studies, such as GlacierMIP, parametric uncertainty is not accounted for, and they aim to show its influence and importance with a dedicated study over the RGI region of Iceland. Overall, the paper is well-written, correctly structured, and the figures are of good quality. Moreover, it addresses an important topic, which has important ramifications for our current understanding of the regional-to-global contributions of glaciers to water resources and sea-level rise.

I mainly see two important issues with this study and two minor ones that I will address in my General comments section (GC). The rest are minor comments which I will address in bullet points in the Specific comments section. With these main shortcomings addressed, I believe this is can be a good contribution to the literature of our field, deserving publication.

1.1 GC1: Simplification and alignment of history matching calibration

The core idea of this paper is to use history matching to calibrate parameter ensembles for the whole region of Iceland and to show how parametric uncertainty drives uncertainties in glacier evolution projections in this region. The authors never hide the fact that their methods are more simple than the ones from Rounce et al.(2020), and they argue that their simplifications enable a better assessment of uncertainties: *”However, there are a number of disadvantages to this approach. One is the additional complexity of considering the aggregation of uncertainties from the glacier to regional scale: treatment of parametric uncertainty as either spatially independent (varying by glacier) or spatially correlated (shared across glaciers, whether fully or partially) requires strong judgements and can substantially affect the degree of uncertainty in projections.”*. This is a very good point, also acknowledged by Rounce and others, but at the same time the simplifications of the authors' approach are a double edged sword. While the region-wide calibration simplifies things on this front, it complexifies any comparison with other GlacierMIP studies by moving away from an apples-to-apples comparison. While the authors argue that their region-wide calibration translates to a positive MB bias, it is still unclear how parametric uncertainty would look like in the real GlacierMIP setup, where the standard is to perform a glacier-wise SMB model calibration based on a single data-point per glacier from the 2000-2020 geodetic SMB from Hugonnet et al. (2021). The authors argue that applying history matching at the glacier scale would be too costly, but if that is outside of the scope of this study, I would then make it very clear that the conclusions drawn from this experiments might not be directly applicable to many of the GlacierMIP models (i.e. GloGEM, OGGM and PyGEM), due to the different nature of their calibration, and due to important differences in their SMB models (see GC2). It would be important that the authors elaborate a bit more on this, beyond the already mentioned cold bias, and investigate how much these results might look like with a glacier-wise calibration. If that remains out of the scope of the study, then I would tone down some of the claims.

1.2 GC2: Over-tuning and SMB model structure

A second important topic addressed in the Discussion is model over-tuning. The authors warn about the dangers of over-tuning in this typically data-scarce setups of SMB model calibration. While everything they say is true, and has been addressed several times previously in the literature, I am not sure that their case is much different. The main equifinality problem in large-scale glacier SMB calibration comes from the fact that mechanistic models such as temperature-index models (TIM), require observations for every glacier that needs to be calibrated. The arrival of the global Hugonnet et al. (2021) dataset changed everything, since now almost any glacier on Earth had one data point for calibration. This means that typically those SMB models calibrate from 2-3 parameters (for the simplest OGGM TIM) to 5-6 for the slightly more complex like GloGEM against a single data point. In this setup, the authors are calibrating 7 free parameters against 4 data points (4 5-year periods out of the 2000-2020 Hugonnet et al. (2021) dataset). This improves things a little bit, but not much compared to e.g. OGGM. Most importantly, in order to truly assess model performance, one needs to use an independent out-of-sample dataset. Typically, those models calibrate on geodetic SMB and validate on glaciological SMB data. This hasn't been done here, so actually all the estimates of parametric uncertainty given here are overconfident, since they are assessed against the very data used to calibrate the model parameters. The authors should re-compute the model error based on an independent data source (e.g. glaciological SMB), in order to compute the model error used in the parametric uncertainty. Another important aspect not taken into account here (related to my next paragraph), is the role of parameter transferability across time. While these simulations are steady state, most GlacierMIP models are normally run under different climate change scenarios, exposing their SMB models to drastically different distribution shifts. This has already been studied by Ismail et al. (2022), but it would be important to at least mention this and how this might affect parametric uncertainty for large-scale models under strong climatic changes.

The authors also claim that *"Physical model parameters should have the same physical meaning for every glacier, so tuning them individually underestimates the inherent model uncertainties that should be incorporated into projections"*. This is true, but only for the case where there are enough high-quality observations to correctly calibrate all those parameters.

Right now, state-of-the-art global glaciers models cannot claim to have "physical" SMB models, but rather empirical. This is not a bad thing, but actually a smart choice given the data constraints of the forcing datasets like ERA5. Deliberately choosing a simple model is a form of regularization, which effectively deals with small data. This is the main reason behind the success of TIMs. I think that the discussion brings important and interesting ideas on this front, but I would encourage the authors to nuance a couple of things. The authors acknowledge the challenge of using more physical SMB models at large scales. A recent paper that made progress on this front is Draeger and Radic (2026), which I believe should be mentioned in this discussion. The second important element, also mentioned in the discussion, is the different nature of statistical models (e.g. Sjursen et al. (2025)). One particular aspect that the authors didn't mention in the discussion, really relevant to the overparametrization and parameter transfer problems, is that statistical models are not affected by the same problems as mechanistic models (e.g. TIMs). Statistical models can be calibrated fully out-of-sample, leveraging at the same time glaciological and geodetic observations. This reduces overparametrization and addresses parameter transferability, by learning mappings between topoclimatic features and SMB changes, applicable to different spatiotemporal configurations.

1.3 GC3: Parametric uncertainty only covers SMB

The study is always presented as assessing parametric uncertainty in a large-scale glacier model, but in fact, only the parametric uncertainty of the SMB component is investigated. While the model uses a simple volume-area scaling, it completely neglects the role of glacier geometry changes (and consequently ice flow dynamics). That's just half of the story. I would suggest being clear about that, and specifically mention that parametric uncertainty of the SMB component of a large-scale glacier model is assessed, not the parametric uncertainty of the whole model.

1.4 GC4: Code availability

There is no code availability statement. In order to align with open science principles and enhance reproducibility, it would be important to share the code used to generate this study and experiments in an open repository (e.g. GitHub). This can of course be done after publication.

2 Specific comments

- **L199** From a computational point of view, how many resources (i.e. RAM, CPU cores and time) did it take to run this? How would that compare to a full-blown MCMC for Bayesian inference? This can shed some light into the added value of your simpler lightweight approach.
- **L203** "Remaining ensemble members were upscaled". What do you mean by this?
- **L206** This subsection could benefit from a little bit more structure. Clearly separating the two calibration methods (e.g. in subsections) would help the reader easily link the

calibration to the two previously presented calibration strategies.

- **L209** See GC3.
- **L260** At this stage, I found it hard to understand what was the purpose of each one of the 3.1.1 and 3.1.2 experiments. This goes with my previous comment asking for a bit more structure to clearly relate each one of the: 2 types of calibrations, 3 types of ensembles: full, tuned and history matching. It would be good to have some strict naming conventions and constantly refer to those throughout the text to make it clear how these are related. Also, for 3.1.1 and 3.1.2 it would be important to clearly explain from the very beginning what is the goal of that experiment and what you intend to show with that ensemble.
- **L296** In here, it would make it easier for the reader again to clearly structure the two types of calibration used and clearly link them the added structure I asked for in 2.4.
- **L339-343** What do you intend to show with each one of these projections? A brief explanation would make it easier for the reader to see where you're going with the results.
- **L375** I find it hard to extract clear conclusions from from a comparison which is not apples-to-apples. See GC1.
- **L483** Would this comparison be the same (or similar) if the calibration was done glacier-wise? See GC1.
- **L578** True, but as far as I know, there's no available product at these large scales.