Reviewer 2

Felikson et al. perform Bayesian calibration using different types of observational datasets as priors for an ensemble of model runs of the Greenland ice sheet under constant present-day climate. In general, better understanding the effect of calibration on the posterior probabilities is a very valuable endeavor. The study presented in the paper sheds some light on that task by showing that using different datasets for calibration of the prior distribution influences the posterior distribution quite substantially. In general, I recommend publishing the paper in TC, however, some comments have been addressed before doing so.

Entire manuscript:

-The underlying model ensemble is called "projections" in many places which will likely cause a misinterpretation of their results. I had to read to the methods to actually understand that the experiments underlying the calibration model the evolution of the Greenland ice sheet under current climate and are not informed by future SSP scenarios. A less careful reader could think that the numbers presented are sea-level projections. Make sure to change your wording, e.g. by adding "commitment under current climate" or "under constant present-day climate". Also add a timeframe over which this commitment is calculated earlier than in the methods (2100 or later?). This includes changing the title, the abstract and checking the rest of the manuscript.

Author response: We agree with the reviewer's comments, and we will make the following changes:

- Change our wording from "projections" to "commitment under current climate."
- Added the simulation timeframe to the abstract and introduction.

Abstract:

- line 6-9: What do you mean with "maximum a posteriori ice sheet mass change"?

Author response: The maximum a posteriori is the ice sheet mass change at the maximum of the posterior distribution. We have reviewed how we present our results and, in response to this comment and other comments below, we have decided to remove all references to the "maximum a posteriori" and, instead, present only the percentiles (5th, 50th, 95th), with the 50th percentile being equivalent to the median.

- end of abstract: what do you propose as a way forward?

Author response: We will add the following sentence to the end of the abstract: "Looking ahead, we present ideas for ways to improve Bayesian calibration of ice sheet projections."

The sentence in the conclusions on lines 323 – 326 states what we propose as potential ways forward: "future work should explore additional choices, such as the method for specifying model structural uncertainty, the timespan over which the calibration is done, the use of time series of observations rather than a snapshot of change, and the use of additional metrics derived from these observations."

Introduction:

- line 23: a "likelihood" of what? And "update" to what? The explanation of Bayesian calibration is quite cryptic, reformulate.

Author response: We will add detail to better describe the Bayesian calibration process in this section, including specifying what the "likelihood" and "update" are referring to.

Reviewer 2

Methods:

- Data/Model output processing: More detail is required on how you handle cells at the margins in your processing and regridding since those show most changes and hence are probably most important for you calibration. For example, is the ice front retreat that you impose part of the "dynamic thinning signal" or is this removed from the signal? And how does this compare to the observational dataset? How is this for the other datasets?

Author response: We will address this by providing additional detail in the manuscript on how mass change at the margins is handled, both in the observations and the model output. The ice front retreat is not part of the "dynamic thinning signal," which is calculated only at model mesh vertices where ice exists throughout the entire calibration period (2007-2015). Thus, the model mesh vertices that contain ice in 2007 but no ice in 2015, due to terminus retreat, are not used in the dynamic thinning calculation. Correspondingly, for observations of dynamic thinning, all observations beyond the ice extent are masked out. In other words, like the modeled dynamic thinning, the observed dynamic thinning is only calculated where there is ice during the entire calibration period (2007-2015).

For velocity change, the same processing is used; that quantity is calculated only where there is ice at the start and end of the calibration period, for both the observations and model results. For modeled mass change, the processing takes into account the ice mass lost due to retreat (and replaced with ocean water mass). For observed mass change, there is a constraint built into each mascon to treat it as either land or ocean. In our processing, only the land mascons are used in the calibration. The land mascons contain mass change signals from both land and grounded ice but in Greenland, the grounded ice mass change greatly outweighs the land mass change. We refer the reviewer to the second paragraph of Section 2.1 in Loomis et al. (2021) for more detail about the processing of the observed mass change over the ice sheet.

- line 134-136: apparently, mass changes are aggregated at a basin-scale, but figure 1 does not show this. You should update Figure 1 to show the basin-scale values you actually used, otherwise this is confusing. That you use basin-scale makes your results for mass changes comparable to other studies mentioned in the introduction that usually use aggregated values of mass change – extend on this in your discussion. If I misunderstood this comment here, and you did not do the calibration using the aggregated values, I suggest you to do this as this is a commonly used methodology and it would be interesting to compare with this.

Author response: This is correct – mass changes were aggregated at the basin scale. We will modify how we present observed mass changes in Figure 1 to show the changes at the basin scale, instead of within individual mascons.

- lines 155 and following: definitively more detail is required on the Bayesian calibration, implicit assumptions you make and how it is applied here. This methods is maybe not clear to all readers from The Cryosphere and this paper should hence make a better effort to introduce the methodology to their reader. Moreover a reference to a standard book that described the methodology should be given. Lines 156-162 need clarification, i.e., that m stands EITHER for model parameters, forgings OR mass change in 2100, and just one of them, and that m does not lump all these together. Lines 163 and following: what is the reasoning behind using these terms to calculate the likelihood terms? What are the underlying assumptions, e.g., on the distribution of the priors? A definition of sigma_{o,i} is missing (is this based on the uncertainties shown in Fig. 1?). Furthermore, you should give the exact numerical form of the equations, for example in an appendix, that you use to calculate the respective terms used, i.e., the sigma, the likelihood, the posterior distribution (do you fit this in Figs 2 and 3 to the histogram resulting from weighting the prior histogram? Which method is used for fitting the curve?).

Reviewer 2

Author response: We will review our description of our Bayesian calibration method and we will provide more detail to make it clearer to the reader. Specifically:

1. We will search for a suitable textbook reference that describes the general principles of Bayesian inference
2. We will clarify our use of the symbol "m" and will introduce other symbols, where necessary, to distinguish the various quantities that "m" represent
3. We will add text to specify that our formulation for the likelihood score, and the terms within, is a typical formulation that has been used in previous literature (e.g., Edwards et al., 2014; Ruckert et al., 2017; Nias et al., 2019; Brinkerhoff et al., 2021)
4. We will clarify in Section 2.1 that the uncertainties that are used for (1) basal friction, (2) ice viscosity, and (30 surface mass balance forcing specify the assumed prior distributions on each of these parameters and forcings
5. We will clarify that sigma_{o,i} is what is displayed in Fig. 1
6. We will add details about how the posterior distribution curve was fit using the likelihood weights

- Line 173: Does this manual adjustment of k influence anything else?

Author response: Yes, the choice of k will influence the width of the posterior, in addition to the peak of the normalized posterior probability distributions. We will add this to this part of the method section.

Results:

-Table 2: Add to the description that the "Prior" is of course not a "posterior".

Author response: We will clarify this in the table.

- Figure 4: Why are there "holes" in the residual plots for dv and dh? Description of RSS missing in legend. Do positive numbers mean that the ensemble member is faster / ticker than the observational dataset or is it the other way around? Why did you flip the colormap in the central pannels in comparison to the other panels (would be easier if they were similar)?

Author response: The "holes" are present where there are missing observations. The velocity change observations are missing data primarily in the southeast and we discuss as an issue that potentially contributes to the differences between the calibrations, in the paragraph starting on line 295. Similarly, the dynamic thickness change observations have holes where laser altimetry measurements are missing, which can happen between satellite tracks and where there is a lack of airborne altimetry. Figure 7 in Schenk and Csatho (2012) shows a representative illustration of the gaps between altimetry measurements over the Greenland Ice Sheet.

We will add the description of "RSS" to the caption of Figure 4.

We describe the interpretation of the colors in the main text, but we will also add these to the caption for clarity. We tried to set up the colormaps so that, around the margin of the ice sheet where outlet glaciers are generally accelerating, dynamically thinning, and losing mass, blues indicate that the model is overestimating acceleration (i.e., there is too much modeled acceleration), overestimating dynamic thinning (i.e., there is too much modeled thinning), and overestimating mass loss (i.e., there is too much modeled mass loss). However, thanks to this reviewer comment, we have realized that we need to also flip the colormap for the last row showing mass loss. This would make the colors consistent in the way that we want. (Note that the text in the paragraph describing this figure is consistent with the figure in the submitted manuscript, so we will change the text accordingly.)

- line 185: “maximum a posterior sea-level” can be misunderstood to mean an upper bound on the sea-level contribution, not the value with maximum probability (if this is a correct assumption from my side?).

Author response: We have decided to remove the “maximum a posteriori” estimate of Greenland mass change from our results and discussion and, instead, will present and discuss the median (50th percentile), along with the 5th and 95th percentiles.

- lines 184-197: I wonder if these large differences between the calibration methods are also partly due to the fact that overall numbers are not very large. Or putting the question the other way around, would you expect similar large percentage differences when calibrating SSP5-8.5 projections instead of “committed mass loss”?

Author response: This is a good point, and it is certainly possible that the percentage differences will be smaller when calibrating an ensemble of the forced mass loss for SSP5-8.5 rather than the committed mass loss. We will add this as a caveat to the final discussion paragraph on lines 303-310.

- line 207-8: clarify sentence.

Author response: We will clarify this sentence.

- line 220: should it be “which overestimates thickness changes around Jacobshaven”?

Author response: We will change this wording to be “overestimates thinning” to make it clearer and more consistent with the previous sentence.

- line 224 and following: Not sure what you can learn from this exercise of comparing the RSS for these three simulations.

Author response: We are using the RSS of the residuals as a measure of the sensitivity of each modeled quantity to the different calibrations. For example, because the RSS of the velocity change residuals differs by 38% across all three calibrations (top row of Figure 4), modeled velocity is more sensitive to the chosen observation that is used for calibration than the modeled thickness change and mass change (RSS residual differences of 16% and 18%, respectively). This indicates that thickness change and mass change are less sensitive to the chosen observation used for calibration.

Discussion:

- From your discussion I get the impression that you do not believe in the thickness change calibration because it is very dependent on the firn thickness change that is hard to constrain. If this is correct, I suggest that you make this one clear conclusion from your study, e.g., by suggesting to rather use the velocity or mass change calibration as these are less prone to these uncertainties.

Author response: Uncertainty in firn thickness change estimates is an important source of error that should be investigated but it is too strong of a statement to say that we do not believe in the dynamic ice thickness change calibration due to this source of uncertainty alone. The dynamic ice thickness change observations are impacted by other sources of error: sparse sampling in space and time, as well as uncertainty in the surface mass balance. And the velocity and mass change observations are similarly impacted by errors that are specific to those observations. Our present study focuses on the combined impact of all error sources on the calibration, and we plan to investigate the sources of observational uncertainty more thoroughly, and their individual impacts on the calibration, for each of the three observation types in a future paper.

In response to this comment, as well as a comment from Reviewer 1 above, we will revise the paragraph on lines 252 – 261 by removing the additional calibration in which we introduce a 10 cm/yr bias to represent a potential error in the firn densification in the interior of the ice sheet. After discussing the reviewer comments, we feel that this unfairly focused too much attention on the possible errors in firn densification, without an equivalent discussion of other sources of error. We will replace this with a discussion of all of the potential sources of uncertainty.

- line 273-278: You mention that open questions remain, but in the sentences following this, I cannot see any open questions that remain from Aschwanden and Brinkerhoff (2022). Is one open question whether it is better to use ice speed data or change in ice speed for calibration (why? Also the model initialisation in their study is different from the inversion used here, so maybe adding the direct velocity data information in their calibration is more like the step of using the velocity data for inversion done here)? Is it a bad that "the second step of calibration using mass change in Aschwanden and Brinkerhoff (2022) does not shift the posterior median estimate of ice sheet mass change" – because you make it sounds like this not a wanted result?

Author response: We agree that our statement that "open questions still remain" is not followed by a clear discussion of the questions that remain. We will revise this section to make it clear what questions are still unanswered in the existing literature, with more clarity on what Aschwanden and Brinkerhoff (2022) was able to answer. The reviewer makes a good point that the use of velocity observations in Aschwanden and Brinkerhoff (2022) is, in some ways, similar to initializing the model using velocity observations. We also do not mean to imply that a lack of a shift in the posterior median of ice sheet mass change, as seen in Aschwanden and Brinkerhoff (2022), is an unwanted result. We will clarify this point, as well.

- line 278: Selling that you account for model uncertainty in contrast to this study is a bit too much, as you only use a very ad-hoc way of including it.

Author response: Our intent was not to imply that our study takes model uncertainty into account, and we will revise the wording in this section to make this more clear.

Using the terminology in Aschwanden et al. (2021), our ensemble samples parametric and aleatoric uncertainty but not model uncertainty. The reviewer makes a good point: our statement that "open questions still remain" (see comment above), followed by this referenced discussion of model uncertainty may lead readers to incorrectly interpret that our ensemble is an advancement over the Aschwanden and Brinkerhoff (2022) work. We will clarify this section accordingly.

- line 288-294: Not sure I understand what you want to imply with this part of the discussion.

Author response: We are glad that the reviewer brought this to our attention, and we will address this comment by revising the paragraph to make it more clear. The purpose of this part of our discussion is to highlight that the calibration method that we used, also used in previous studies, estimates structural model uncertainty by using a multiplier (*k*) to inflate the observational uncertainty, typically by an order or two orders of magnitude. With this approach, the spatial structure of the observational uncertainty (i.e., the relative magnitudes between data points) becomes more important than the absolute magnitude of the uncertainty in each individual observational data point. As the modeling community develops better methods for quantifying structural model uncertainty, we can move away from using this multiplier approach and towards having an estimate of structural model uncertainty that is independent of observational uncertainty. At that point, this argument will be moot. Currently, however, calibration that is used in the literature uses this approach and, here, we are drawing attention to this to emphasize to the observational communities that the current Bayesian calibration approaches are highly dependent on the relative errors between observation data points.

We will clarify this paragraph by doing the following:

- Revising the topic sentence to: “Current approaches to estimating structural model error in Bayesian calibration results in the spatial structure of observational uncertainty to be more important than the magnitudes of the observational uncertainty in each individual data point.”
- Referencing the multiplicative factors used in previous studies.
- Ensure that we are using the wording of “observational uncertainty” and “structural model uncertainty” consistently in this paragraph and throughout the entire manuscript.

- How much does adding the structural uncertainty in the ad-hoc manner (proportional to the observational uncertainty) affect your results, i.e., how would your results look like without accounting for this term?

Author response: This is a fair question. We will perform an additional set of calibrations, setting structural model uncertainty to zero to form the likelihood. We expect that neglecting structural model uncertainty will lead to the calibration applying high weight to a relatively smaller number of ensemble members, resulting in narrower posterior probability distributions for sea-level contribution. In other words, the calibration will “hone in” on a smaller number of ensemble members and will de-weight the rest. We will add this to the supplementary figures, with text added to the results and discussion in the manuscript.

Conclusion:

- line 315: I suggest rewording here, because it is not only the mass change calibration that leads to highest scoring members with undesired behavior, you found the same also for the other calibrations. I would rather write “As we show, using the mass change calibration – or any other single dataset for calibration - does not necessarily mean …”

Author response: We agree, and we will make the suggested change.

- line 318: “right answer” to what question?

Author response: We will remove this sentence. It is paraphrasing the previous sentence but, as the reviewer points out, it is unclear.

- line 328: you claim that you have shown that “utilizing different observation types in separate calibrations can yield additional insight into biases in the model ensemble”, but the paragraph on that in the discussion (lines 261 and following) contains only once sentence on this (“For example, the highest-weighted ensemble member from the mass change ensemble overestimates acceleration (Fig. 4c) and underestimates dynamic thinning along almost the entirety of the GrIS margin (Fig. 4f).”) and this appears to be more a problem of the observational dataset (the firn correction uncertainty) rather than a bias in the mode ensemble?

Author response: We will develop additional examples for how the different observation types can yield additional insights, and we will add this to the discussion. The specific example about the underestimate of dynamic thinning along the GrIS margin does not necessarily correspond to a potential bias with the firn densification because most of the margin is an ablation zone, without firn. We will explicitly state this in this part of the discussion.

- What do you recommend, based on your results, to the modeling community to make meaningful model calibration in the future?

Author response: We have provided some guidance for future work on lines 323 – 330, but we will also add text to state that the modeling community should develop robust methods to quantify structural model uncertainty for velocity change, dynamic ice thickness change, and mass change, which could then be used to perform one multi-variate calibration using all three observation types simultaneously.

References:

Aschwanden, A., Bartholomaus, T. C., Brinkerhoff, D. J., & Truffer, M. (2021). Brief communication: A roadmap towards credible projections of ice sheet contribution to sea level. *The Cryosphere, 15*(12), 5705-5715.

Aschwanden, A., & Brinkerhoff, D. J. (2022). Calibrated Mass Loss Predictions for the Greenland Ice Sheet. *Geophysical Research Letters*, *49*(19), e2022GL099058.

Brinkerhoff D, Aschwanden A, Fahnestock M (2021). Constraining subglacial processes from surface velocity observations using surrogate-based Bayesian inference. Journal of Glaciology 67(263), 385–403. https://doi.org/10.1017/jog.2020.112

Edwards, T. L.; Fettweis, X.; Gagliardini, O.; Gillet-Chaulet, F.; Goelzer, H.; Gregory, J. M.; Hoffman, M.; Huybrechts, P.; Payne, A. J.; Perego, M.; Price, S.; Quiquet, A. and Ritz, C. (2014). Probabilistic parameterisation of the surface mass balance–elevation feedback in regional climate model simulations of the Greenland ice sheet. The Cryosphere, 8(1) pp. 181–194.

Loomis, B. D., Felikson, D., Sabaka, T. J., & Medley, B. (2021). High-Spatial-Resolution Mass Rates From GRACE and GRACE-FO: Global and Ice Sheet Analyses. *Journal of Geophysical Research: Solid Earth, 126*(12), e2021JB023024.

Nias, I. J., Cornford, S. L., Edwards, T. L., Gourmelen, N., & Payne, A. J. (2019). Assessing uncertainty in the dynamical ice response to ocean warming in the Amundsen Sea Embayment, West Antarctica. *Geophysical Research Letters*, 46(20), 11253-11260.

Ruckert, K. L., Shaffer, G., Pollard, D., Guan, Y., Wong, T. E., Forest, C. E., & Keller, K. (2017). Assessing the impact of retreat mechanisms in a simple Antarctic ice sheet model using Bayesian calibration. *PLoS One*, *12*(1), e0170052.

Schenk, T., & Csatho, B. (2012). A new methodology for detecting ice sheet surface elevation changes from laser altimetry data. IEEE *Transactions on Geoscience and Remote Sensing*, 50(9), 3302-3316.