Response to reviewer 2

Review of "Quantifying and attributing the role of anthropogenic climate change in industrial-era retreat of Pine Island Glacier" by Bradley et al

This study applies an uncertainty quantification framework called "calibrate-emulate-sample" to the multi-centennial evolution of Pine Island Glacier. The motivation for the "calibrate-emulate-sample" method is its iterative nature, which results in more members that agree with observations, compared to the standard Latin Hypercube Sampling plus Importance Sampling. The experimental design is very thoughtful, the manuscript is very well written and the authors do an excellent job highlighting potential caveats when necessary. The methods are clearly explained, and I was able to understand the gist of the "calibrate-emulate-sample" framework (I am not an expert in this field and thus not qualified to assess the details of the method).

We are grateful to the referee for their review of our manuscript and pleased to receive a positive review. Here, we respond to their comments in detail.

Main comments:
Focusing only on Figure 7a, one is tempted to conclude that all four ensembles explain the two observations of grounding line position equally well. This makes me wonder how robust your attribution to anthropogenic climate change really is.

While we agree that each of the ensembles reproduces the retreat to some extent, we disagree that they all explain the observations equally well. Indeed, the 18% that we quote through manuscript is the difference between the “all forcings” and “no anthropogenic trend” ensembles, evidencing that there is a difference between them.

Our choice of plotting style in the original manuscript may have obscured the differences between the different ensembles. To enable the differences between the ensembles to be seen more clearly, we have added figures showing mean and standard deviation of grounding line retreat between 1940 and 2015, and grounded volume in 2015 relative to observed (figures 7c and d in the updated manuscript). In these, the differences between the ensembles can be more clearly seen – anthropogenic forcing shifts the distributions to the right (corresponding to larger retreat and higher ice volume loss). We elaborate on this point in the updated manuscript, writing:

“In figure 7c, we show distributions of grounding line change between 1930 and 2015 in the four ensembles. Only in the all-forcings and no-trend ensembles is the ensemble mean plus one standard deviation within the observational range, accounting for the observational error (i.e. the right hand of the bar for the ensembles lies within the red shaded area). From this we see clearly how both the anthropogenic trend in forcing and 1940s event shift the distribution of grounding line retreat towards enhanced retreat. However, consistent with the retreat of all ensembles, as outlined above, these shifts are not extreme and the distributions still overlap.”

Are there any additional observations that could be used to further constrain the ensembles? (e.g. Sentinel images that provide the front position, observed velocities). How well does your ice sheet model reproduce reality besides retreat, e.g., velocities, dhdt? In addition, mapping

grounding line position onto a center-line is a relatively weak metric. Would you get a different result by using the floating/grounded mask and a Jaquard Score?

There are of course many different (satellite) observations that we could assimilate into the CES procedure. However, most of these are already used in the inversion procedure (including dh/dt), so are already taken into account. In addition, for the CES procedure to work well, all we need are large scale bulk metrics that are well spread over time, to constrain the centennial evolution. Satellite observations are good to constrain the present-day inversion, but weaker for the CES machinery as their changes over the modern observational era are small on the scale of the 1940s-present changes. In addition, including further present day observations into the procedure naturally downweights the 1940s observations which are crucial and only create one data point (we have already prioritised the present day somewhat by including both volume and grounding line position for 2015, versus grounding line only for 1940). We clarify this in the updated manuscript, as well as making a distinction between the spatially varying fields used in the inversion, and temporally varying fields used in the CES machinery. We clarify these points in the updated manuscript:

"*We also made several choices during the CES procedure, which should be noted. Firstly, we chose to use observations of grounding line position in 1930 and 2015, and grounded volume in 2015. There are a wide variety of satellite observational datasets available which we which could have further assimilated into the procedure. However, we elected not to as most of these are already used in the inversion, so are already taken* into account*. In addition, for the CES procedure to work well in this instance, all we need are large scale bulk metrics that are well spread over time, to constrain the centennial evolution. Satellite observations are good to constrain the present-day inversion, but weaker for the CES machinery as their changes over the modern observational era are small on the scale of the 1940s-present changes. In addition, including further present day observations into the procedure naturally down-weights the 1940s observations which are crucial and only create one data point (we have already prioritised the present day somewhat by including both volume and grounding line position for 2015, versus grounding line only for 1940).*"

- You construct three emulators, one for each target (GL 1930, 2015, Volume 2015). Would your findings change if you used one emulator that predicts all three targets?

The referee raises a good point, which we had not addressed in the original manuscript. We elected to construct different emulators for each of the targets because, when attempting to emulate all three targets simultaneously, we encountered convergence issues during emulator training (specifically, the training process requires us to invert a matrix which is poorly scaled).

It is difficult to assess the sensitivity of our results to the choice of number of emulators because it requires us to re-run the entire set of posterior simulations again. This is because a different choice of emulator(s) will give slightly different posterior distributions of model parameters and thus different samples for the posterior ensembles. Given that our emulators display good performance in regard to RMSE and coverage, we believe it is a good representation of the underlying simulation space and any other choice of emulator(s) with similar performance would yield the same posterior distributions.

The reviewer's comment is a good one and further work should investigate the sensitivity of posterior distributions to the choice of emulator.

We have added a note of these points in the updated manuscript, writing:

"Secondly, we used Gaussian processes to emulate these observations and, in particular, using individual Gaussian processes for each of the observational constraints. We chose to emulate the observational constraints individually as, when attempting to emulate them all simultaneously, we encountered convergence issues associated arising from the fact that the matrix required to be inverted during the training process was poorly scaled. Our choice of Gaussian processes enables us to obtain uncertainty estimates in emulated values of these quantities, which are propagated through to attribution assessments via posterior distributions of model parameters. Given that our emulator displays good performance, and our results include emulator uncertainties, we do not believe that the results would change if another, different emulator with similar performance was chosen, but future work should investigate the sensitivity of posterior distributions to the choice of emulator."

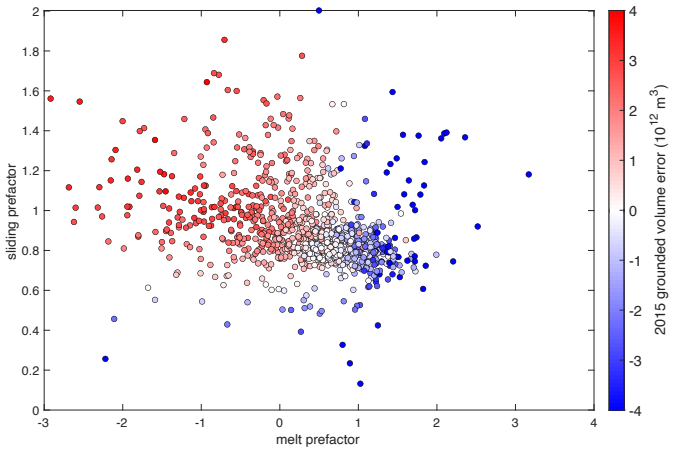
Kind regards,
Andy Aschwanden
University of Alaska Fairbanks

Minor comments:

Melt prefactor vs Sliding prefactor. The sampler seems most opinionated about these two parameters, are they strongly anti-correlated?

This is a good point. Below, we include below a scatter plot of the melt prefactor and sliding prefactor from all 1400 simulations in the EKI, with colours corresponding to the dimensionless error in the grounded volume. Visually, there is a weak negative correlation, though the melt prefactor appears a stronger control on the output than the sliding prefactor. This is confirmed by a partial correlation between the two of -0.0668. This correlation is to be expected physically: a higher (lower, respectively) sliding prefactor would reduce (promote) retreat, while a higher (lower) melt prefactor has the opposite effect, promoting (reducing) retreat. We note this in the updated manuscript, writing:

"*The sampler is most opinionated on the sliding prefactor and melt prefactor, and a partial covariance analysis reveals them to be weakly anti-correlated (R = -0.0668); physically this is to be expected: a higher (lower, respectively) sliding prefactor would reduce (promote) retreat, while a higher (lower) melt prefactor has the opposite effect, promoting (reducing) retreat*."



L 135: "which premultiplies the . $A_$..." (remove ".")

Thank-you for spotting this typo, we have fixed it in the updated manuscript.

L 219 and 220: There is no Figure 2d, I assume you mean 2c.

Fixed, thanks.

L 343 ...by Cleary et al (2021)...

Fixed, thanks.

L 393 (and elsewhere) "hasn't" -> "has not"

Fixed, thanks.

Figure 3: "...as a function of (c-e) model and (f-h) climate parameters..." I think those are switched, (c-e) are climate and (f-h) are model parameters.

Thanks for spotting this – you are correct. We have fixed this in the updated manuscript.