Point by point response to review#1

In the following we will respond to all comments of review #1, the original comments in blue, response in black and changes to the manuscript is quoted at the end of each response, where appropriate. We believe we can address all concerns in a convincing manner and think that the manuscript would greatly benefit from this revision.

Anonymous Referee #1

Received and published: 18 September 2019

This paper presents a new approach to probabilistic forecasting of future ice flow. [...] However I have serious concerns about the conclusions that the authors made from the application of their methods and cannot recommend the paper for publication. These methods have not yet been benchmarked on representative synthetic problems and this step is a necessary prerequisite for the publication of results using new methods.

We have now added a benchmark on representative synthetic problems as suggested, and adjusted our conclusions accordingly. This has allowed us to improve clarity on what we can, and what we cannot, achieve with this calibration approach, and how it compares to other approaches.

General comments:
The statistical methods that the authors use are comparatively new in glaciology. The authors cite several precedents from other fields and a paper by Chang and others from 2016 that used a similar combination of emulation and calibration. Chang et al 2016 and the current paper apply these methods to different datasets, however, and the success of the method at making certain inferences from one data set is no guarantee that the inferences from a different one are accurate.

To establish the correctness and capability of a new method on real data, it is common practice to first test it on a synthetic problem where the ground truth values of all fields and the signal-to-noise ratio of the synthetic observations are both known exactly. Without going through this preliminary testing step, you cannot be sure if the method improves on existing approaches, if the posterior density assigns non-zero probability to ground truth values, or even if the code to implement it is correct.

We agree and have now added a synthetic model test which we have applied to our proposed calibration approach as well to two other approaches for comparison. This analysis shows, very much in agreement with your remarks below, that the sliding law is not correctly inferred with any of the approaches tested here. The same is true for the ocean melt rate and we propose an explanation in the following (see below).

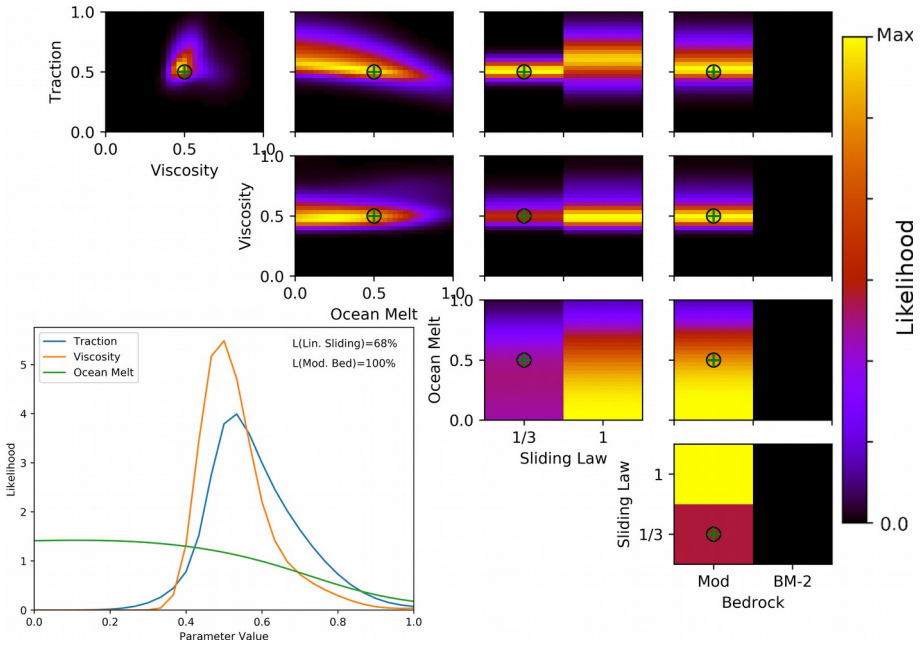
As the calibration does not adequately constrain these two parameters, we base the calibration only on the remaining parameters, namely the bedrock and basal traction and viscosity scalings. We use a uniform prior for basal melt and select nonlinear sliding by expert judgment (see below for reasoning).

Parts of the synthetic model test to be added to the manuscript:

“In this section we test our calibration approach on synthetic observations to see whether our method is capable of finding known-correct parameter values. We select one member of the BISICLES model ensemble at a time and add 14 different realizations of noise to it. The noise is added to see how the calibration performs if the observations cannot be fully represented by the ice sheet model.

We define the noise as spatially independent, zero-mean, normally distributed, random noise with variance equal to the local variance from the 14 periods of satellite observations. This way the variance incorporates dynamic changes (acceleration/deceleration of the ice thickness change) and technical errors (e.g. measurement and sampling errors). For each selected model run we generate 14 noise fields and add them to the single model ice thickness change field. These 14 realizations are used in exactly the same way as described before for the 14 periods of satellite observations.

For figure \ref{fig:mtest} the model run with central parameter values ($=0.5$) for basal traction, viscosity and ocean melt scaling factors, nonlinear sliding and modified bedrock has been selected, as indicated by black circles. This parameter set has been selected as it highlights the limitations of the calibration, but the results of many other synthetic model tests are shown in the supplement.



Caption: Likelihood of parameter combinations of synthetic test case (evaluations of Equation \ref{equ:7}). Upper right panels show likelihood values marginalized to pairs of parameters, normalized to the respective maximum for clarity. Lower left panel shows likelihood values marginalized to individual parameters for the three scalar parameters (line plots), and sliding law and bedrock topography map (text and quotation within), normalized to an integral of one, consistent with Probability Density Functions. The central values for traction, viscosity and ocean melt as well as nonlinear sliding and modified bedrock are used. The parameter values are also shown by the black circles, while the values of the set of parameters with highest likelihood are shown by green crosses. \label{fig:mtest}

As can be seen from figure \ref{fig:mtest}, marginal likelihoods of our calibration approach can favour linear sliding even if the synthetic observations use nonlinear sliding. In addition, the ocean

melt parameter is often virtually unconstrained or, as in this case, biased towards small melt factors. In contrast, the basal traction coefficient and viscosity scaling factors have a strong mode at, or close to, the correct value of 0.5 and the bedrock map is always clearly identified (Figure \ref{fig:mtest} and supplement). Different values of basal traction and viscosity have been tested in combination with both bedrock maps and show similar performance (see supplement). The fact that the parameter setup used for the test is attributed the maximal likelihood (green cross on top of black circle) supports our confidence in the implementation as the real parameter set is identified correctly as best fit. Relative ambiguity with respect to sliding law and ocean melt overrules the weak constraints on these parameters in the marginalized likelihoods.

The higher total likelihood of linear sliding can be traced back to a higher density of central ensemble members for linear sliding. Nonlinear sliding produces more extreme ice sheet simulations as fast simulations will have reduced (compared to linear sliding) basal drag and become even faster (and vice versa for slow simulations). The frequency distribution of total sea level contribution \citep{nias2016} and basis representation (Fig. S???) are therefore wider for nonlinear sliding. The relative density of ensemble members around the mode of the frequency distribution can, as for this test case, be the cause of a smaller marginal likelihood for nonlinear sliding compared to linear sliding (32\% to 68\%).

But why is the signal of sliding law and ocean melt not strong enough to adequately constrain the calibration, even though both parameters are known to have a strong impact on model simulations? This is likely related to the delayed impact of those parameters compared to the others. The perturbation of ocean melt from the start of the model period has to significantly change the ice shelf thickness before the ice dynamics upstream are affected. The fields of basal traction coefficient are adjusted to the sliding law by the inversion of surface ice velocities. It is only after the ice velocities change that the sliding law has any impact on the simulations. A change in bedrock, basal traction or viscosity have, however, a much more immediate effect on the ice dynamics and are therefore expected to dominate the calibration on short time scales.

From this test we conclude that basal sliding law and ocean melt scaling cannot be inferred from our calibration approach. We will therefore only calibrate the bedrock map as well as basal traction and viscosity scaling factors. We use a uniform prior for ocean melt scaling and select nonlinear sliding by expert judgment."

My most serious concern is with the authors' finding that a linear sliding relation gave the best fit to observational data using their calibration procedure. This result disagrees with recent published work using model-data comparison. Gillet-Chaulet et al. 2016 found that m = 1/5 or smaller gave the best fit to several years of velocity measurements for Pine Island Glacier. Joughin et al. 2019 tested the linear viscous, Weertman, and Schoof sliding laws against several years of velocity and thickness change measurements at Pine Island Glacier and found that the Schoof sliding law, which is asymptotic to m = 0 in the limit of high sliding speed, gave the best fit to observations. Other studies through the years have found evidence for nonlinear sliding using methods ranging from laboratory studies to seismic sensing. The authors state that their calibration procedure gave the best fit with m = 1 with little further discussion. Is this an assertion that glacier sliding really is linear viscous, despite numerous studies showing nonlinear and even near-plastic sliding? Or is it an artifact of the calibration? If it's the latter then the calibration procedure should be fixed, as other published methods do not come to this same conclusion.

Agreed, as explained above, the preference to linear sliding was not a robust finding. We follow your argument to justify the selection of nonlinear sliding by expert judgment.

"Since the calibration is not able to constrain the sliding law exponent, it will be represented solely by its prior. Several studies used the observed dynamical changes of parts of the ASE to test different sliding laws. Gillet-Chaulet et al. (2016) find a better fit to evolving changes of Pine Island Glacier surface velocities for smaller m, reaching a minimum of the cost function from around m=1/5 and smaller. This is supported by Joughin et al. (2019) who find m=1/8 to capture the PIG speed up from 2002 to 2017 very well, matched only by a regularized Coulomb (Schoof-) sliding law. It further is understood, that parts of the ASE bed consist of sediment-free, bare rocks for which a linear Weertman sliding law is not appropriate (Joughin et al. 2010). We therefore set the sliding law exponent prior so that only m=1/3 is used. "

Moreover, the finding that m = 1 gave the best fit to observations compared to other parameter choices that were tried does not imply that it gives a good fit to observations in any absolute sense. If the errors in the thickness change measurements are, for example, normally distributed with known variance, then the normalized sum of squared errors should come out to around 1/2. The Konrad et al 2017 paper only offers some range of possible measurement errors but this could be handled in a hierarchical Bayesian framework and the idea is the same. The question is not just what parameter combination gave the best fit to observations, but also whether that fit is good enough in an absolute sense given what we know about the error statistics. Otherwise we are merely choosing the best among bad options. This issue is discussed in MacAyeal et al. 1995 and Habermann et al. 2012.

This issue is now addressed by an initial history matching where for each parameter combination the implausibility parameter is calculated and only those parameter combinations with an implausibility<14.86 (threshold based on 99.5% of a chi-squared distribution with 4 degrees of freedom) are considered for the probabilistic calibration. This initial history matching ensures that the probabilistic calibration is only based on parameter combinations which are sufficiently close to the observations that they cannot be easily ruled out. About 20% of the input space cannot be ruled out in this way.

The following paragraph has been added:
" *\subsubsection{History matching}*
Probabilistic calibrations search for the best input parameters, but stand-alone probabilistic calibrations cannot guarantee that those are also 'good' input parameters in an absolute sense. While 'good' is subjective, it is possible to define and rule out implausible input parameters. The Implausibility parameter is commonly defined as \citep[e.g.][]{salter2018}:
\begin{equation}
\mathcal{I}(\vec{\theta}) = (\vec{\omega(\theta)}-\vec{\hat{z}})^T \mathbf{\Sigma}_T^{-1} (\vec{\omega(\theta)}-\vec{\hat{z}})
\end{equation}

A threshold on $\mathcal{I}(\vec{\theta})$ can be found using the before mentioned 'three sigma rule' (i.e. a threshold of nine is used for $\mathcal{I}(\vec{\theta})$ with one degree of freedom). Since $\vec{\omega(\theta)}$ is Gaussian, we can set an approximately equivalent threshold for the implausibility from the 99.5\% interval of a chi-squared distribution with $k=4$ degrees of freedom. Therefore we rule out all $\vec{\theta}$ with $\mathcal{I}(\vec{\theta}) >14.86$. By adding this test, called history matching, we ensure that only those input parameters are used for a probabilistic calibration which are reasonably close to the observations. In the worst case the

whole input space could be ruled out, forcing the practitioner to reconsider the calibration approach and uncertainty estimates.“

Part of the problem might stem from the choice of which parameters to calibrate. The only means by which the viscosity and basal traction can be adjusted is by scaling the amplitude of the optimal results from an inversion computed in Nias et al. 2016. The emulation method can capture the sensitivity of model outputs to variations in this amplitude scaling, but amplitude scaling as such is not necessarily a good way to capture additional modes of spatial variability. Several papers (Isaac et al. 2015, Petra et al. 2014) have successfully applied a dimensionality reduction approach in inverse problems by using the largest several eigenvalues of the Gauss-Newton approximation to the Hessian of the log-posterior. The unusual results from the calibration procedure might be ameliorated by a different choice of basis.

We agree that scaling an optimized input field, as has been done for the dataset used here, is inferior to fully exploring the ice sheet response to more flexible, higher dimensional variations to the input fields. However, computational and methodological challenges make simple scaling approaches more feasible and a common approach to represent basal traction coefficient uncertainty in forward ice sheet model simulations (see e.g. Schlegel et al. 2018, Nias et al. 2019). That is, if this uncertainty is represented at all.
The focus of this manuscript is on how spatial observations can be used for calibration of an existing set of ice sheet model simulations. Here it is not our intention to improve the initial design of ensemble experiments. Therefore higher dimensional perturbations are not possible in this case, this focus will be clarified in the revision

Schlegel, Nicole-Jeanne, et al. "Exploration of Antarctic Ice Sheet 100-year contribution to sea level rise and associated model uncertainties using the ISSM framework." Cryosphere 12.11 (2018): 3511-3534.

Nias, I. J., et al. "Assessing uncertainty in the dynamical ice response to ocean warming in the Amundsen Sea Embayment, West Antarctica." Geophysical Research Letters. (2019)

Finally, the authors state that the prediction uncertainty is greatly reduced by using their method. However, they apply a constant climate forcing, which is difficult to justify given recent trends of CO2 release that more follow the RCP8.5 scenario. The authors also state that future ocean warming is uncertain, but recent results from ocean GCMs suggest that the warming trend around the Amundsen Sea is likely to continue into the future, see Holland et al. 2019.

We agree that the simulations used here should not be understood as predictions and we have made this more clear in the manuscript now. We are not using the word ‘prediction’ for the model simulations used here anymore. We also take up the findings of Holland et al. 2019 but it has to be clear that it is one thing to suggest a long therm anthropogenic influence on the ocean melt in the ASE and a very different challenge to robustly represent climate scenarios in model simulations. To quote Holland et al. (2019):
“*Owing to the unpredictable phasing of internal climate variability, there is significant variance in wind trends between ensemble members, with the 1 s.d. range for LENS and MENS extending between no trend and twice the mean trend (Supplementary Fig. 6). Internal variability is therefore of comparable importance to radiative forcing in determining the magnitude of PITT wind changes during the twenty-first century. In the CMIP5 ensembles, inter-model differences add further uncertainty to the future trajectory of PITT winds (Supplementary Fig. 6). To deliver meaningful projections of the WAIS over this period, ice-*

sheet models will need to adopt an ensemble approach forced by multiple realizations of ocean melting."

Note that our projections are even shorter than the time scales considered in the quote, increasing the role of internal variability even more.
We therefore note that climate scenarios are expected to have small net impact on 50 year simulations and add:

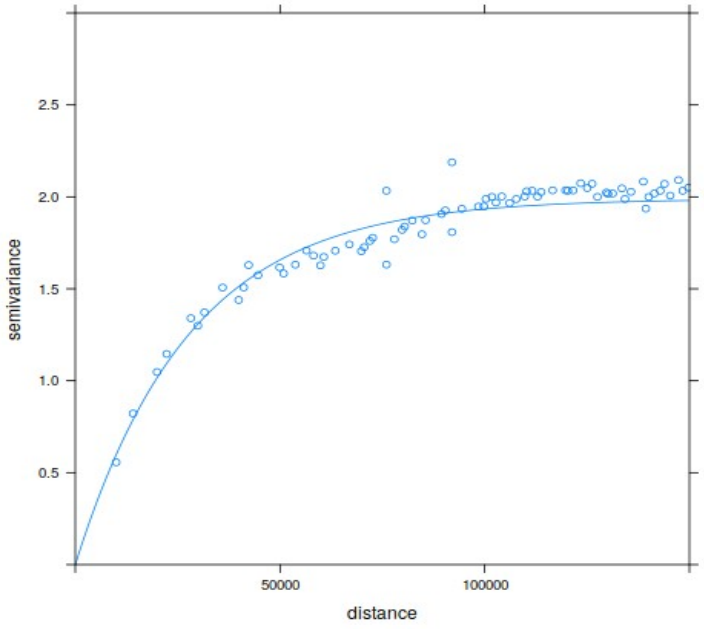
"*Relating climate scenarios to local ice shelf melt rates is associated with substantial uncertainties itself. The latest set of CMIP6 climate models are inconsistent in predicting Antarctic shelf water temperatures, so that the model choice can make a substantial (>50\%) difference in the increase of ocean melt by 2100 for the ASE \citep{naughten2018}. Melt parameterisations, linking water temperature and salinity to ice melt rates, can add variations of another 50\% in total melt rate for the same ocean conditions \citep{favier2019} and hence add another level of uncertainty. The treatment of melt on partially floating grid cells further impacts ice sheet models significantly, even for fine spatial resolutions of 300~m \citep{yu2018}. It is therefore very challenging to make robust climate scenario-dependent ice sheet model predictions. Instead we use projections of the current state of the ASE for a well defined set of assumptions for which climate forcing uncertainty is simply represented by a halving to doubling in ocean melt.*

Naughten, Kaitlin A., et al. "Future projections of Antarctic ice shelf melting based on CMIP5 scenarios." Journal of Climate 31.13 (2018): 5243-5261.
Favier, L., Jourdain, N. C., Jenkins, A., Merino, N., Durand, G., Gagliardini, O., Gillet-Chaulet, F., and Mathiot, P. (2019). Assessment of sub-shelf melting parameterisations using the ocean–ice-sheet coupled model nemo (v3. 6)–elmer/ice (v8. 3). Geoscientific Model Development, 12(6):2255–2283.
Yu, Hongju, et al. "Retreat of Thwaites Glacier, West Antarctica, over the next 100 years using various ice flow models, ice shelf melt scenarios and basal friction laws." The Cryosphere 12.12 (2018): 3861-3876."

In general the study has been re-framed towards a methods test which reduces the importance of the SLR projections.

Specific comments:
Page 2: 10-11: Worth mentioning some of the paleoglaciology literature, see Hein et al. 2016.
Done

Page 3: 9-11: How nearby and how correlated? A standard approach in geostatistics would be to assume that the correlations between the error made in measurements at point x and point y is proportional to exp(-|x - y|/L) for some correlation length L. What is the correlation length for the observational data you're using? You assert that model-to-observation comparisons on a cell-by-cell basis are not statistically independent, but that depends on whether the model resolution is large or small compared to the correlation length.



Using the seven year mean, gridded (10X10km) dh/dt data from Konrad et al. (2017) for the ASE we derived the above semivariogram which has a range value for the shown exponential fit of approximately 28000 m. Therefore the covariance of measurements 28km apart from each other reaches about 63% of the far field variance (the sill= 2 m^2 year^-2). This is in agreement with visual inspections for Figure 1 of Konrad et al. (2017) and means that L>10km.

Page 4: 15-16: Why should scaling the viscosity and friction coefficients up and down be a good way to capture variability in these fields that was not captured in the original study by Nias et al.? The true misfit might instead have a completely different spatial pattern.

The model ensemble, including the scaling, is performed by Nias et al. (2016). We tried to make this is more clear. See above discussion on the use of scaling factors

Page 10: 3: The fact that the most likely fields match the inversion from Nias only tells us that the fit can't be improved within the much lower-dimensional parameter space that you've chosen, not that it can't be improved through the addition of a completely different mode of spatial variability.

We did not intend to claim that there cannot be further improvements. The referenced note about "suggested good model consistency" was directed towards the absence of basin wide velocity biases, which could be balanced by scaled traction or viscosity fields. However, we removed this statement.

References:
Gillet-Chaulet et al. 2016, Assimilation of surface velocities acquired between 1996 and 2010 to constrain the form of the basal friction law under Pine Island Glacier, Geophysical Research Letters
Habermann et al. 2012, Reconstruction of basal properties in ice sheets using iterative

inverse methods, Journal of Glaciology
Hein et al. 2016, Evidence for the stability of the West Antarctic Ice Sheet Divide for 1.4 million years, Nature communications.
Holland et al. 2019, West Antarctic ice loss influenced by internal climate variability and anthropogenic forcing, Nature Geoscience.
Isaac et al. 2015, Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the Antarctic ice sheet, Journal of Computational Physics
Joughin et al. 2019, Regularized Coulomb Friction Laws for Ice Sheet Sliding: Application to Pine Island Glacier, Antarctica, Geophysical Research Letters
MacAyeal et al. 1995, Basal friction of Ice Stream E, West Antarctica, Journal of Glaciology.
Petra et al. 2014, A Computational Framework for Infinite-Dimensional Bayesian Inverse Problems, Part II: Stochastic Newton MCMC with Application to Ice Sheet Flow Inverse Problems, SIAM Journal of Scientific Computing

References have been added to the manuscript