1 Summary of comments and our proposed changes

In this document, we respond to every comment from the three reviewers. The reviews were extremely thorough and valuable, and addressing the comments given has substantially improved our manuscript. Below we address all 187 comments. However, first we summarize here what we inferred were the main concerns raised by the reviewers.

Please note that while this stage of the review process did not require that we prepare a new revision, we choose to do so anyway to ensure that sufficient thought was given to each comment. Consequently, our responses below use the past tense. Also note that, we changed the acronym used in this paper from MFUQ to MFSE, which refers to multi-fidelity statistical estimation. We believe this term better captures the algorithms used in this paper than term multi-fidelity uncertainty quantification (MFUQ), which is broader and could be perceived to include methods such as multi-fidelity surrogate modeling. Ideally, we would use the term multi-fidelity Monte Carlo (MFMC), instead of either of the two aforementioned terms. However, MFMC is used in the literature to to a particular type of MFSE algorithm.

First, all three reviewers stated that the Section 4, which introduced multi-fidelity methods, needed to be rewritten more clearly. We were happy to do this, as we want to make the paper as easily accessible as possible. We also thank the reviewers for their constructive advice on how to improve our exposition.

Second, the reviewers asked to more clearly and precisely establish the novel contributions of our work, especially compared to a recent paper that we were unaware of. Additionally, we were asked to add further discussion of the limitations of our study to the paper. We have rewritten the introduction and discussion sections to address this concern.

Third, but not least, the third reviewer (listed in this document) asked how MFSE compares to a recent method introduced in a paper referred to as BR23. The method in BR23 linearized the parameter-to-QoI map of an ice-sheet model to computationally efficiently estimate the distribution of the QoI conditioned on observations. Moreover, the reviewer also remarked that our procedure for selecting the prior distribution of our uncertain parameters could be improved by pointing to the tuning method proposed in BR23. Specifically, the reviewer correctly pointed out that not using a rigorous method for tuning the hyper-parameters of the prior distribution, such as that proposed in BR23, can lead to uncertainty being underestimated. However, the method requires automatic differentiation to linearize the parameter-to-QoI map, which our codes (and many others) do not support, so we were unable to make a comparison or adopt the tuning procedure proposed in the cited paper. Secondly, linearizing the parameter-to-QoI map introduces an error when the map is nonlinear. As the cited paper points out this, error depends on the strength of the nonlinearity in the map. As we could not construct the linearization used in BR23, we could not compute the size of this error. However, we have included plots in the paper and response that demonstrate the parameter-to-QoI map is nonlinear. Moreover, we now remark in the paper that if linearization is possible, then it will likely be a computationally efficient and accurate low-fidelity model that could potentially improve the performance of MSE further.

In the following we list the reviewers comments with *"black italic"* font. We list text from the original document with *"blue italic"*. Additionally, we include new text added to the revised document in red. Lastly, we made additional edits to the paper to improve readability, but we only highlighted such changes if they were in response to a reviewer comment or change the narrative substantially.

1.1 Miscellaneous

Please note that there was an issue with section numbering in Section 4 which we have now corrected. In the revised document we created a new subsection 4.4 called *"Computational considerations for multi-fidelity uncertainty quantification"* which discusses the important computational aspects of the exploration and exploitation phases of multi-fidelity UQ. Section 4.4 now includes subsections 4.2.3, 4.2.4, 4.25, 4.2.6 in the original submission. Moreover, Remark 4.1 in the original submission has now been moved to form the basis of the introduction in Section 4.4.

2 Reviewer 1 (Douglas Brinkerhoff)

2.1 General comments

"In this manuscript, Jakeman et al. present the application of multi-fidelity uncertainty quantification to accelerate – and hopefully provide more accurate – estimates of first- and second-order ensemble statistics. They begin by describing two approximations to the Stokes' equations, which trade solution expressivity for computational expense and serve as the basis for their multifidelity methods. They next describe a mechanism by which to characterize an approximate posterior distribution (based on a low-rank Laplace approximation) over basal traction conditioned on surface velocity observation, which serves as the source of samples for Monte Carlo sampling of ice volume evolution, the primary quantity of interest in this work. The primary methodological advance is the introduction of an adaptive control variate (ACV) estimator for the mean and variance of mass change after approximately a century of ice evolution. This estimator leverages correlations between so-called low- and high-fidelity models (which have different computational expense) to effectively reduce the error in Monte Carlo estimates of ice volume change relative to predictions made using a limited number of high-fidelity model predictions on its own. They present this method for a single high and low-fidelity model, and then extend the analysis to the case where there exists one high-fidelity model and a hierarchy of multiple low-fidelity models. Such methods require the establishment of statistical relationships between the low and high-fidelity models, and also careful selection of the number of samples evaluated for each constituent model: the authors carefully present strategies for these tasks within the framework of a fixed computational budget, and explore the implications of these strategies when they are better informed by empirical analysis than theory. The manuscript applies these methods to the Humboldt Glacier basin of Northwest Greenland and show that the present methods can be used to perform effective uncertainty quantification, at least over the limited subset of uncertain parameters that the authors' consider. This work is an important and timely contribution to the growing effort towards robust uncertainty quantification in ice sheet modeling. I have no major objections to the scientific content of this work, which I find to be well-motivated and defensible. I do think that the work relies on language and a presentation style that will be challenging for many readers, particularly those without a specialized statistical background. As a general comment, I would encourage the authors to try to provide more intuition and plain-language summaries, particularly in Section 4. I provide more specific examples, alongside other detailed comments below."

2.2 Specific comments

1. *"L26 Here and elsewhere, 'effects' should be 'affects"'.*

   Fixed.

2. *"L97 Should $H(x, y, z)$ be $H(x, y, t)$?"*

Yes. We fixed this typo.

3. *"L98 The MOLHO (or the Blatter-Pattyn approximation) doesn't neglect vertical velocity, it is just eliminated from the system of equations via mass conservation and the assumption of hydrostatic pressure. It can always be determined a posteriori from the horizontal velocity components."*

   We changed the sentence *"In contrast, using the observation that ice-sheets are typically shallow, i.e. their horizontal extent is much greater than their thickness, the MOLHO model neglects the vertical velocity w and only simulates the horizontal velocities $u(x, y, z, t), v(x, y, z, t)$ but still as functions of the three spatial coordinates"* to In contrast, the MOLHO model makes simplifications based on the observation that ice-sheets are typically shallow, i.e. their horizontal extent is much greater than their thickness. These simplifications lead to a model that does not explicitly estimate the vertical velocity w and only simulates the horizontal velocities $u(x, y, z, t)$, $v(x, y, z, t)$ as functions of the three spatial coordinates.

4. *"I don't think Dukowicz is the best reference here. Pattyn (2003) is more commonly cited, or if the preference is for something that clearly describes the hierarchy of approximations, Schoof and Hindmarsh (2009)."*

   We have added a citation to Schoof and Hindmarsh (2009).

5. *"L137 I think that citet should be used rather than citep here."*

   We now use citet.

6. *"L152 The discretization of the continuity equation is non-trivial and should be described here. How was it stabilized? How was positivity (even in the absence of negative forcing) ensured?"*

   We added the following details:

   Specifically, the continuity equation was discretized with nodal finite elements, using streamline upwind stabilization. Additionally, the advection term was integrated by part and the thickness was treated implicitly. Using this time evolution process, we did not observe any numerical instabilities when using the time-step sizes adopted in this study.

7. *"L172 Who specifically considers friction to be such a large source of uncertainty? Plenty of recent work has shown that forcing terms are the most important uncertainty sources, particularly at long time scales. I don't have a problem with focusing on traction here, but I think it is important to contextualize this choice a little bit more fully."*

   We agree that forcing terms are large sources of uncertainty. Carr et al. (2024) in particular have recently made this argument. However, basal sliding is widely acknowledged to also be a large source of uncertainty (see e.g., Nias et al. (2018); Joughin et al. (2019); Brondex et al. (2019); Åkesson et al. (2021); Hillebrand et al. (2022)), especially when using model configurations with active calving, which Carr et al. (2024) ignored. While we also do not account for active calving except to prevent advance, our methodology could be extended to situations that include physically based calving laws.

   We changed the statement *"While all sources of uncertainty may significantly impact predictions of mass change from ice sheets, this study focused on quantifying uncertainty due to the unknown basal friction, which is considered one of the largest sources of prediction uncertainty. This singular focus was made to improve our ability to assess whether MSE is useful for ice-sheet modeling for a very high-dimensional source of uncertainty, which cannot be*

tractably tackled using most existing UQ methods. This ensures that the conclusions drawn by our study can be plausibly extended to studies considering additional sources of uncertainty." to While all these sources of parametric uncertainty may significantly impact predictions of mass change from ice sheets, this study focused on quantifying uncertainty due to unknown basal friction, which is considered one of the largest sources of prediction uncertainty after future environmental forcing (Nias et al., 2018; Joughin et al., 2019; Brondex et al., 2019; Åkesson et al., 2021; Hillebrand et al., 2022; Nias et al.,190 2023). This singular focus was made to improve our ability to assess whether MFSE is useful for quantifying uncertain in ice-sheet modeling with high-dimensional parameter uncertainty, which most existing UQ methods cannot tractably address. By doing so, we ensured that the conclusions drawn by our study can be plausibly extended to studies considering additional sources of uncertainty.

8. *"L180 I think that the community is using the term 'Gaussian process' with some frequency now, so it would be good to at least mention that as a name for what is going on here (and a reference to, say Rasmussen and Williams (2006))."* We did not add this or a similar citation. We want to avoid using the term Gaussian process because the method we used to generate the Gaussian random field, which the reviewer is referring to, and condition it on data differs from the approach used by Gaussian process used in machine learning, e.g. (Rasmussen and Williams, 2006).

9. *"L212 For what it's worth, it's a stretch to call BedMachine 'data' – it is the result of a PDE-constrained optimization scheme that relies on assumptions of climate, smoothness, and a variety of other things. Again, nothing different needs to be done, but it is important to state that this inferred geometry is assumed error-free. "*

   We clarified that we assumed the geometry was error-free.

10. *"L232 This scaling term is less mysterious when reported with units 'number of observations per area'."*

    We think that the "scaling term" the reviewer is referring to is the inverse of our coefficient α in the original submission. In our case, the observations are given as a spatial field, not as a finite number of data points, therefore α has the units of area, which is consistent with the reported units of km^2. In an attempt to make the scaling coefficient more intuitive, we changed the definition of α and considered its inverse, which is more inline with the typical definition of scaling terms in objective functionals. We also expanded Remark 5.1 to discuss how the scaling parameter α has been selected.

11. *"L263–267 Here and elsewhere, please be sure to use a consistent tense. This switches from present to past inside a sentence."*

    We have corrected tense here and throughout the document.

12. *"L298 Should these sums have $N-1$ in front of them?"* Yes. We corrected this mistake.

13. *"L298 For the inner sum, please use a different index variable than n."*

    We now use the symbol j.

14. *"L298 Is the optimization of η mandatory or does this work with arbitrary η? What is the objective that is optimized?"*

    The optimization of η is not mandatory. Any value can be used, indeed MLMC uses $\eta = -1$, however using a non-optimized value of η can substantially degrade the accuracy of the

estimator. We now mention that η can be set a priori. To see the exact changes refer to the added text in red in response to comment 15.

Also note that, in the original manuscript we stated that *"the MSE of the ACV estimator can be minimized by optimizing the determinant of the estimator covariance matrix"*. For a fixed sample size, the optimization of η is analytical as reported in Equation (14) of the original manuscript.

15. *"L298 The reader would benefit from a description of what this equation means and some intuition of why this works. It appears to be that the low-fidelity terms yield a correction to the high fidelity statistic, but it is somewhat surprising that this doesn't need to include any explicitly quantified relationship between the two models. It would also be helpful to emphasize that the Θ_0 and Θ_1 can have different set sizes."*

We added the following text to the revised document to provide the intuition requested and discuss the sizes of the sample sets Θ_0 and Θ_1.

Using only high-fidelity model simulations to estimate a statistic with single-fidelity MC produces an unbiased estimator of Q_0. However, when the computational cost of running a high-fidelity model limits the number of model simulations that can be used, the variance and thus the MSE, of the MC estimator will be large. Fortunately, the MSE error of the estimator can be reduced by correcting the high-fidelity estimator with statistics computed using lower-fidelity models. For example, given a high-fidelity model $f_0(\theta)$ and a single low-fidelity model $f_1(\theta)$, an MFMC ACV estimator approximates the mean of the high-fidelity model using

$$Q^{\mu}_{\mathrm{ACV}}(\Theta_0,\Theta_1) = N_0^{-1}\sum_{n=1}^{N_0} f_0(\theta_0^{(n)}) + \eta\left(N_0^{-1}\sum_{n=1}^{N_0} f_1(\theta_0^{(n)}) - N_1^{-1}\sum_{j=1}^{N_1} f_1(\theta_1^{(j)})\right) \tag{1}$$

$$= Q^{\mu}_0(\Theta_0) + \eta(Q^{\mu}_1(\Theta_0) - Q^{\mu}_1(\Theta_1)) \approx \mathbb{E}_{\Theta}\left[f_0\right]. \tag{2}$$

The two-model ACV estimator in Eq.(21) uses a weighted combination of a high-fidelity MC estimator and two low-fidelity estimators. The high-fidelity model evaluations are used to ensure the ACV estimator is unbiased, i.e. $\mathbb{E}_{\Theta}\left[Q^{\mu}_{\mathrm{ACV}}(\Theta_0,\Theta_1)\right] = \mathbb{E}_{\pi}\left[f_0\right]$, while the low-fidelity evaluations are used to reduce the variance of the estimator. The estimator of the low-fidelity mean $Q^{\mu}_1(\Theta_0)$ is referred to as a control variate because it is a random variable, which is correlated with the random estimator $Q^{\mu}_0(\Theta_0)$, and can be used to control the variance of that high-fidelity estimator. The term $Q^{\mu}_1(\Theta_1) \approx Q^{\mu}_1$ is an approximation of the true low-fidelity statistic Q_1 that is used to ensure that the ACV estimator is unbiased, i.e. $\mathbb{E}_{\Theta}\left[Q^{\mu}_{\mathrm{ACV}}(\Theta_0,\Theta_1)\right] = \mathbb{E}_{\Theta}\left[Q^{\mu}_0(\Theta)\right] + \eta\left(\mathbb{E}_{\Theta}\left[Q^{\mu}_1(\Theta_0)\right] - \mathbb{E}_{\Theta}\left[Q^{\mu}_1(\Theta_1)\right]\right) = Q^{\mu}_0 + \eta\left(Q^{\mu}_1 - Q^{\mu}_1\right) = Q^{\mu}_0$. The weight η can either be fixed – e.g. MLMC sets $\eta = 1$ – or optimized to minimize the MSE of the estimator. However, an ACV estimator will always be unbiased, with respect to Q_0, regardless of the value of η, because the expected values of the second and third terms will always cancel.

Computing the ACV estimate of the high-fidelity mean in Eq. (21) requires two different sets of model evaluations. These evaluations must be obtained by first drawing two sets of samples $\Theta_0 = \{\theta_0^{(n)}\}_{n=1}^{N_0}, \Theta_1 = \{\theta_1^{(n)}\}_{n=1}^{N_1}$ from the distribution of the random variables. In our study, we draw random samples from the posterior distribution of the log basal friction, i.e $p(\boldsymbol{\theta} \mid \mathcal{M}, \boldsymbol{y})$. The high-fidelity model must be evaluated on all the samples in Θ_0 and the low-fidelity model must be evaluated on both the sets Θ_0 and Θ_1. Typically $N_0 < N_1$. In

most practical applications, such as this study, the model f_0 used with an ACV estimate is chosen to be the highest-fidelity model that can be simulate $O(10)$ times, However, when a model utilizes a numerical discretization that can be refined indefinitely, MLMC can be used to adaptively set Q_0 such that the discretization error $Q_0 - Q_\infty$, in Eq. (18), is equal to the variance $\mathbb{V}_\Theta[Q_{\text{ACV}}]$ of the MLMC estimator.

16. *"L323 The line about some samples being shared is vague. Please elaborate on what this means."*

    We replaced this line with:

    Different ACV estimators can be produced by changing the way each sample set is structured. For example, MFMC estimators sample the uncertain parameters such that $\Theta_\alpha^* \subset \Theta_\alpha$ and $\Theta_\alpha^* = \Theta_{\alpha-1}$ and MLMC estimators sample such that $\Theta_\alpha^* \cap \Theta_\alpha = \emptyset$, and $\Theta_\alpha^* = \Theta_{\alpha-1}$.

    We also know state in Section 4.4.2:

    Each existing ACV estimator was developed to exploit alternative sample structures $\mathcal{T}$ to improve the performance of ACV estimators in different settings. For example, a three model ACVMF estimator performs well when the low-fidelity models are conditionally independent of the high-fidelity model. Imposing this conditional independence is useful when knowing one-low-fidelity does not provide any additional information about the second low-fidelity model, given enough samples of the high-fidelity model. This situation can arise when the low-fidelity models use different physics simplifications of the high-fidelity model. In contrast, MLMC assumes that each model in the hierarchy is conditionally independent of all other models given the next highest fidelity model. This allows MLMC to perform well with with a set of models ordered in a hierarchy by bias relative to the exact solution of the governing equations.

17. *"L328 'statistics' → 'statistic'."*

    Fixed.

18. *"Eq. 14 As before, are there alternatives to using this value for η?"*

    See responses 14 and 15.

19. *"Eq. 16 Split this into two equations, and add matrix sizes for each."*

    Fixed.

20. *"L352 I'm not sure I understand this sentence."*

    We changed this sentence when responding to the next comment.

21. *"L356 I think it would be better to include more detail about how these expressions are used to compute Eq. 15 than just referencing Dixon (2023). Otherwise, it sort of feels like a lot of space gets used describing Eqs. 16 and 17, but they never really go anywhere."*

    We removed equation 16 and 17, as we agree that they did not really build intuition. Moreover, including the expressions in Dixon et al (2023) are extremely complicated and would likely turn off all but the most mathematically inclined reader.

22. *"Eq. 20 Why is it the case that minimizing the determinant of the covariance determines an optimial sampling strategy?"*

Because we are computing a vector valued statistic, comprised of the mean and variance of the mass loss at the final time, the MSE error of the statistic is a matrix. The determinant was first proposed in as a scalar metric that quantifies the error of a vector-valued estimator. It is likely possible to use alternative measures, such as the trace to quantify error, however to date only the determinant has been used. Indeed, most MSE literature only focuses on estimating a single statistic of a scalar function, in which case the trace and the determinant are equal.

We now state in Section 4.4.2: Unfortunately, a tractable algorithm for solving Eq. (29) has not yet been developed, largely due to the extremely high number of possible sample allocations in the set $\mathbb{A}$. Consequently, various ACV estimators have been derived in the literature that simplify the optimization problem, by specifying what we call the sample structure $\mathcal{T}$, which restricts how samples are shared between the sets $\Theta_\alpha, \Theta_\alpha^*$. For example, optimizing the estimator variance, Eq. (23), of a two model MFMC (Peherstorfer et al., 2016) mean estimator, Eq. (21), requires solving

$$\begin{aligned}
&\min_{N_0,N_1} N_0^{-1}\mathbb{V}\left[f_0\right]\left(1-\frac{N_1-N_0}{N_1}\mathbb{C}\mathrm{orr}[f_0,f_1]^2\right)\\
&\text{s.t.} \qquad N_0w_0+N_1w_1\le W_{\max},\\
&\mathcal{T}=\{N_{0\cap 1*}=N_0, N_{0\cup 1*}=N_0, N_{0\cap 1}=N_0, N_{0\cup 1}=N_1, N_{1*\cap 1}=N_0, N_{1*\cup 1}=N_1\}.
\end{aligned}$$

Alternatively, minimizing the estimator variance of the two model MLMC (Giles, 2015) mean estimator requires solving

$$\begin{aligned}
&\min_{N_0,N_1} N_0^{-1}\mathbb{V}\left[f_1-f_0\right]+(N_1-N_0)^{-1}\mathbb{V}\left[f_1\right]\\
&\text{s.t.} \qquad N_0w_0+N_1w_1\le W_{\max},\\
&\mathcal{T}=\{N_{0\cap 1*}=N_0, N_{0\cup 1*}=N_0, N_{0\cap 1}=0, N_{0\cup 1}=N_1, N_{1*\cap 1}=0, N_{1*\cup 1}=N_1\}.
\end{aligned}$$

MLMC and MFMC employ sample structures $\mathcal{T}$ that simplify the general expression for the estimator covariance
$\mathbb{C}\mathrm{ov}_\Theta\left[\boldsymbol{Q}_{\mathrm{ACV}},\boldsymbol{Q}_{\mathrm{ACV}}\right]$ in Eq. (27). These simplifications were used to derive analytically solutions of the sample allocation optimization problem in Eq. (29) when estimating the mean, $\mathbb{E}_\Theta\left[f_0\right]$ in Eq. (16), for a scalar-valued model. However, the optimal sample allocation of MLMC and MFMC must be computed numerically when estimating other statistics, such as variance $\mathbb{V}_\Theta\left[f_0\right]$ in Eq. (17). Similarly, numerical optimization must be used to optimize the estimator covariance, $\mathbb{C}\mathrm{ov}_\Theta\left[\boldsymbol{Q}_{\mathrm{ACV}},\boldsymbol{Q}_{\mathrm{ACV}}\right]$ in Eq. (27), of most other ACV estimators, including the ACVMF and ACVIS (Gorodetsky et al., 2020), as well as their tunable generalizations (Bomarito et al., 2022).

23. *"Fig. 4 and accompanying text I don't think that the text does a sufficient job of describing the principles behind these different sampling strategies. Figure 4 tells me that Θ_0 and Θ_1^* share their samples, and different schemes use entirely different or appended different samples for Θ_1, but I cannot grasp from the text why this is significant. This needs to be motivated fully or de-emphasized and more carefully referenced."*

We removed the figure and instead now try to provide intuitive descriptions on why different sample structures target different relationships between models.

See our response to comment 16.

24. *"L394 'model' appears twice."*

    Fixed.

25. *"L433–439 Is there an argument that can be made here to reassure a reader that the observed changes are due to real climate/ice dynamic effects and not so-called 'transients' resulting from inconsistency between initial conditions, physics, and input fields? Fig. 7 (left) has some rather surprising high-frequency noise in the surface elevation change – it would be helpful to know where this comes from."*

    The high-frequency content is largely due to the oscillations in the posterior sample of the basal friction (Figure 6). We have added a similar statement to the document when discussing Figure 7.

26. *"L453 Just to clarify, was Θ_{pilot} shared across models, or were the samples different for each model?"*

    We now state:

    First, we evaluated each of our 13 models at the same 20 random pilot samples of the model inputs Θ_{pilot}

27. *"L458 'significant' is an unfortunately subjective term here - to my eye, the differences in the heights of the referenced bars seems rather insignificant. Is it possible to elaborate on the meaning of 'significant' here, and why it should be viewed as such?"*

    We updated the figure to more clearly show the differences between the means and standard deviations of each model. We also now state:

    The middle and lower panels of Figure 8 show that the means and standard deviations of each model differ.

28. *"L467 This sentence is a bit challenging, with 5(!) nested prepositions and two uses of 'variance' each describing different things. I recognize that it is challenging to compactly describe the statistics of statistics, but is it possible to relax this sentence a bit?"*

    The previous text was *"Given estimates of the pilot statistics, (18) and (19), we used (20) to predict the determinant of the variance of the ACV estimator of the mean and variance of the mass change. "*

    we change this to We used Eq. (20), with estimates of the pilot statistics obtained using Eq. (18) and Eq. (19), to predict the determinant of the the ACV estimator covariance.

29. *"L476 This assertion is surprising to me, and a citation describing the assumption of pilot statistic exactness would be useful."*

    We have added a citation to (Peherstorfer et al., 2016) which was cited elsewhere in the paper for another reason. In the last paragraph from Page A3181 the authors state "We use the sample variances and the sample correlation coefficients to determine the number of model evaluations m and the coefficients α. Table 2 compares sample variances and sample correlation coefficients computed from 10, 100, and 1000 samples. The different number of samples leads to different estimates."

    The authors also state that "the variations in the sample variances have only a minor effect on the coefficients α" where α refers to the control variate weights denoted η in our paper. However our results show that the impact of the size of the pilot is problem dependent.

30. *"L494 Is there an interpretation of why some models appear to be more informative than others, or is this just random chance? I can't identify a mechanism for why some low-fidelity models were chosen more frequently, but understanding that (or being able to predict it a priori) would be exceptionally useful."*

    We included the following statement at the end of section 4:

    Whether a model is useful for reducing the MSE error of a multi-fidelity estimator depends on the correlations between that model, the high-fidelity, and the other low-fidelity models. For toy parameterized PDE problems, such as the diffusion equation with an uncertain diffusion coefficient, theoretical convergence rates and theoretical estimates of computational costs can be used to rank models. However, for the models we used in this study, and likely many other ice-sheet studies, ordering models hierarchically, that is, by bias or correlation relative to the highest-fidelity model, before evaluating them is challenging. Indeed, the best model ensemble for multi-fidelity UQ may not be hierarchical (see Gorodetsky, 2020). Yet, estimators such as MLMC and MFMC only work well on model hierarchies. Consequently, having a practical approach for learning the best model ensemble is needed. Yet, to date this issue has received little attention in the multi-fidelity literature. Section 5 provides a sorely needed discussion of the impact of the pilot study on model selection and the error a multi-fidelity estimator.

31. *"L496 I don't understand what is meant by a 'hierarchical relationship' here."*

    See response 30.

32. *"L525 I get where these numbers come from after some digging back through the other sections, but it would be helpful to remind the reader where each of the terms in the cost expression represent."*

    We added the following to the text:

    The cost of constructing our final estimator was equal to the sum of the pilot cost (197.13 hours) and the exploitation cost (160×4.18) hours, which was approximately 36 days. The pilot cost was the sum of evaluating all 13 models on the initial 20 pilot samples and 8 models on an additional 10 pilot samples (see Section 5.2). The exploitation cost was fixed at the beginning of the study to the computational cost equivalent to evaluating the high-fidelity model 160 times, which takes a median time of 4.18 hour to simulate.

33. *"L536 Is this extreme asymmetry between the number of high- and low-fidelity model evaluations typical? I think that this is a significant and interesting result if the high-fidelity model is only really needed to, e.g. characterize the spatial variability of the mean solution, but the low-fidelity models are sufficient to characterize all of the uncertainty."*

    We added the following statement to Section 5.5:

    The allocation of the small number of samples to the highest-fidelity model is due to the extremely high-correlation between that model and the model $\text{MOLHO}_{1km,36days}$. This high-correlation suggests that the temporal discretization error of the highest-fidelity model is smaller than the spatial discretization error.

34. *"Table 1 I don't think this needs to be a table."*

    We removed the table and now state the following in the main text:

    The mean and standard deviation computed using the best ACV estimator were -639.06 ± 0.23 and 17.68 ± 6.67, respectively.

35. *"L589 The variance doesn't have the same units as the mean, so I'm not sure what numbers I'm looking at. Is 17.68 the standard deviation?"*

    You are right. We now state the values in question represent standard deviation.

3 Reviewer 2 (Vincent Verjans)

3.1 General comments

"This study proposes a multi-model method for evaluating uncertainties in ice sheet model projections. This method uses models of different degrees of fidelity to simulate glacier mass change projections, therefore referred to as multi-fidelity uncertainty quantification (MFUQ). It exploits correlation between the different model realizations to approximate the statistics that would be obtained by the highest-fidelity model available, but at reduced computational cost. Here, the study focuses on uncertainty arising from the uncertain basal friction input field, and shows an application at Humboldt glacier, Greenland. Random samples of the basal friction field are drawn from a Laplace approximation of the posterior probability distribution, which is calibrated to match output from an ice flow model to the present-day Humboldt glacier configuration. The study then compares the MFUQ method with Monte Carlo sampling using the highest-fidelity model only, which is referred to as single-fidelity Monte Carlo (SFMC). Results show that, applied to this problem, MFUQ can serve to infer the mean and variance statistics with large computational savings compared to SFMC. The MFUQ procedure splits the computational burden by using only few high-fidelity model runs and a large number of lower-fidelity model runs, and then exploiting the correlation between both sets of runs. This study is a valuable contribution to the field of uncertainty quantification in ice sheet modeling. It demonstrates that combining multiple levels of model fidelity can serve to improve uncertainty estimates in useful quantities, which is an approach scarcely used in this field. The science presented in this study uses elaborate statistical techniques, which is a good thing. And I evaluate the scientific aspects of this study positively. However, I believe that major efforts should be made on two presentation aspects. First, more clarity is needed in the presentation of the MFUQ method. I needed to re-read and go back-and-forth between different sections multiple times to really un- derstand the procedure. Second, the authors should try to guide the reader in understanding the procedure, and to provide some intuitive explanations of the different steps in addition to the mathematical details. This latter aspect would better align with the readership of Earth System Dynamics, which is not primarily focused on methodological developments per se. I separate this re- view in one Major comment, focused on the most important clarifications required, and line-by-line comments focused on less important aspects that need elaboration, as well as on scientific aspects that could be slightly adjusted or more thoroughly explored. Line numbers (L) refer to lines in the preprint. Although my review insists a lot on presentation aspects, I find it also important that the science-related comments are addressed. I encourage the authors to revise their manuscript following comments from other reviewers and me. Given the strong scientific basis of this study, I am certain that a revised version of this manuscript will be a valuable contribution to the literature.

Major comment: mathematical presentation. There is no single specific aspect that makes the mathematical presentation unclear. Instead, it is the accumulation of various elements that renders understanding the methods challenging. I try to identify some of these elements here."

We want to make the paper as easily accessible as possible and have rewritten the paper. Specific focus was given to improving clarity and providing intuition wherever possible.

3.1.1 Equations should be better explained and without errors.

36. *"In all equations with matrices, please provide explicitly the dimensions of the matrices involved. This would help to understand, for example, Eqs. 13, 16, 17. It would also be helpful to explicitly mention if a quantity is a scalar, vector, or matrix when it is used for the first time."*

    We added matrices sizes to all relevant equations in the revised document. To further improve clarity, we our revised paper uses bold italics to represent vectors and bold regular font for matrices. This is inline with the instructions on how to denote matrices and vectors n the journal's guide for authors. We apologize for missing these instructions in our previous submission.

37. *"In Eq. 12, the covariance term has twice the same argument, and can therefore not be a covariance. Furthermore, I am not convinced of the validity of the $Var(Q^{\sigma^2})$ formula, so please provide a reference and/or a detailed derivation in the response."*

    Eq 12 is a valid covariance. $\mathbb{C}\text{ov}\left[X, Y\right] = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(Y - \mathbb{E}\left[Y\right])\right]$ and $\mathbb{C}\text{ov}\left[X, X\right] = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(X - \mathbb{E}\left[X\right])\right] = \mathbb{V}\left[X\right]$ is just a special case that occurs when $X = Y$. We want to use the notation $\mathbb{C}\text{ov}\left[X, X\right]$ so to emphasize that the covariance is a matrix when the random variable X is a vector. However since $X = (f_\alpha - \mathbb{E}\left[f_\alpha\right])^2$ is only a scalar we replaced $\mathbb{C}\text{ov}\left[(f_\alpha - \mathbb{E}\left[f_\alpha\right])^2, (f_\alpha - \mathbb{E}\left[f_\alpha\right])^2\right]$. We also added the following note after equation 15 that states:

    Note that, in (14) and (15), and the remainder of this paper we use $\mathbb{C}\text{ov}\left[X, X\right]$ as long hand for $\mathbb{V}\left[X\right]$ to emphasize that the covariance is a matrix when the random variable X is a vector.

    A detailed derivation of the expression can be found in (Dixon et al., 2023). See equation 2.3 and the proof of proposition 3.4 on page 10 of their ARXIV manuscript. We added the following to the paper:

    A detailed derivation of the expression for $\mathbb{V}\left[Q_\alpha^{\sigma^2}(\Theta)\right]$ can be found in (Dixon et al., 2023).

38. *"In L298, to be valid, this equation requires some normalizing terms (1/N0,1/N0,1/N1)."*

    We corrected these equations.

39. *"In the first part of Eq. 16, one term should be $\mathbb{C}ov\left[\Delta_\alpha^{\sigma^2}, \Delta_\beta^{\mu}\right]$."*

    Fixed.

40. *"In Eq. 11, both Q_α and Q are referred to as MC estimators (in Eq. (10) and on L271, respectively)."*

    We now state the following before equation 11.

    Consequently, any MC estimator $Q_\alpha(\Theta)$ of an exact statistic Q, such as $Q_\alpha^{\mu}(\Theta)$ and $Q_\alpha^{\sigma^2}(\Theta)$, is a random variable and the mean-squared error (MSE)

41. *"It would be nice define the MC estimator precisely, as well as the quantity that it is estimating."* Eq. (16) and Eq. (17) define the MC estimators of the mean and variance of mass change, respectively. We now reference these equations throughout the paper to make clear which quantities we are talking about as suggested comment 44.

42. *"Please be consistent in the notation. For example, QACV is bold in Eq. 15, but not in Eq. 20."*

    Fixed.

43. *"Please use equation numbers for all equations."*

    We would to only number equations if they are referenced in the text. I cannot find guidance online about the journals requirement to number all equations. We will do so if the editor requests we do so. However, we have increased the number of numbered equations to allow the reader to easily refer back to relevant equations, when discussing them in the text to address comment 44.

44. *"Throughout the text, try as much as possible to refer to the relevant equations and/or mathematical variables. That would be incredibly helpful for the reader to understand the methods more easily. For example, refer to Eq. 15 every time the "ACV estimator covariance" is mentioned, refer to L197 when mentioning sampling from the prior, refer to L227 when mentioning sampling from the posterior. And there are many more instances, which I will not enumerate here. But I encourage the authors to look for every instance where the reader would benefit from knowing clearly which quantity or equation a certain statement relates to."*

    We have added references throughout the revised paper.

3.1.2 Adding some intuitive explanations

"Here and there, it would help to add a simple sentence to give a better intuition about some concepts. I provide a few examples here below. Again, this is not an exhaustive list. So, I encourage the authors to actively look for similar statements, equations, or paragraphs that could benefit from some intuitive explanations."

We added intuitive explanations whenever suggested by the reviewer. We also took the opportunity to provide intuitive explanations at other places in the text. Any such changes are marked in red in the revised document.

45. *"Towards the beginning of the manuscript, please provide one short paragraph to explain what the statistics of interest are, and why they are uncertain. I believe that all readers might not intuitively understand the concept of variance of a variance."*

    We added the following statement to the introduction: *"However, the substantial computational cost of evaluating ice-sheet models limits the number of model simulations that can be run, and thus the accuracy of uncertainty estimates."* For example, when estimating the mean of a model with Monte Carlo, the mean squared error (MSE) in the estimated value only decreases linearly as the number of model simulations increases.

    We also added the following statement in Section 4.1:

    MC estimators converge to the true mean and variance of f_α as the number of samples tends to infinity, but using a finite number of samples N introduces an error into the MC estimator that depends on the sample realizations used to compute the estimators. That is, two different realizations of N parameter samples Θ, and the associated QoI values, will produce two different mean and variance estimates (see Figure 4). Consequently, any MC estimator $Q_\alpha(\Theta)$ of an exact statistic Q, such as $Q^\mu_\alpha(\Theta)$ and $Q^{\sigma^2}_\alpha(\Theta)$, is a random variable.

46. *"L228: Please add one or two sentences to explain that $g(\theta)$ can be computed without time stepping model solves, and why this is the case."*

We added the following text. After line 228. We were able to calibrate the model using only a steady model without time-stepping because we assumed that the velocity data were collected when the ice sheet which was in equilibrium.

47. *"L247: Please explain that Σ_{post} characterizes the balance between the prior uncertainty in the friction field estimate, and the model-observation mismatch weighted by the observational noise."*

    We added the following statement at the end of section 3:

    The posterior characterizes the balance between the prior uncertainty in the friction field and the model-observation mismatch, weighted by the observational noise. In the limit of infinite observational data, the posterior distribution will collapse to a single value. However, in practice when using a finite amount of data, the posterior will only change substantially from the prior in directions of the parameter space informed by the available data, which were captured by our low-rank approximation.

48. *"L298: Why is this valid regardless of how truthfully f_1 approximates f_0?"*

    Please see response 15 above.

49. *"L322: What do the control variates represent?"*

    We now state:

    The estimator of the low-fidelity mean $Q_1^\mu(\Theta_0)$ is referred to as a control variate because it is a random variable, which is correlated with the random estimator $Q_0^\mu(\Theta_0)$, and can be used to control the variance of that high-fidelity estimator. The term $Q_1^\mu(\Theta_1) \approx Q_1^\mu$ is an approximation of the true low-fidelity statistic Q_1 that is used to ensure that the ACV estimator is unbiased, i.e. $\mathbb{E}_\Theta\left[Q_{\text{ACV}}^\mu(\Theta_0,\Theta_1)\right] = \mathbb{E}_\Theta\left[Q_0^\mu(\Theta)\right] + \eta\left(\mathbb{E}_\Theta\left[Q_1^\mu(\Theta_0)\right] - \mathbb{E}_\Theta\left[Q_1^\mu(\Theta_1)\right]\right) = Q_0^\mu + \eta\left(Q_1^\mu - Q_1^\mu\right) = Q_0^\mu$.

50. *"Figure 4: How do results between these different sample allocation strategies differ? For example, does one approach prioritize minimizing the diagonal entries of the ACV covariance, versus another better constraining the correlation between different models?"*

    See our response to comment 16.

3.2 Line-by-line comments

51. *"General (1): The text would benefit from the use of many more commas. I encourage the authors to, at least, double the number of commas in the manuscript. The general writing level is good, so I have no doubt that the authors can find sentences that need (or would benefit) from commas."*

    While revising the document, we broke compound sentences into multiple sentences where appropriate and used additional commas for complex sentences where needed.

52. *"General (2): The quality of the figures is low. Color scales should be more informative, units should be provided, span of y axes should be appropriate for the range of values shown, a scale in km should be added when showing Humboldt, labels should be added to colorbars, etc. "*

    We regenerated all figures in the document to improve readability.

53. *"L4 Here and elsewhere, the term "accuracy" is used very loosely, and encompasses a wide range of concepts. When used to describe a model degree of fidelity, please always use "fidelity" since this is the technical term used for the name of the method (MFUQ). When describing the amount of variance, please rather use precision, which is mathematically the inverse of the variance."*

    We now use precision when referring to the accuracy of a statistic. Instances of its use are highlighted in red in the revised document. We also use fidelity when approriate.

54. *"L4 Replace ice sheet by glacier."*

    Fixed.

55. *"L5 The problem size is not "representative" of continental scale studies. Please use more careful wording."*

    We now state:

    The problem size and complexity were chosen to reflect the challenges posed by future continental scale studies while still facilitating a computationally feasible investigation of MSE methods.

56. *"L11 prediction should be plural."*

    Fixed.

57. *"L15 Add report after IPCC."*

    Fixed.

58. *"L15 Ice sheets are all land-based."*

    Thanks. We removed "land-based".

59. *"L26 Throughout the manuscript, affect should be used as a verb instead of effect."*

    Fixed.

60. *"L28 Replace inadequacy by uncertainty."*

    We now state

    In addition, while the comparison of model outputs has been used to estimate uncertainty arising from model inadequacy

61. *"Throughout the manuscript, there is confusion in the wording of "parameters" and "inputs". For example, both terms are used interchangeably to characterize the basal friction field. Please (i) always use the same term for a same meaning, and (ii) clearly define the difference between parameters and inputs in the Introduction."*

    We now always use the term parameters.

62. *"L39 There are also methods that have been developed to reduce the problem dimensionality. Please cite Brinkerhoff (2022)."*

    We added a reference to Brinkerhoff (2022) to Section 3, which discusses Bayesian calibration. The introduction now focuses on quantifying uncertainty in predictions, whereas the focus of the cited paper is on Bayesian inference.

63. *"L48 When using the notion of MSE, it is important to clearly define with respect to which quantity the error is considered. In this study, I believe that the error is considered with respect to the expec- tation of the mass change from the high-fidelity model with respect to the posterior distribution of the basal friction field. I realize that this is not straightforward to include. But I recommend adding a couple of sentences to give the definition, and possibly explain its meaning."*

We added the following to Section 4.1:

The MSE of an MC estimator, Eq. (18), consists of two terms referred to as the estimator variance (I) and the estimator bias (II). The bias term of the MSE is caused by using a numerical model, with inadequacy and discretization errors, to compute the mass change. More specifically, letting Q_∞ denote the exact value of the statistic of a numerical model with zero discretization error but non-zero model inadequacy error, and Q_0 denote the highest-fidelity computationally tractable model approximation of Q_∞, then the bias can be decomposed into three terms

$$\begin{aligned}(\mathbb{E}\left[Q_\alpha(\Theta)\right] - Q) &= (\mathbb{E}_\Theta\left[Q_\alpha(\Theta)\right] - Q_\infty + Q_\infty - Q_0 + Q_0 - Q) &(3)\\ &= (\mathbb{E}_\Theta\left[Q_\alpha(\Theta)\right] - Q_0) + (Q_0 - Q_\infty) + (Q_\infty - Q) &(4)\end{aligned}$$

The first term is caused by using a model f_α with numerical discretization that is inferior to that employed by the highest fidelity model f_0. The second term represents the error in the statistic introduced by the numerical discretization of the highest-fidelity model. The third term quantifies the model inadequacy error caused by the numerical model being an approximation of reality.

Later in that section we now also state:

Constructing a SFMC estimator of statistics, such as the mean Eq. (16) or variance Eq. (17), with a small MSE ensures that the value of the estimator will be likely close to the true value, for any set of model parameters samples. However, when using numerical models approximating a physical system, constructing an unbiased estimator of Q is not possible. All models are approximations of reality and thus the model inadequacy contribution $Q_\infty - Q$ to the bias decomposition in Eq. (3) can never be driven to zero. Additionally, it is impractical to quantify the discretization error $Q_\infty - Q_0$. Consequently, SFUQ methods focus on producing unbiased estimators of Q_0, such that $\mathbb{E}_\Theta\left[Q_\alpha(\Theta)\right] = Q_0$.

64. *"L50 I think that the authors might not be aware of the study of Bulthuis et al. (2020). Please consider referring to it."*

The study cited uses a surrogate based multi-fidelity method which has no direct relationship to the statistical estimation methods discussed in this paper. Moreover, the paper's reliance on a surrogate means that it can not be applied to ice-sheet models with large numbers of parameters as is the focus of this paper. Consequently, we did not cite this paper in the revised manuscript.

However, we did find the a paper using MFSE on an ice-sheet model Gruber et al. (2022). This paper, uses one type of ACVMF estimator, i.e. MFMC, to estimate uncertainty in a steady-state Blatter-Pattyn model of an ice sheet defined on rectangular a rectangular parallelepiped domain. Specifically, MFMC was used to estimate uncertainty in the L^2 norm of the ice-velocity caused by uncertainty in two parameters, specifically a scalar representing basal friction and a variable parameterizing the simple bed topography.

We added a citation to this work in the revised document. Specifically, we now state: Note, Gruber et al. (2022) previously applied MFSE to a ice-sheet model; however, their study was highly simplified, as it only quantified uncertainty arising from two uncertain parameters of an ice-sheet model define on a simple geometric domain.

65. *"L61 I find the changes between past and present tense somewhat confusing. I recommend consistently using a single tense."*

    We tried to improve the consistency of tense wherever possible. In revising our paper we used the following guidelines.

66. *"L85 Model simulations do not only capture the "melting", since they represent the dynamic response of the glacier as well. This should induce changes in the amount of ice flowing out of the simulated domain."*

    We revised the sentence.

67. *"L102-104 In this sentence, the summary of the Stokes and MOLHO models sound identical to me."*

    We now state: In summary, the simpler 2D SSA model is formulated to simulate grounded ice with significant sliding at the bed or ice shelves, while the 3D MOLHO model is designed to capture the evolution of ice sheets over frozen and thawed beds, as well as ice shelves.

68. *"L113 Replace exorbitant by impractical."*

    Fixed.

69. *"L121 Provide a reference for $q = 1/3$."*

    Done.

70. *"L125. Define the $\|$ notation here."*

    Done.

71. *"L139 The ψ term is already multiplied by n above, so this multiplication should not be included in the definition of ψ. Also, why is there an extra term $\rho g(s - z)$ in the boundary condition on Γ_m here compared to the Stokes model?"*

    Fixed.

72. *"L145 $\partial\Sigma$ is not defined."*

    Fixed.

73. *"Figure 2 Show the meshes side-by-side (+ all comments from General (2))."*

    The paper now includes plots of all four meshes side-by-side.

74. *"L172 The statement "one of the largest sources of prediction uncertainty" should be quantified and referenced with a citation."*

    Please see response 7.

75. *"L180 "we set $\mu = 0$". I believe that this is only for the prior. It seems strange to me that the posterior is forced to have zero mean. Please specify."*

    We now state: In this study, we adopted a fully distributed approach that treated the friction as a log-Gaussian random field that is $\theta = \log(\beta)$ with a Gaussian prior distribution $p(\theta) \sim \mathcal{N}(\mu, \mathcal{C})$ with $\mu = 0$.

76. *"L183 There is a switch from C to Σ without mentioning it. Specify that Σ_{prior} is a covariance."*

We now state:

a finite-dimensional discretization of the operator $\Sigma_{\text{prior}} \approx \mathcal{C}$ with

77. *"L185 Why is the source term only integrated over Γ_g and not over Γ_f ? I would assume that snow accu- mulation and surface melting should also be computed over the floating parts of the domain."*

That was a typo, the integrals are over the lower surface of the ice sheet $\Gamma_l = \Gamma_g \cup \Gamma_f$. Fixed.

78. *"L199 Please specify "this Gaussian prior"."*

Fixed.

79. *"L199 Replace "on" by "constrained with"."*

Fixed.

80. *"L207 The authors sort of sweep under the rug the possible influence of ocean melt on their methods. Melt at the boundary can induce strong dynamical responses by a marine-terminating glacier. It can be expected that differences between models of different levels of fidelity would be exacerbated, potentially diminishing the advantages of the MFUQ. Please discuss this more thoroughly in the Discussion."*

We now state in the discussion: Moreover, introducing more complicated physics in the highest-fidelity model, such as calving, could degrade the performance of MFSE. For example, ice melt at the boundary can induce strong dynamical responses in a marine-terminating glacier, which could potentially reduce the correlation between models that do not capture this phenomenon.

81. *"L213 Please clarify why this assumption is required in the procedure. I believe that it is needed to compute the $g(\theta)$ function represented by the Blatter-Pattyn flow model. And that without this assumption, the PDE-constrained optimization cannot be solved."*

We now state: Before continuing, we wish to emphasize two important aspects of the calibration used in this study that mean our results must be viewed with some caution. First, we assumed the observational data to be uncorrelated, as assumed in most ice-sheet inference studies, including (Recinos et al., 2023; Isaac et al., 2015). Moreover, we also assumed our Gaussian error model to be exact. However, neither of these assumptions are likely to be perfect in reality. For example, Koziol et al. (2021) showed that, for an idealized problem, ignoring spatial correlation in the observational noise can lead to uncertainty being underestimated. Second, our optimization of the MAP point was constrained by the coupled velocity flow equations and steady-state enthalpy equation, which is equivalent to implicitly assume that the ice is at thermal equilibrium. Theoretically, this assumption could be avoided if the temperature tendencies were known, but they are not. Alternatively, transient optimization over long time periods, comparable to the temperature time scales, could be used. However, this approach would be computationally expensive and would require including time-varying temperature data (e.g., inferred by ice cores) which are very sparse.

82. *"L217 "However, such approaches ignore the uncertainty in the model parameters due to using a finite amount of noisy observational data". This statement is incorrect. Observational*

uncertainty can be incorporated in cost functions. See for example Eq. (1) from Goldberg (2015).”

That’s correct. We also account for the uncertainty in our deterministic inversion to compute the MAP point. We have rephrased our statement: However, such approaches only use a single optimized parameter value to represent the uncertainty in the model parameters that arises from using a finite amount of noisy observational data.

83. *“L222 “the likelihood distribution”: the likelihood is a function, not a distribution.”*

    We disagree. The likelihood is the probability of the data given the parameters θ $p(y|\theta)$. However, we have dropped the word distribution from the sentence.

84. *“L232 Please add a justification for this choice of α.”*

    We expanded Remark 5.1 to discuss how we heuristically chose parameter α.

85. *“L233 Please specify “samples from the posterior of* $\log(\beta)$*”.”*

    Fixed.

86. *“L249 Again, this statement is likely not obvious to most readers. At first sight, the computation that is referred to here is a simple addition of two matrices $(H_{MAP} + \Sigma^{-1})$ prior explain why this is intractable would be beneficial.”*

    We rephrased the sentence and explained why it is intractable to compute H_{MAP} and invert high-dimensional dense matrices.

87. *“L254 and 255 Replace ice sheet by glacier.”*

    Fixed.

88. *“L263 What do the authors mean by “robust”?”*

    We now state:

    SFMC quadrature is a highly versatile procedure that can be used to estimate a wide range of statistics for nearly any function regardless of the number of parameters involved.

89. *“L264 “three-step””*

    Fixed.

90. *“L265 The m superscript should be n (which would preferably be another letter than n, see Major comment).”*

    We changed m to n. However, we like to use capital N to denote the number of samples and lower case n to denote the index $n = 1, \ldots, N$.

91. *“L266 Specify “basal friction field”.”*

    Fixed.

92. *“L278 “The bias term of the MSE (11) is caused by using a numerical model, with inadequacy and discretization errors, to compute the mass change.” Here also, I ask for clarification: bias with respect to what? If it is with respect to observations, then observational uncertainty should also be discussed. If it is with respect to the highest-fidelity model, then the latter is also a “numerical model”, and the sentence is inappropriate. If it is with respect to the unknown*

true dynamical behavior of Humboldt glacier, then there is a philosophical question of how to compute a mean squared error with respect to a quantity that cannot be known."

See our response to comment 63.

93. *"L287 Typo estimated."*

    Fixed.

94. *"L296 Two-model"*

    Fixed.

95. *"L316 QoI is not defined."*

    Fixed.

96. *"L324 Concerning $\Theta_\alpha^* \cup \Theta_\beta = \emptyset$. (i) I believe that $\cup$ should be $\cap$, (ii) I believe that "for $\alpha \neq \beta$" should be specified."*

    We have added additional clarifying text.

97. *"Eq. (16) Is $\mathbb{C}ov\left[Q_0, \Delta_0\right](\Theta_{ACV})$ a covariance matrix? If so, it should be symmetric. However, the $(0,1)$ and $(1,0)$ entries of the right-hand-side seem different to me. Please explain."*

    The quantity is a covariance matrix. It is symmetric in the arguments μ and σ^2.

98. *"L351 "following standard practice": provide citation."*

    We added a citation

99. *"L352 Please add an additional explanatory statement, for example: This involves computing the high- fidelity and all the low-fidelity models for the same set of samples Θ pilot."*

    We added the following statement This involves computing the high-fidelity and all the low-fidelity models at the same set of samples θ_{pilot}.

100. *"L360 Specify "introduce sampling errors"."*

     Fixed.

101. *"L367 I believe that the same should be specified for $\alpha \cup \beta^*$ and $\alpha \cap \beta^*$"*

     Fixed.

102. *"L382 There is no verb in this sentence."*

     Fixed.

103. *"L391 Please add an additional explanatory statement, for example: This can happen if a subset of the low-fidelity models correlate much better with the high-fidelity model than the rest of the low-fidelity models, for example."*

     We added the statement:

     This occurs when a subset of the low-fidelity models correlate much better with the high-fidelity model than the rest of the low-fidelity models. For instance, some low-fidelity models may fail to capture physical behaviours that are important to estimating the QoI.

104. *"L394 I think this should be estimator types."*

     Fixed.

105. *"L394 "model models subsets" is either a typo, or very confusing language."*

We fixed the typo.

106. *"L402 was should be were."*

Fixed.

107. *"L413 ice-sheet should be glacier."*

Fixed.

108. *"L415 Typo an an."*

Fixed.

109. *"L424 Specify MALI ice-sheet code with the Blatter-Pattyn flow model."*

See response 182.

110. *"Figure 9 I provide here a concrete example of how to help the reader navigate through the technical details of the study. The caption should specify: "... MAP point (θ_{MAP} in Eq. (9)) ... prior variance (Σ_{prior} in Eq. (xxx)) ... posterior variance (Σ_{post} in Eq. (xxx))". Using more such links between text and mathematics would really help reading the study."*

We have added symbols and equation references throughout the paper.

111. *"L437 "speeds up as it thins": I think that this statement is incorrect, although I see what the authors mean. A glacier does not speed up because of thinning. It speeds up because of increasing surface slope, caused by enhanced thinning at the front. Also, the inverted relation holds: as a glacier speeds up, it discharges more ice into the ocean, leading to thinning."*

We now state In general, the glacier speeds up as negative surface mass balance causes the surface to steepen near the terminus. The largest speedup occurs in the region of fast flow in the north where basal friction is small.

112. *"Remark 5.2 I believe that this is an important scientific aspect, which is also somewhat swept under the rug. In their results, the authors demonstrate that the simulated mass change is sensitive to high-frequency variability in the basal friction field. As such, the interpolation method from fine to coarse meshes is potentially very influential. Which interpolation method has been used here? If it is simple linear interpolation, then all the high-frequency variability will be smoothed out. This would affect the behavior of low-fidelity models with coarser meshes. I recommend that the authors try interpolation methods that better preserve high-frequency variability (nearest neighbor, or maybe polynomial interpolation) and evaluate the impacts on their results."*

We used the finite element mesh of the high-fidelity model to interpolate the basal friction from that mesh onto the coarser meshes. We added a note to this effect to the text.

However, we do not believe that a different interpolation strategy is needed. Our results show that multi-fidelity models are able to use coarser meshes despite those meshes not being able to accurately representing the basal friction. See Figure 13 in the original submission. We make this point on paragraph starting on line 551 in the original submission.

113. *"L458 "significant differences": the word significant is misused here, because no statistical test has been performed. If a statistical test has been performed, please specify which one, and provide p-values. Furthermore, by eye, the differences do not seem very large in Figure*

8 compared to the standard deviations. However, this is difficult to say because of the terrible choice of y-axes span in Figure 8, which should be changed."

See response 27.

114. *"L460 The meaning of accuracy is not clear here (see comment on L4)."*

See response 53.

115. *"L476 Provide a citation to support this statement."*

See response 29.

116. *"477. Please quantify "the error introduced". "not insignificant": this wording is misused here, because no statistical test has been performed. If a statistical test has been performed, please specify which one, and provide p-values."*

We now state:

Moreover, we found that the error introduced by using a small number of pilot samples can be substantial, yet it is typically ignored in existing literature.

117. *"L479 Please specify the number of realizations per bootstrap. From the rest of the text, I believe that it is 20 realizations per bootstrap samples, but this should be clarified explicitly."*

Fixed.

118. *"Figure 10 (1) I am puzzled by the very high upper bound on the variance reduction of the variance. In the ratio, the SFMC variance estimator is the denominator, which should therefore be the same for all the bootstrap samples. As such, the very high upper bound is caused by an unrealistically low estimated ACV variance via Eq. (15). This leads me to the question: is the approximation on pilot samples via Eqs. (15,20) unstable when using bootstrap with replacement? In any case, please provide an explanation about the very high value of the 95% quantile."*

We now report the 10% and 90% quantiles to more clearly show how variance reduction distributes. We also refer the reviewer to the statement on page 26 *"However, the box plots in Figure 10a highlight that using only 20 samples introduces a large degree of uncertainty into the estimated variance reduction."* We also refer the reviewer to the statement on page 28 *"While, increasing the number of pilot samples decreased variability, we believed that the benefit of further increasing the number of pilot samples would be outweighed by the resulting drop in the variance reduction."*

119. *"Figure 10 (2) It is not immediately clear why a same model combination would give different estimates of the variance reduction, since the ice sheet models are deterministic. If I understand correctly, some of this variability comes from the random bootstrapping within the pilot samples, and some of the variability comes from the ACV estimator selected (MLMC, MFMC, ACVMF). Is it possible to quantify how these two sources of variability compare? And in turn, is it possible to quantify how much of the boxplot spread in Fig. 10a is due to these two factors versus the fact that different subsets of low-fidelity models have been selected?"*

The variability in the plots is induced entirely by the bootstrapping. This plot was included to demonstrate that using a small number of pilot samples introduces a non-trivial error. Each estimator does have a different estimator variance. However, the box plots just report the smallest estimator variance, across all estimators, for each bootstrapped sample.

We now state in the paper: Please note that, while we enumerate over numerous estimators, each with a different variance reduction, the variability in the plots is induced entirely by the bootstrapping procedure we employed. The box plots report the largest variance reduction, across all estimators, for each bootstrapped sample.

120. *"L489 Specify subsets of model combinations."*

Fixed.

121. *"L491 the original 20 pilot samples are used."*

Fixed.

122. *"L491 Specify were determined useful to include for reducing. (Probably that individually, all the models would be useful. But they are not relative to including other better-correlated or computationally- cheaper models.)"*

We now state:

Moreover, bootstrapping the estimators also revealed that using all models simultaneously to reducing the variance of the ACV estimator was not as effective as using a smaller subset of models.

123. *"L496-499 I could not understand the end of this paragraph. It would be helpful if the authors defined the notion of hierarchical relationship."*

See response 30.

124. *"Figure 11. Please specify the number of samples for each case (2, 3, and 4 models)."*

The number of samples allocated to each model depends on the bootstrapped realization of the pilot data to compute the pilot statistics. Consequently, there is no one number that we can provide.

125. *"L520 Again, the meaning of "accurate" is not well-defined."*

We now state: MSE of the final ACV estimator we would construct would be much smaller than the MSE of a SFMC estimator of the same cost because even the smallest variance reduction was greater than 14.

126. *"L520 "even the smallest variance reduction was greater than 20". This is not what is shown in Fig. 11. Certainly not for the cases of 2 and 3 models. And for the case of 4 models, it seems to me that even the 5% quantile is below 20, suggesting that the smallest value is definitely smaller than 20."*

See our response to comment 125.

127. *"L522 Replace that by which (with a comma, see General comment (1))."*

Fixed.

128. *"L523 The three models listed do not include $MOLHO^*_{1km,9days}$. As such, I believe that it corresponds to the case "4 models" in Fig. 11. I find the discrepancy between the number of low-fidelity versus the total number of models confusing. Please use a consistent manner to quantify the number of models used."*

Throughout the paper, We now always include the highest-fidelity model when counting the number of models used by an estimator. The particular statement highlighted by the reviewer

had a typo, which we corrected. Three models were chosen despite the estimator being allowed to use four.

We now state in the paper: Note, an estimator allowed to choose four models may still choose less than four models, which will happen when some of those models are not highly-informative.

129. *"L525 Please remind the readers where these numbers come from."*

See response 32.

130. *"L527 I do not see any right or left panel."*

Fixed.

131. *"535 Please specify another estimator (i.e., MLMC or ACVMF)."*

Fixed.

132. *"In Discussion. This question relates to my curiosity concerning the complementarity between this method and stochastic ice sheet modeling. Here, the MFUQ samples uncertainty from a single time-constant uncertain input. In contrast, stochastic modeling (e.g., Verjans et al., 2022) samples uncertainties between multiple correlated uncertain inputs, and at different time steps (for example SMB variability in time is prescribed as stochastic). However, since the statistical properties of the time-varying stochastic inputs (i.e., the auto-correlation, the covariance structure and the mean of each stochastic input) can be specified a priori, I suppose that, in theory, the MFUQ method could be applied. But I wonder if this is practically feasible. I think that the Discussion would benefit from a short paragraph about this point."*

We added the follwing to the discussion:

This study focused on investigating the efficacy of using MFSE to accelerate the quantification of parametric uncertainty using deterministic ice-sheet models. We did not quantify the uncertainty arising from model inadequacy. Recently Verjans et al. (2022), attempted to quantify model uncertainty by developing stochastic ice-sheet models designed to simulate the impact of glaciological processes that exhibit variability that cannot be captured by the spatiotemporal resolution typically employed by ice-sheet models, such as calving and subglacial hydrology. The MFSE algorithms presented in this paper can be applied to such stochastic models, by sampling the model parameters and treating the stochasticity of model as noise. However, the noise typically reduces the correlation between models and thus the efficiency of MFSE (Reuter et al., 2024). Moreover, this study only focused on estimating the mean and variance of mass change. Consequently, the efficacy of MFSE may change when estimating statistics – such as probability of failure, entropic risk, and average value at risk (Rockafellar and Uryasev, 2013; Jakeman et al., 2022) – to quantify the impact of rare instabilities and feedback mechanisms in the system. We anticipate that larger number of pilot samples than the amount used in this study will be needed to estimate such tail statistics, potentially reducing the efficiency of MFSE.

133. *"L567 Appendix B."*

Fixed.

134. *"Figure 13 (1) Changing the color scale here is absolutely necessary."*

It was difficult to find a color scheme that more clearly highlighted the difference. However, we believe the other two plots in Figure 13 supported our argument that the models did not produce exactly the same predictions. Consequently, we removed the left panel of Figure 13.

135. *"Figure 13 (2) If I understand correctly, the basal friction field should be model-independent. The differences only stem from the interpolation method. This should be specified in the caption. Furthermore, this Figure seems to confirm my comment about Remark 5.2."*

See response 112

136. *"L589 "variance" should be standard deviation here, since Gigaton units are specified."*

Fixed.

137. *"L590 "significance": the word significant is misused here, because no statistical test has been performed. If a statistical test has been performed, please specify which one, and provide p-values. Furthermore, even the meaning of "the significance of these numbers" is not clear to me."*

We now state:

However, the exact values of these statistics were impacted by our modeling choices.

138. *"L593 In this study, the basal friction field distribution was derived assuming that all other variables were perfectly known. In reality, different sources of uncertainty can mix. Please cite Gudmundsson and Raymond (2008) and add one or two sentences about this to the Discussion."*

We added further discussion of the limitations of our approach. For more details, see our response to comment 80.

139. *"L614 Please mention here that this study explores the use of MFUQ for low-order moments only. One can wonder if this method can be used for statistics such as skewness or quantiles in the tails of the distribution. This can be particularly important for evaluating the response to an input that could introduce instabilities and feedback mechanisms in the system, such as ocean or SMB forcing."*

We have added the following to the discussion:

Moreover, this study only focused on estimating the mean and variance of mass change. Consequently, the efficacy of MFSE may change when estimating statistics such as probability of failure, entropic risk, average value at risk, etc. to quantify the impact of rare instabilities and feedback mechanisms in the system. We anticipate that larger number of pilot samples, than used here, will be needed to estimate such tail statistics, thus potentially reducing the efficiency of MFSE.

140. *"L616 Here, and in many other instances, the authors insist about the fact that MFUQ can be used at continental scale to estimate uncertainty on ice-sheet mass change statistics. However, such a statement is not well-supported by their results. Just looking at the results, one can argue that the MFUQ framework presented here requires 36 CPU days for a single glacier. Scaling this linearly to the Greenland ice sheet results in $O(1-10)$ years of computation. Thus, there should be a slightly more in-depth explanation of why MFUQ is applicable for studies at the ice-sheet-scale."*

The following statement from the original paper, is one example that raised the reviewers concerns:

"Thus, MFSE reduced the cost of estimating uncertainty from over two and a half years of CPU time to just over a month, assuming the models are evaluated in serial."

To address the reviewers concern, we had added the following statement to the discussion.

Note that while applying MFSE to the Humboldt Glacier took over a month of serial computations, the clock time needed for MFSE can be substantially reduced because MFSE is embarrassingly parallel. Each simulation run in the pilot stage can be run in parallel without communication between. Similarly, for the exploitation phase. Moreover, each simulation can be computed in parallel. Consequently, while using MFSE for continental scale UQ studies may require years of serial CPU time, distributed computing could substantially reduce this cost, potentially one to two orders of magnitude. The exact reduction would depend on the number of CPUs used.

141. *"L618 "substantially": please quantify and provide a citation."*

We now state: Mass loss from ice sheets is anticipated to contribute 10s of cm to sea-level rise in the next century under all but the lowest emission scenarios (Edwards et al., 2021).

142. *"L638 I do not understand the underlying meaning of this sentence. Please expand or remove it."*

We now state:

Moreover, while the utility of the lower-fidelity models ultimately chosen for MFSE were not clear at the onset of the study, we were still able to estimate uncertainty at a fraction of the cost of single fidelity MC. This was achieved despite the need to conduct a pilot study that evaluated all models a small number of times.

143. *"L641 Antarctica and Greenland."*

Fixed.

144. *"L661 Again, the meaning of accurate is unclear here. It would be more correct to explain that the approximation level depends on the variance retained in the truncation."*

We now state: In order to compute a low-rank approximation of the matrix T we truncated the spectral decomposition of $\mathbf{W} = \mathbf{U}\left((\mathbf{\Lambda}+\mathbf{I})^{-\frac{1}{2}} - \mathbf{I}\right)\mathbf{U}^\top$ by discarding the eigenvalues λ_i such that $\left|1 - \frac{1}{\sqrt{\lambda_i+1}}\right| \ll 1$. This ensured that the low-rank approximation of T well approximated T in the spectral norm sense.

145. *"L668 Please define K here as well. Otherwise, the reader needs to go back to the main text."*

Fixed.

146. *"L674 I do not see why the representation is "bi-Laplacian". I wonder if this term is not inadvertently misused here. Could this please be clarified? I believe that applying the Laplace approximation has no link with the bi-Laplacian operator, but sorry if I am misunderstanding here."*

The bi-laplacian is used to define the prior-distribution of the log normal basal friction field. We dropped the mention of bi-laplcian to avoid confusion.

147. *"L685 Typo: in this study"*

Fixed.

148. *"L686 Typo: modes"*

Fixed.

149. *"L700 I believe that MF estimator should be ACV estimator"*

Fixed.

150. *"L701 I believe that MFUQ estimator should be ACV estimator"*

Fixed.

151. *"L702 This should be: The mean and variance bootstrapped (...)."*

Fixed.

152. *"L706 This should be: the uncertainty in the mean mass change (...)"*

Fixed.

153. *"L706-708 Please refer to Figure B5."*

Fixed.

4 Reviewer 3 (Dan Goldberg)

4.1 General comments

"The manuscript, An evaluation of multi-fidelity methods for quantifying uncertainty in projections of ice-sheet mass change by Jakeman et al, uses a new computational approach to determining the posterior uncertainty of ice mass change in a glacier forecast conditioned on observational data and uncertainty. The main contribution is a Multi Fidelity Uncertainty Quantification (MFUQ) scheme which samples from a probability distribution (see below) and provides an inexpensive means of Monte Carlo variance reduction in the calculated statistics that requires far less simulation time. This is achieved through generating ensembles from models that are of lower fidelity (coarser resolution / longer time steps) whose dependencies on the input parameters are similar. The probability distribution which is sampled – that of the sliding parameter conditioned on observations and model physics – would be too expensive to find via Monte Carlo methods. Rather, a method introduced by others in the literature – which approximates this posterior as Gaussian and finds a low rank approximation to the inverse covariance matrix to make the problem tractable – is used.

The methodology introduced in the paper – the MFUQ scheme – is fairly well described and seems quite useful, and its results deserve to be shown.

However, there are a number of major issues I have with the manuscript. Aside from a number of writing issues, such as inconsistent statements and introducing of terms and symbols without definition or explanation (see specific comments), I feel that the messaging of the paper in the introduction is not in line with what the authors have actually done. Furthermore they have downplayed or overlooked recent works in the literature – works which, in some cases, bring the methodology of this study into question. I will highlight these in general comments below.

Finally I should point out first though that it monte carlo methods are not my area of expertise. I have some specific comments about certain things that looked as thought they might be typos or need more explanation. Largely however I do not have much to say about the actual MFUQ methodology and its presentation, and I hope that other referees can assess it better. "

4.2 High-level comments

154. *"The paper sets out to deal with parametric uncertainty, which is the case. But the introduction is written in a way that makes it seem that MFUQ is used to solve the "full" problem – that is, quantifying the probability density of mass change conditioned on the model and observations, which can be termed $p(Q|m,U)$ where Q is the mass change, m is the model and U is the observations. But in truth a different method (Hessian-based) was used to find the posterior density of the frictional field q, and then this was sampled from to find the posterior of Q i.e. $p(Q|m,q)p(q|m,U)$ – and $p(Q|q)$ is the only component being determined by MFSE. I think this could be potentially very misleading and give the impression that MFSE is capable of the "full" problem when from the results of the paper it definitely is not. This is very important: given the newness of the fields of ice-sheet modelling and ice-sheet uncertainty quantification there is extensive misunderstanding about which problems can be tackled by sampling methods and which require alternative methods. Although this is somewhat covered in lines 61-71 of the manuscript, the passage requires familiarity with the field and with both MC and Hessian-based UQ. It needs to be much more clear – with mathematical formality – which distribution is being quantified using MFSE."*

We now state the following in the introduction:

"This study investigated the efficacy of using MFUQ methods to reduce the computational cost needed to accurately estimate statistics summarizing the uncertainty in predictions of sea-level rise obtained using ice-sheet models parameterized by large numbers of inputs." To facilitate a computationally feasible investigation, we focused on reducing the computational cost of estimating the mean and variance of mass change in the Humboldt Glacier in northern Greenland. This mass change was driven by uncertainty in the spatially varying basal friction between the ice sheet and land mass, under a single climate change scenario between 2007 and 2100. Specifically, letting f denote the mass change at 2100 computed by a mono-layer higher-order (MOLHO) (Dias dos Santos et al., 2022) model $\mathcal{M}$, $\boldsymbol{\theta}$ denote the parameters of the model characterizing the Basal friction field, and $\boldsymbol{y}$ denote the observational data, we estimated the mean and variance of the distribution $p(f \mid \mathcal{M}, \boldsymbol{y}) = p(f \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathcal{M}, \boldsymbol{y})$ in two steps. First, using a piecewise linear discretization of a log-normal basal friction field, we used Bayesian inference to calibrate the resulting 11,536 dimensional uncertain variable to match available observations of glacier surface velocity. Specifically, we constructed a low-rank Gaussian approximation (Isaac et al., 2015; Recinos et al., 2023; Barnes et al., 2021; Johnson et al., 2023; Perego et al., 2014) of the Bayesian posterior distribution of the model parameters $p(\boldsymbol{\theta} \mid \mathcal{M}, \boldsymbol{y})$ using a Blatter-Pattyn model (Hoffman et al., 2018). Second, we estimated the mean and mass of glacier mass change using 13 different model fidelities (including the highest-fidelity model), based on different numerical discretizations of the MOLHO and shallow-shelf (SSA) physics approximations (Morland and Johnson, 1980; Weis et al., 1999).

155. *"The manuscript is also misleading about contributions in this paper versus in the literature. Specific examples are given below, but the manuscript does not acknowledge previous authors' attempts to quantify the uncertainty of high-dimensional parametric uncertainty. In particular, a recent paper in The Cryosphere (Recinos et al, 2023, hereby shortened as BR23) has been overlooked. The authors can certainly be forgiven for this of course as the paper came out only last year, but it is extremely relevant to many of the assumptions and calculations within the manuscript (and is mentioned extensively in the specific comments below). Additionally, based on this paper there are several assumptions and/or approximations that give me serious reservations about this paper's results – these are easily identifiable in the specific comments*

where BR23 is mentioned."

We apologize for missing the work in BR23. We agree that the paper is highly relevant to this manuscript. We have edited the paper to address the reviewers reservations about the results we present. First, in our response to comment 156 we show that despite linearizing when computing the posterior distribution of the parameters, the parameter-to-QoI map is nonlinear. Second, we acknowledge that BR23 presents a more rigorous approach to setting the hyper-parameters of the prior imposed on the Basal friction field which is then used during Bayesian inference and we now state the positive benefits of this approach in multiple places in our paper. However, aside from our approach for choosing the prior hyper-parameters, the method we used for Bayesian inference, which is also used in BR23, is state-of-the-art in the ice-sheet community. Moreover, the goal of our study was to demonstrate the efficacy of MFSE on a ice sheet problem that has challenges representative of papers used to predict ice-sheet evolution and not to produce scientifically meaningful values for sea-level rise. We detail the limitations of our study in the discussion. Finally, we do not believe our approach and BR23 are mutually exclusive. Indeed, we believe that when adjoints are available in a ice sheet code, then the linearization approach in BR23 could be used to provide a highly-computationally efficient and accurate low-fidelity model that could be used to further increase the accuracy and computational gains of MFSE reported in this manuscript.

We now state in the introduction:

Most recent studies have focused on estimating uncertainty in the predictions of ice-sheet model with small numbers of parameters, e.g. (Nias et al., 2023; Ritz et al., 2015; Schlegel et al., 2018; Jantre et al., 2024), despite large numbers of parameters being necessary to calibrate ice sheet model to observations (Barnes et al., 2021; Johnson et al., 2023; Perego et al., 2014). However, recently Recinos et al. (2023) used the adjoint sensitivity method to construct a linear approximation of the map from a high-dimensional parameterization of uncertain basal friction coefficient, and ice stiffness, to quantities of interest (QoI) – specifically the loss of ice volume above flotation predicted by a shallow-shelf approximation model at various future times. The linearized map and the Gaussian characterization of the distribution of the parameter uncertainty was then exploited to estimate statistics of the QoI. While this method is very computationally efficient, linearizing the parameter-to-QoI map will introduce errors (bias) into estimates of uncertainty, which will depend on how accurately the linearized parameter-to-QoI map approximates the true map (Koziol et al., 2021). Moreover, the approach requires using adjoints or automatic differentiation to estimate gradients, which many ice-sheet models do not support. Consequently, in this study we focused on Multi-fidelity statistical estimation (MFSE) methods that do not require gradients.

We removed our claim that we were the first to compute uncertainty in QoI when using Bayesian inference to calibrate a large number of model parameters. We now state in the introduction:

Our study makes two novel contributions to previously published glaciology literature. First, it represents the first application of MFSE methods to quantify the impact of high-dimensional parameter uncertainty on transient projections of ice-sheet models defined a realistic physical domains. Our results demonstrate that MFSE can reduce the serial computational time required for a precise UQ study of ice-sheet contribution to sea-level rise from years to a month. Note, Gruber et al. (2022) previously applied MFSE to a ice-sheet model; however, their study was highly simplified, as it only quantified uncertainty arising from two uncertain parameters of an ice-sheet model define on a simple geometric domain. Second, our paper

provides a comprehensive discussion of the practical issues that arise when using MFSE, which are often overlooked in all MFSE literature.

This statement also includes reference to an earlier attempt at using MFSE with ice-sheet models. Please refer to our response to comment 64 for more details on the limitations of that study.

156. *"The underlying premise of the paper is that, given a Hessian-based approximation of the posterior parameter density has already been carried out, "traditional" means of sampling from this posterior density is too expensive. But another such approach – using the linearization of the mass change model f(q) (using either Automatic Differentiation or some other form of differentiation) to project the posterior uncertainty of q onto the quantity of interest – exists, and is not at all mentioned. Playing devil's advocate, such an approach assumes near-linearity of f(q), but linearity has already been assumed in the posterior calculation of q. Moreover at least two prior papers – Isaac et al (2015) and BR23 – have used this method (see eq. 24 of Isaac et al 2015, or eq. 15 of BR23), and the latter comprehensively tested the linearity assumption. Given this, I would expect acknowledgement of this very relevant and related approach, its drawbacks and benefits, and fit (or lack thereof) to the current problem"*

Again, we apologize for missing the work in BR23. We agree that the method we used to compute the Laplace approximation of the posterior (also used in Isaac et al (2015) and BR23 is only exact if the model used to predict the observations is linear. This approach was only computationally feasiable for us (and in the other papers) because we could solve adjoint equations to compute the action of the Hessian of the misfit (between the model and the observations) on a vector using our steady-state model implemented in MALI. However, our codes do not have the ability to solve adjoint equations to compute gradients of the transient model used to predict mass change at year 2100. Consequently, we could not linearize the parameter-to-QoI map as done in BR23. A forward finite difference approximation of the gradient would require 11,537 model evaluations. However, if the parameter-to-QoI could be linearized computationally efficiently, using it to compute the mean and variance of mass change would introduce an error because the map is nonlinear see Figure 1). Specifically, Figure 1) plots one-dimensional sweeps through the 11,536 dimensional parameter space used to represent Basal friction computed using the lowest fidelity model in our 13 model ensemble, that is $\text{SSA}_{3km,365days}$. Each sweep is along a random direction through the parameter space that pass through the origin (which corresponds to using the mean friction field). The extremes of the sweep correspond to $\pm\sigma$, where σ is the standard deviation of the posterior along the sweeps. It is clear that the parameter-to-QoI map is nonlinear but because we cannot linearize the map we cannot compute the error introduced when using it to estimate the mean and variance of mass change. However, if a linearized map was available we could use it as an additional computationally efficient low-fidelity model when using MFSE to compute statistics. The exact benefit of doing so would depend on the correlation between the new model and the other models we used in the paper.

We added the following remark to Section 5.1 of the revised manuscript:

The models we used here are all different numerical discretizations of the two different physics models. However, in the future we could also use alternative classes of low-fidelity models if they become available. For example, we could use linearizations of the parameter-to-QoI map (as done in (Recinos et al., 2023), if our MOLHO and/or SSA codes become capable of efficiently computing the gradient of the map by solving adjoint equations or by using automatic differentiation. Such an approach would require only one non-linear forward transient

solve of the governing equations followed by lone linear solve of the corresponding backward adjoint. Once constructed, the linearized map could then be evaluated very cheaply and used to reduce the MSE of the MFSE estimators, provided the error introduced by the linearization was not substantial. Other types of surrogates could also be used in principle, however, the large number of parameters used pose significant challenges to traditional methods such as the Gaussian processes used in Jantre et al. (2024). Recently developed machine-learning surrogates (Jouvet et al., 2021; Brinkerhoff et al., 2021; He et al., 2023) could be competitive alternatives to the low fidelity models considered in this work.

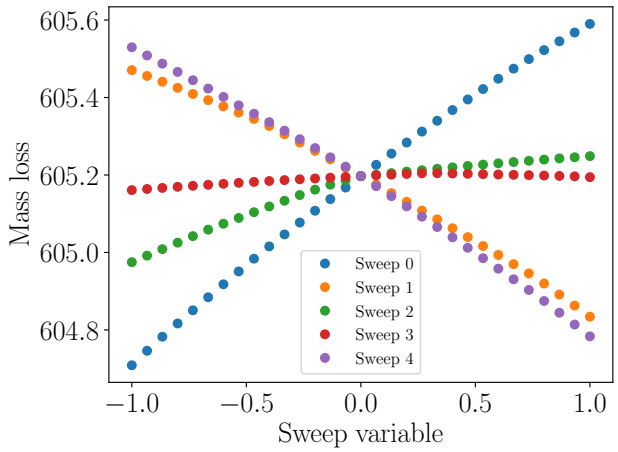


Figure 1: One-dimensional sweeps through the 11,536 dimensional parameter space used to represent Basal friction. Each sweep is along a random direction through the parameter space that pass through the origin (which corresponds to using the mean friction field). The extremes of the sweep correspond to $\pm\sigma$, where σ is the standard deviation of the posterior along the sweeps.

We did not include Figure 1 in the revised manuscript because we believe it distracts from the main message of the manuscript, that is that MFSE can substantially reduce the computational cost of computing statistics of prediction uncertainty. However, we did add the following statements and plots.

The left panel of Figure 13 (Figure 2 in this response document) plots the time evolution of mass loss predicted by the three models selected by our final ACV estimator. The right panel plots the distribution of mass loss at the final year, 2100, computed using the $\text{SSA}_{1.5km,365days}$ model. The bias of the $\text{SSA}_{1.5km,365days}$ is clear in both plots, for example, in the right panel the mean of the blue distribution is not close to the mean computed by the ACV estimator. However, we must emphasize that, by construction, the ACV estimate of the mean mass loss, and its variance, is unbiased with respect to the highest-fidelity model $\text{MOLHO}_{1km,9days}$. We also point out that while our Laplace approximation of the posterior is a Gaussian, the push-forward of this distribution through the $\text{SSA}_{1.5km,365days}$ model model is nonlinear. Specifically, the push-forward of a Gaussian through a linear model remains Gaussian; however, in this case, the right tail of the push-forward density is longer than the left tail, indicating that it is not Gaussian. This suggests that the mapping from the basal friction parameters to the quantity of interest is nonlinear. We were unable to compute reasonable push-forward densities with the simulations obtained from the other two models used to construct the ACV estimator due to an insufficient number of simulations. However, we believe it is reasonable to assume that the parameter-to-QoI map if these models is also non-linear.

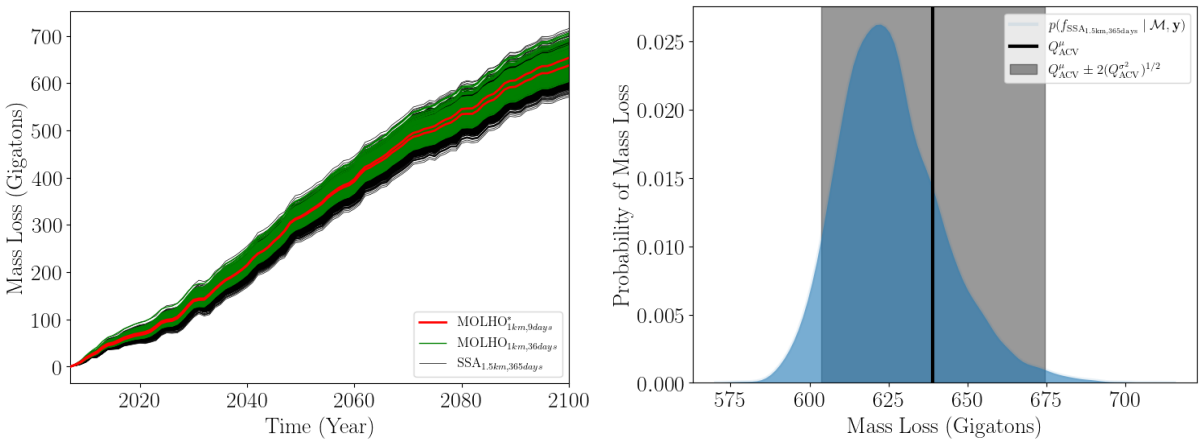


Figure 2: (*Left*) The evolution of mass loss predicted by the three models we used in our final ACV estimator, corresponding to each of the simulations used to construct the estimator. (*Right*) The probability of mass loss computed using the $\text{SSA}_{1.5km,365days}$ model. The black vertical line represents the ACV estimate of the mean, while the gray shaded region represents plus and minus 2 standard deviations, again computed by the ACV estimator.

4.3 Specific comments

157. *"L23-25. This is a good outlay of the different sources of uncertainty. What is missing is a definitive statement that the only type of uncertainty being quantified in this paper is parametric uncertainty."*

    We now state explicitly that we are quantified parametric uncertainty in the introduction. We also now point out that model form error and model discretization error impact the bias of our MFSE estimates of mean and variance in section 4. See our response to comment 63 for details.

158. *"L26-27. "but the impact of discretization errors has not been explicitly considered with other sources of uncertainty". And it has not in this study either, right? As I understand it the MFUQ scheme is solely to estimate parameter uncertainty of the 1km, 9-day MOLHO model – it did not quantify disc. uncertainty despite using different discretizations."*

    You are correct. We apologize for our sloppy statement that may have led reviewers to believe we were quantifying discretization uncertainty. We have edited the introduction accordingly. See our responses to comments 155 and 63.

159. *"L37-41. This is a good place to cite works such as Isaac et al 2015 (and various papers by Noemi Petra e.g. Petra et al 2013), and BR23."*

    See our response to comment 155 that explains our changes to the introduction intended to clarify which studies are low-dimensional and which are high-dimensional.

160. *"L60-61. As noted above, quantifying the impact of a high-dimensional parameterization of basal friction on long-term projections is not novel (cf. BR23 – unless you are distinctly saying that 40 years is not long-term and 80 years is!)"*

    We were not aware of BR23 and now we have read agree that quantifying the impact of a high-dimensional parameterization of basal friction on long-term projections is not novel.

Consequently, we removed that claim and now state in the introduction: We discussed the changes made to the introduction in our response to comment 155.

161. *"L62-64. As noted above, Isaac et al, whose methodology you cite and use, arguably did this."* We do not think that Isaac et al *"quantified the impact of a high-dimensional parameterizations of basal friction on long-term ice-sheet projections"* as stated in the first sentence on line 62. However, we agree we are using the method from Isaac et al for inference. We hope our revised introduction clarifies these facts.

162. *"L66. I'm not sure why you include Isaac 2015 in a list of papers using low-dimensional parameterisations – they used $O(10^6)$ parameters in their basal sliding parameterization."*

The original manuscript stated *"In contrast, previous UQ studies (Nias et al., 2023; Ritz et al., 2015; Schlegel et al., 65 2018; Jantre et al., 2024) only employed low-dimensional parameterizations despite high-dimensional parameterizations being necessary to calibrate ice-sheet models to observational data (Barnes et al., 2021; Isaac et al., 2015; Johnson et al., 2023; Perego et al., 2014)"*. We cited Isaac et al to point out that high-dimensional parameterizations are needed to calibrate ice-sheet models well not saying they used a low-dimensional parameterization. Again we hope the new introduction clarifies this point.

163. *"Fig 2, 3, 5, 6, 7, and 13: you need to show the coordinate axes in all visualisations of the model domain – and there should be one figure showing the placement of Humboldt in Greenland."*

We regenerated all figures in the revised document to improve readability.

164. *"L181: "covariance" – prior or posterior?"*

We changed it to prior covariance.

165. *"L190-193. I have deep concerns about your parameter choices. Firstly, what is the pointwise variance? Secondly, how did you arrive at this correlation length as suitable – on what basis? I do not see any physical reasoning leading to it. You are saying that the data essentially does not need to constrain variability on a scale smaller than this, which I don't think is an accurate statement. BR23 chose far smaller autocorrelations (3km) using some degree of physical inference, and moreover showed that it was necessary to give reasonable values of posterior uncertainty (see comment on TABLE 1 regarding this assessment), and it is possible that in choosing such large numbers you are making the posterior uncertainty artificially small by choosing an overly-informative prior. This may be why you only needed ¡ 1000 eigenvalues to represent the posterior as shown in the appendix. (see BR23 for details.)"*

Thank you for pointing us to the reference BR23. We believe it provides a computationally tractable method for tuning the hyper-parameters of the prior when automatic differentiation (AD) is available with an ice-sheet code. Unfortunately, our codes do not have AD capabilities for transient simulations. (We do have them for steady state simulations. Indeed we use AD to compute the action of the hessian when constructing the Laplace approximation of the posterior.) Because we did not have AD, we had to tune the hyper-parameters heuristically. We informally varied the hyper-parameters of the prior and used our judgment to pick a correlation length and variance that resulted in a posterior MAP point that was able to match the observations well. We have expanded Remark 5.1 to better discuss how we heuristically choose the parameters: "In this study we used our domain experience to determine the best values of the prior hyper-parameters γ, δ, η reported in Section 2.3 and the likelihood hyper-parameter α reported in Section 3. However, varying these hyper-parameters, would likely

change the estimates of uncertainty in ice-sheet predictions produced by this study. Similar to previous studies (Isaac et al., 2015), we did not investigate these sensitivities extensively. We heuristically chose the prior hyper-parameters so that the prior samples would have a variance and spatial variability that we deemed inline with our experience. Further, we found that reducing α substantially from the value we ultimately used while keeping the prior hyper-parameter fixed prevented the MAP point from capturing the high-frequency content of the basal friction field needed to accurately match the observed surface velocities. Future studies should investigate the sensitivity of mass change to the values of the hyper-parameters more rigorously using an approach such as the one developed by Recinos et al. (2023)."

Despite not using the rigorous tuning procedure in BR23, we believe that our calibration is still close to state-of-the-art and is sufficient to demonstrate the ability of MFSE to reduce the computational cost of computing uncertainty. Moreover, the results in BR23 have further motivated us to consider adding AD capabilities to our models for future studies. However, doing so will require substantial human hours and so will not be feasible for this paper.

166. *"L205. How did you generate mass balance? Did you run a regional climate model that incorporates firn and snow processes? If so, say so. Did you use a parameterization? If so, state it and the source."*

We added the following to Section 2.4: *"Second, the MIROC5 climate forcing from the CMIP5 for the Representative Concentration Pathway (RCP) 2.6 scenario was used to generate the surface mass balance (difference between ice accumulation and ablation) f_H and drive the ice-sheet evolution from 2007 to 2100."* This surface mass balance was provided by the Ice Sheet Model Intercomparison Project for CMIP6 (ISMIP6), which down-scaled output from Earth system models using the state-of-the-art regional climate model MAR (Nowicki et al., 2020).

167. *"L231. On what basis do you assume they are uncorrelated? The fact that the products are not posted with spatial correlations of error is not a reason – this is simply too difficult for them to calculate. Please highlight this, and state what the consequences of such an assumption could be for estimating posterior uncertainty."*

We have added the following statement.

In this study we assumed that the observational data are independent, as also assumed in (Recinos et al., 2023) Moreover, we also assumed our Gaussian error model to be exact. However, neither of these assumptions are likely to perfect in reality. Consequently, our results must be viewed with some caution. For example, Koziol et al. (2021) showed that, for an idealized problem, ignoring spatial correlation in the observational noise can lead to uncertainty being underestimated.

168. *"Section 4: in general I think this section should be read over very carefully to look for typos and variables introduced without definition. Ill mention several below but these sections (the ones that I read closely) seem to have been written hastily."*

We have corrected mistakes pointed out by you and the other reviewers as well as some additional ones. We will also spend considerable effort to improve section 4 in the revised manuscript.

169. *"L263, mean Q^μ: mean of what?? And what is Q? and are these "true" statistics or estimators since they have no subscript?"*

We have made extensive edits to this section, including clarifying what Q^μ represents.

170. *"L265. Try to be consistent with tense throughout, and definitely within a sentence: "The second step simulates the model at each realization ... and computed the mass change.."*"

We have corrected tense here and throughout the document. See comment 66.

171. *"L271: "Any MC estimator Q" – do you mean Q_α^μ or $Q_\alpha^{\sigma^2}$, or both or neither?"*

See response 40.

172. *"Eq 11 – can you show how this is derived? At first glance it looked similar to the identity $E[(X - E[X])^2] = E[X^2] - (E[X])^2$ but I could not derive it using similar reasoning."* The right most expression is correct. However the middle expression had a typo which has been corrected. You can find a derivation of the final expression for bias in one of our software tutorials. `https://sandialabs.github.io/pyapprox/auto_tutorials/multi_fidelity/plot_monte_carlo.html`. We did not include this proof in the paper.

173. *"L279 did you mean MSE (II), rather than MSE (11)?"*

We mean Eq. 11. All in text equation references in the paper have been changed from (EQNO) to Eq. (EQNO).

174. *"L279: I don't believe that all of these sources of uncertainty go into the bias term. My interpretation is that, for the purpose of your MFUQ, you are given a density of q arising from the Isaac methodology. You then have a deterministic function fa(X) which is given by your high fidelity model and its discretization, and is therefore deterministic. You are seeking properties of the probability distribution induced by fa and the only actual uncertainty is how fast the MC converges. Model uncertainty and discretization uncertainty, while very real, are not accounted for in such a calculation."*

Line 279 stated *"The bias term of the MSE (11) is caused by using a numerical model, with inadequacy and discretization errors, to compute the mass change."*

We hope that our response to comment 63 answers this question.

175. *"L280 what does MSE (10) mean?"*

We now say Constructing a SFMC estimator with a small MSE, Eq. (10), ...

176. *"L280 ensures, for any set of model input samples,"*

Fixed.

177. *"First eq in 4.2.1 (not numbered) – is the 2nd term in brackets not divided by N1?"* You are correct. See response 15.

178. *"L316 – QoI not defined previously."*

We now define QoI as quantities of interest the first time it is introduced.

179. *"L324 – for the union of these sets to be null, both need to be null. Should it be an intersection symbol?"*

We fixed this mistake.

180. *"Eq 18. You seem to be estimating these statistics using straighforward (Naïve) MC. Why is this OK given the whole thrust of your study is that MC is too expensive to apply to the statistics of the ice model?"*

You raised an important aspect of MF UQ. All current theory assumes that quantities such as (18) in the original manuscript are known exactly. However, in practice they must be estimated using estimates such as (18), typically with a small number of so called pilot samples. An important contribution of this paper is to show that estimating these quantities introduces an error that can be non-trivial. We also provide a strategy for estimating the impact of this pilot error. See figures 10a, 11, a, b and c in the initial submission. In the initial submission we stated the following in the second paragraph after equation (18).

"Unfortunately, using a finite P introduces errors into (16) and (17), which in turn induces error in the ACV estimator covariance. This error can be decreased by using a large P but this would require additional evaluations of expensive numerical models, which we were trying to avoid. Consequently, in this study we investigated the sensitivity of the number of pilot samples on the accuracy of ACV MC estimators."

181. *"L418-422. State # of elements In models"*

We now state:

The number of elements associated with the four meshes with characteristic element sizes 1km, 1.5km, 2km and 3km, were 2611, 9238, 13744, 22334, respectively. The number of nodes for the same four meshes were 1422, 4846, 7154, 11536.

182. *"L424 in the 1st para of 2.4 you state you use FEniCS. MALI is a C++ model with Fortran libraries and not, to my knowledge, written with fenics. Which model(s) did you use???"*

We have added the following to the paper to Section 5.1:

Lastly, note that we used a different model, to the 13 described above, for the Bayesian calibration of the basal friction parameters. Specifically, we used the C++ code `MALI` (Hoffman et al., 2018), which can solve the Blatter-Pattyn equations (Pattyn, 2003; Dukowicz et al., 2010) and compute the action of the Hessian on a vector. `MALI` efficiently computed these Hessian-vector products, needed to compute our Laplace approximation of the posterior in Eq. (14), by solving the adjoint equations for the steady state Blatter-Pattyn equations. However, SSA equations (Section 2.1.3) are not currently implemented in `MALI` and the MOLHO (Section 2.1.2) equations have only recently been implemented (after the simulations for this work were perfomed). Consequently, we used `FEniCS` (Alnæs et al., 2015) to implement both MOLHO and SSA to ensure that the relative computational timings of these models would be consistent. Solving the Blatter-Pattyn model using the C++-based `MALI` code and solving MOLHO and SSA using the python based `FEniCS`, would have corrupted the MFSE results. Moreover, implementing SSA in `MALI` would be time consuming because it is currently only used to solve 3D models and not 2D models, such as SSA. Indeed, a partial motivation for this study was to to determine the utility of implementing the SSA equations in `MALI`.

183. *"Fig 10 – I might be misunderstanding the methods but shouldn't there be units??"*

Figure 10a plots a dimensionless quantity it is the ratio (variance of the MC estimators) of two quantities with the same units. We added Eq (30) to clarify the quantities plotted.

184. *"Table 1. This value is presented without validation. It is possible to do a "sanity check". BR23 use 2 essentially independent measurements of velocity (ITS_LIVE and MEaSUREs) to invert for parameters and simulate mass loss. If the difference seen is of almost negligible probability under the calculated posterior for mass loss - then there must be an issue with the calculated posterior. You are capable of doing this as well ..."*

We do not have access to two different observational data sets for the Humboldt Glacier for the year 2007. We used the best available data set (MEaSUREs) for our Bayesian calibration starting in 2007. Yet, while ITS_LIVE velocity data exists for this year, its coverage at Humboldt is limited, so we could not perform an inversion using ITS_LIVE data alone.

Additionally, while such a sanity check is indeed valuable, computing exactly the correct posterior is not the focus of this paper. Rather the goal of this paper is to demonstrate the utility of using MFSE to quantify uncertainty in the predictions of ice sheet models using a problem setup (data, calibration etc.) that is close to what is used in practice by the best papers in the literature, e.g. in BR23. Changing the prior would likely change the exact values of mass change we reported, but it would likely not substantially change the variance reduction we observed. The variance reduction of a statistic is the ratio of the single-fidelity MC estimator variance divided by the MFSE estimator variance. Thus, when estimating the mean mass change, the exact value of the statistic cancels. We have expanded the discussion of the limitations of our study in the summary and conclusions sections.

185. *"L550. "the SSA model was not.." can you provide an example or evidence of this?"* The original manuscript stated: *"While the highest-fidelity model MOLHO was capable of capturing ice-sheet dynamics that the SSA model was not, that is vertical changes in the horizontal velocities,"* Moreover, Figure 13 in the original manuscript shows that these two model produce different estimates of thickness at the final time. A reference to figure 13 has been placed in the discussion.

186. *"L565. Im confused – I thought that the MFUQ was needed as you are sampling from a distribution of 600 dimensions (the number of Eigenvals retained in the Hessian based UQ). If you have only 10 dimensions can you not use standard (naïve) MC?"*

The original submission stated. *"Our study used a high-dimensional representation of the basal friction field that is capable of capturing high-frequency modes, however it has been common in previous studies to use lower-dimensional parameterizations. Consequently, we investigated the impact of using a low-frequency/lower-dimensional representation of the friction field on the efficiency of ACV estimators using ice-sheet models. Specifically, we estimated the mean and variance of the mass change using a 10 dimensional Karhunen Loeve expansion (KLE) to represent the posterior uncertainty of the basal friction field (complete details are presented in B). "*

Note, while we retained 1125 modes from the prior-preconditioned hessian, computing uncertainty in the QoI still required sampling the 11,536 variables used to parameterize the friction field. Only the KLE study required sampling 10 variables. Moreover, the dimensionality of the parameter space does not explicitly effect the MSE of a MC estimator, see equations (11) and (12) in the original manuscript. We explored the use of MFMC because the number of variables parameterizing the friction field was 11,536 (reported in the introduction of the original submission) prevented us from using surrogate methods. We also do not have the capability to compute gradients of mass loss using adjoint methods such as done in BR23. Our investigation of the impact of using a 10 term KLE was to show that one must avoid the temptation to use a lower-dimensional parameterization of the friction field, to enable the use of surrogates or increase the performance of MFSE, as doing so severely underestimates uncertainty. BR23 shows this very well and we have added a citation to that paper. Specifically, we now state in the discussion

(Recinos et al., 2023) also demonstrated that lower-dimensional parameterizations of uncer-

tainty cause uncertainty to be estimated.

187. *"L567: Appendix B"*.

Fixed.

An evaluation of multi-fidelity methods for quantifying uncertainty in projections of ice-sheet mass-change

John D. Jakeman[1], Mauro Perego[2], D. Thomas Seidl[2], Tucker A. Hartland[3], Trevor R. Hillebrand[4], Matthew J. Hoffman[4], and Stephen F. Price[4]

[1]Optimization and Uncertainty Quantification, Sandia National Laboratories, Albuquerque, NM, 87123
[2]Scientific Machine Learning, Sandia National Laboratories, Albuquerque, NM, 87123
[3]Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA, 94550
[4]Fluid Dynamics and Solid Mechanics Group, Los Alamos National Laboratory, Los Alamos, NM, 87544

Correspondence: J. D. Jakeman (jdjakem@sandia.gov)

Abstract. This study investigated the computational benefits of using multi-fidelity statistical estimation (MFSE) algorithms to quantify uncertainty in the mass change of Humboldt Glacier, Greenland, between 2007 and 2100 using a single climate change scenario. The goal of this study was to determine whether MFSE can use multiple models of varying cost and accuracy to reduce the computational cost of estimating the mean and variance of the projected mass change of a glacier. The problem size and complexity were chosen to reflect the challenges posed by future continental scale studies while still facilitating a computationally feasible investigation of MFSE methods. When quantifying uncertainty introduced by a high-dimensional parameterization of basal friction field, MFSE was able to reduce the mean-squared error in the estimates of the statistics by well over an order of magnitude when compared to a single-fidelity approach that only used the highest-fidelity model. This significant reduction in computational cost was achieved despite the low-fidelity models used being incapable of capturing the local features of the ice flow fields predicted by the high-fidelity model. The MFSE algorithms were able to effectively leverage the high correlation between each model's predictions of mass change, which all responded similarly to perturbations in the model inputs. Consequently, our results suggest that MFSE could be highly useful for reducing the cost of computing continental scale probabilistic projections of sea-level rise due to ice-sheet mass change.

1 Introduction

The most recent Intergovernmental Panel on Climate Change (IPCC) report predicts that the melting of ice sheets will contribute significantly to future rises in sea level (Masson-Delmotte et al., 2021), but the amount of sea-level rise is subject to a large degree of uncertainty. For example, estimates of the sea-level rise in 2100, caused by melting of the Greenland Ice Sheet, range from 0.01 m to 0.18 m. Moreover, projections of the Antarctic Ice Sheet's contribution to sea-level rise are subject to even larger uncertainty (Bakker et al., 2017; Masson-Delmotte et al., 2021; Edwards et al., 2019). Consequently, there is a strong need to accompany recent improvements in the numerical modeling of ice-sheet dynamics with rigorous methods that quantify uncertainty in model predictions.

Accurately quantifying uncertainty in ice-sheet predictions requires estimating the impacts of all sources of model variability. Prediction uncertainty is caused by three main factors: 1) the inadequacy of the governing equations used by the model to approximate reality; 2) the errors introduced by the numerical discretization used to solve the governing equations; and 3) the uncertainty in model parameters used to parameterize future climate forcing and the current condition of the ice sheet, among others. Several studies have demonstrated that model discretization significantly affects model predictions (Cornford et al., 2013; Durand et al., 2009), but the impact of discretization errors has not been explicitly considered with other sources of uncertainty. In addition, while the comparison of model outputs has been used to estimate uncertainty arising from model inadequacy (Goelzer et al., 2018), such studies are not guaranteed to estimate the true model inadequacy (Knutti et al., 2010). Consequently, several recent efforts have focused solely on quantifying parametric uncertainty (Nias et al., 2023; Edwards et al., 2021; Ritz et al., 2015; Schlegel et al., 2018; Recinos et al., 2023), as we do in this study.

Parametric uncertainty is often estimated using Monte Carlo (MC) statistical estimation methods, which compute statistics or construct probability densities using a large number of model simulations evaluated at different random realizations of the uncertain model parameters. However, the substantial computational cost of evaluating ice-sheet models limits the number of model simulations that can be run, and thus the precision of uncertainty estimates. For example, when estimating the mean of a model with MC, the mean squared error (MSE) in the estimated value only decreases linearly as the number of model simulations increases. Therefore, recent UQ efforts constructed emulators (also known as surrogates) of the numerical model from a limited amount of simulation data, and then sampled the surrogate to quantify uncertainty (Berdahl et al., 2021; Bulthuis et al., 2019; Edwards et al., 2019; Jantre et al., 2024). While surrogates can improve the computational tractability of UQ when uncertainty is parameterized by a small number of parameters, their application becomes impractical when there are more than 10-20 variables. This limitation arises because the amount of simulation data required to build these surrogates grows exponentially with the number of parameters (Jakeman, 2023). Consequently, there is a need for methods capable of quantifying uncertainty in ice-sheet models parameterized by a large number of uncertain parameters, such as a those used to characterize a spatially varying basal friction field.

Most recent studies have focused on estimating uncertainty in the predictions of ice-sheet model with small numbers of parameters, e.g. (Nias et al., 2023; Ritz et al., 2015; Schlegel et al., 2018; Jantre et al., 2024), despite large numbers of parameters being necessary to calibrate ice sheet model to observations (Barnes et al., 2021; Johnson et al., 2023; Perego et al., 2014). However, recently Recinos et al. (2023) used the adjoint sensitivity method to construct a linear approximation of the map from a high-dimensional parameterization of uncertain basal friction coefficient and ice stiffness, to quantities of interest (QoI) – specifically the loss of ice volume above flotation predicted by a shallow-shelf approximation model at various future times. The linearized map and the Gaussian characterization of the distribution of the parameter uncertainty was then exploited to estimate statistics of the QoI. While this method is very computationally efficient, linearizing the parameter-to-QoI map will introduce errors (bias) into estimates of uncertainty, which will depend on how accurately the linearized parameter-to-QoI map approximates the true map (Koziol et al., 2021). Moreover, the approach requires using adjoints or automatic differentiation to estimate gradients, which many ice-sheet models do not support. Consequently, in this study we focused on Multi-fidelity statistical estimation (MFSE) methods that do not require gradients.

MFSE methods (Giles, 2015; Peherstorfer et al., 2016; Gorodetsky et al., 2020; Schaden and Ullmann, 2020) utilize models of varying fidelity, that is models with different inadequacy, numerical discretization and computational cost, to efficiently and accurately quantify parameteric uncertainty. Specifically, MFSE methods produce unbiased statistics of a trusted highest-fidelity model by combining a small number of simulations of that model with larger amounts of data from multiple lower-cost models. Note that while low-fidelity models with different discretization and inadequacy error are used, MFSE does not quantify the impact of these two types of errors on the high-fidelity statistics. Furthermore, provided the low-fidelity models are highly correlated with the high-fidelity model and are substantially cheaper to simulate, the mean squared error (MSE) of the MFSE statistic will often be an order of magnitude smaller than the estimate obtained using solely high-fidelity evaluations, for a fixed computational budget. However, such gains have yet to be realized when quantifying uncertainty in ice-sheet models.

This study investigated the efficacy of using MFSE methods to reduce the computational cost needed to estimate statistics summarizing the uncertainty in predictions of sea-level change obtained using ice-sheet models parameterized by large numbers of parameters. To facilitate a computationally feasible investigation, we focused on reducing the computational cost of estimating the mean and variance of mass change in the Humboldt Glacier in northern Greenland. This mass change was driven by uncertainty in the spatially varying basal friction between the ice sheet and land mass, under a single climate change scenario between 2007 and 2100. Specifically, letting f denote the mass change at 2100 computed by a mono-layer higher-order (MOLHO) (Dias dos Santos et al., 2022) model $\mathcal{M}$, $\boldsymbol{\theta}$ denote the parameters of the model characterizing the Basal friction field, and $\boldsymbol{y}$ denote the observational data, we estimated the mean and variance of the distribution $p(f \mid \mathcal{M}, \boldsymbol{y}) = p(f \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathcal{M}, \boldsymbol{y})$ in two steps. First, using a piecewise linear discretization of a log-normal basal friction field, we used Bayesian inference to calibrate the resulting 11,536 dimensional uncertain variable to match available observations of glacier surface velocity. Specifically, we constructed a low-rank Gaussian approximation (Isaac et al., 2015; Recinos et al., 2023; Barnes et al., 2021; Johnson et al., 2023; Perego et al., 2014) of the Bayesian posterior distribution of the model parameters $p(\boldsymbol{\theta} \mid \mathcal{M}, \boldsymbol{y})$ using a Blatter-Pattyn model (Hoffman et al., 2018). Second, we estimated the mean and mass of glacier mass change using 13 different model fidelities (including the highest-fidelity model), based on different numerical discretizations of the MOLHO and shallow-shelf (SSA) physics approximations (Morland and Johnson, 1980; Weis et al., 1999).

Our study makes two novel contributions to previously published glaciology literature. First, it represents the first application of MFSE methods to quantify the impact of high-dimensional parameter uncertainty on transient projections of ice-sheet models defined a realistic physical domains. Our results demonstrate that MFSE can reduce the serial computational time required for a precise UQ study of ice-sheet contribution to sea-level rise from years to a month. Note, Gruber et al. (2022) previously applied MFSE to a ice-sheet model; however, their study was highly simplified, as it only quantified uncertainty arising from two uncertain parameters of an ice-sheet model defined on a simple geometric domain. Second, our paper provides a comprehensive discussion of the practical issues that arise when using MFSE, which are often overlooked in MFSE literature.

This paper is organized as follows. First, Section 2 details the different ice-sheet models considered by this study and the parameterization of uncertainty employed. Second, Section 3 presents the calibration of the ice sheet model and how the posterior samples were generated. Third, Section 4 presents the MFSE methods that were used to quantify uncertainty. Fourth,

Section 5 presents the numerical results of the study and Section 6 presents our findings. Finally, conclusions are drawn in Section 7.

2 Methods

This section presents the model formulations (Section 2.1) and the numerical discretization of these models (Section 2.2) we used to model ice-sheet evolution, as well as the sources of model uncertainty we considered (Section 2.3) when quantifying uncertainty in the mass change from Humboldt Glacier between 2007 and 2100.

2.1 Model Formulations

Ice-sheets behave as a shear thinning fluid and can be modeled with the nonlinear Stokes equation (Cuffey and Paterson, 2010). This section details the Stokes equations and two computationally less expensive simplifications, MOLHO (Dias dos Santos et al., 2022) and SSA (Morland and Johnson, 1980; Weis et al., 1999) which were used to quantify uncertainty in predictions of the contribution of Humboldt glacier to sea-level rise.

Let x and y denote the horizontal coordinates and z the vertical coordinate, chosen such that the sea level, assumed to remain constant during the period of interest, corresponds to $z = 0$. We approximated the ice domain at time t as a vertically extruded domain Ω defined as

$$\Omega(t) := \{(x, y, z) \text{ s.t. } (x, y) \in \Sigma, \text{ and } l(x, y, t) < z < s(x, y, t)\},$$

where $\Sigma \subset \mathbb{R}^2$ denotes the horizontal extent of the ice, $\Gamma_l(t) := \{(x, y, z) \text{ s.t. } z = l(x, y, t), (x, y) \in \Sigma\}$ denotes the lower surface of the ice at time t, and $\Gamma_s(t) := \{(x, y, z) \text{ s.t. } z = s(x, y, t), (x, y) \in \Sigma\}$ denotes the upper surface of the ice[1].

The Stokes, MOLHO, and SSA models defined the thickness of the ice $H(x, y, t) = s(x, y, t) - l(x, y, t)$ as the difference between the ice-sheet surface $s(x, y, t)$ and the bottom of the ice-sheet $l(x, y, t)$. The bottom of the ice-sheet was allowed to be both grounded to the bed topography $b(x, y)$, such that $l(x, y, t) = b(x, y)$, or floating such that $l(x, y, t) = -\frac{\rho}{\rho_w} H(x, y, t)$, where ρ and ρ_w are the densities of ice and ocean water, respectively. Different boundary conditions were then applied on the grounded portion Γ_g of the ice bottom and on the floating portion Γ_f of the ice bottom, where $\Gamma_g \cap \Gamma_f = \emptyset$ and the ice bottom was given by $\Gamma_g \cup \Gamma_f$. The lateral boundary of Ω was also partitioned into the ice-sheet margin (either terrestrial or marine margin) Γ_m and an internal (artificial) boundary Γ_d marking the interior extent of the Humboldt Glacier that was considered. The relevant domains of the ice-sheet are depicted in Figure 1.

The Stokes equations model the horizontal ice velocities $(u(x, y, z, t), v(x, y, z, t))$, vertical ice velocity $w(x, y, z, t)$ and thickness $H(x, y, t)$ of an ice-sheet as a function of the three spatial dimensions (x, y, z). In contrast, the MOLHO model makes simplifications based on the observation that ice-sheets are typically shallow, i.e. their horizontal extent is much greater than their thickness. These simplifications lead to a model that does not explicitly estimate the vertical velocity w and only simulates the horizontal velocities $u(x, y, z, t)$, $v(x, y, z, t)$ as functions of the three spatial coordinates. Contrasting again, the

[1] For simplicity here we assume that Σ does not change in time. This implies that the ice-sheet cannot extend beyond Σ but it can become thicker or thinner (to the point of disappearing in some regions).

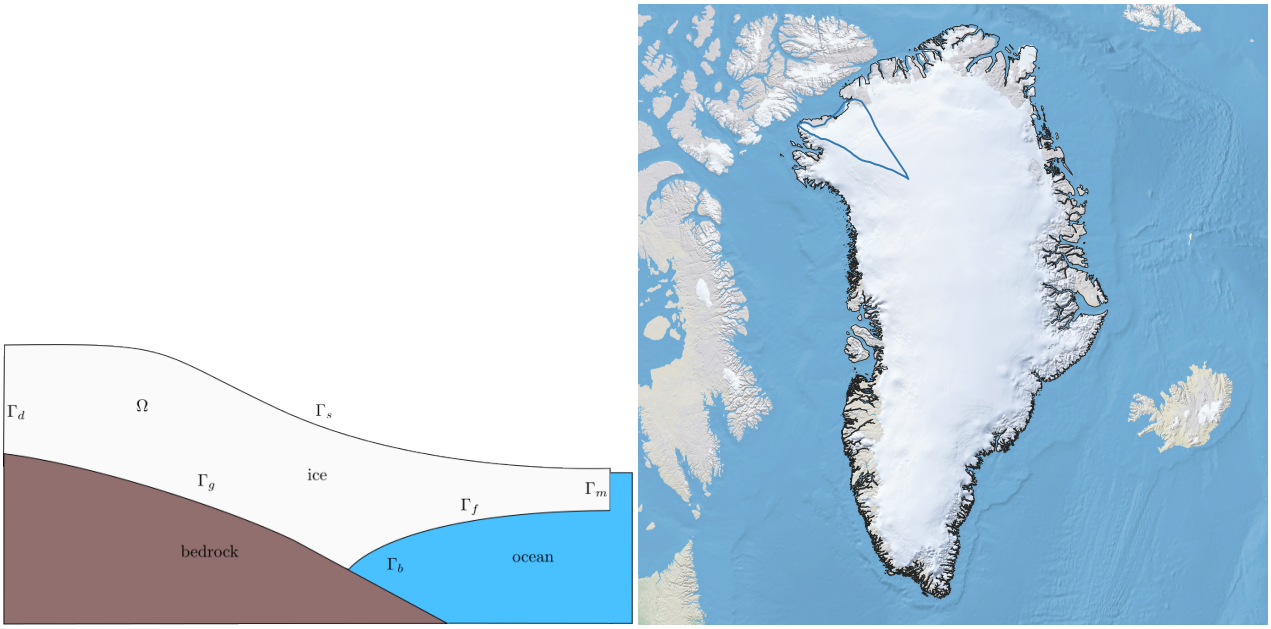


Figure 1. (*Left*) Conceptual model of an ice sheet in the $x-z$ plane. (*Right*) The boundaries (blue lines) of Humboldt Glacier in Greenland.

SSA model makes the additional assumption that the horizontal components of velocity do not vary with thickness (a reasonable approximation in regions where motion is dominated by basal slip) so that the horizontal velocities $u(x,y,t), v(x,y,t)$ are solved for only as functions of (x,y). In summary, the simpler 2D SSA model is formulated to simulate grounded ice with significant sliding at the bed or ice shelves, while the 3D MOLHO model is designed to capture the evolution of ice sheets over frozen and thawed beds, as well as ice shelves.

The Stokes, MOLHO, and SSA models all evolve ice thickness $H(x,y,t)$ according to

$$\partial_t H + \nabla \cdot (\bar{\mathbf{u}} H) = f_H, \qquad H \geq 0, \tag{1}$$

where $\bar{\mathbf{u}} := \frac{1}{H} \int_l^s \mathbf{u}\, dz$ is the thickness-integrated velocity and f_H is a forcing term that accounts for accumulation (e.g. snow accumulation) and ablation (e.g. melting) at the upper (s) and lower (l) surfaces of the ice sheet. However, each model determines the velocities of the ice sheet differently. The following three subsections detail how each model computes the velocity of the ice sheet.

2.1.1 Stokes model

This section introduces the Stokes model, which while not used in this study due to its impractical computational cost, forms the basis of the other three models used in this study. Specifically, the governing equations of the Stokes model are

$$-\nabla \cdot \sigma = \rho \mathbf{g} \tag{2}$$

$$\nabla \cdot \mathbf{u} = 0. \tag{3}$$

The velocities $\mathbf{u} = (u, v, w)$ are dependent on the pressure p, ρ denotes the density of ice, $\sigma = 2\mu\mathbf{D} - pI$ denotes the stress tensor of the ice and $\mathbf{D}_{ij}(\mathbf{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$ denotes the strain rate tensor of the ice; here we used the shorthand $\mathbf{u} = (u, v, w) = (u_1, u_2, u_3)$. The stress tensor is dependent on the non-linear viscosity of the ice which satisfies

$$\mu = \frac{1}{2} A(T)^{-q} D_e(\mathbf{u})^{q-1}, \tag{4}$$

where A is the ice flow factor that depends on the ice temperature T and $q \leq 1$; in our study we set $q = \frac{1}{3}$, which is a typical choice (Hillebrand et al., 2022). In addition, the effective strain rate $D_e(\mathbf{u})$ satisfies $D_e(\mathbf{u}) = \frac{1}{\sqrt{2}}|\mathbf{D}(\mathbf{u})|$, where $|\cdot|$ denotes the Frobenius norm.

When used to model ice sheets, the Stokes equation must be accompanied by the following boundary conditions:

$$\begin{cases} \sigma\mathbf{n} = 0 & \text{on } \Gamma_s & \text{stress free, atmospheric pressure neglected} \\ \sigma\mathbf{n} = \rho_w g \min(z, 0)\mathbf{n} & \text{on } \Gamma_m & \text{boundary condition at the ice margin} \\ \mathbf{u} = \mathbf{u}_d & \text{on } \Gamma_d & \text{Dirichlet condition at internal boundary (ice-flow divide)} \\ \mathbf{u}\cdot\mathbf{n} = \mathbf{0},\ (\sigma\mathbf{n})_{\parallel} = \beta\mathbf{u}_{\parallel} & \text{on } \Gamma_g & \text{impenetrability + sliding condition} \\ \sigma\mathbf{n} = \rho_w g z \mathbf{n} & \text{on } \Gamma_f & \text{hydrostatic pressure of ocean under ice shelves} \end{cases}$$

Here $\beta(x, y)$ is a linearized sliding (or friction) coefficient and $\mathbf{n}$ the unit outward-pointing normal to the boundary and the subscript $\parallel$ denotes the component tangential to the bed. The boundary condition at the margin includes an ocean back-pressure term when the margin is partially submerged ($z < 0$). For a terrestrial margin, $z > 0$, the boundary condition becomes a stress-free condition.

2.1.2 Mono-layer higher-order (MOLHO)

The MOLHO model (Dias dos Santos et al., 2022) is based on the Blatter-Pattyn approximation (Pattyn, 2003; Dukowicz et al., 2010) which can be derived by neglecting the terms w_x and w_y (the derivatives of w with respect to x and y, respectively) in the strain-rate tensor $\mathbf{D}$ and using the incompressibility condition ($\nabla \cdot \mathbf{u} = 0$) such that w_z can be expressed solely in terms of u_x and v_y and

$$\mathbf{D} = \begin{bmatrix} u_x & \frac{1}{2}(u_y + v_x) & \frac{1}{2}u_z \\ \frac{1}{2}(u_y + v_x) & v_y & \frac{1}{2}u_z \\ \frac{1}{2}u_z & \frac{1}{2}v_z & -(u_x + v_y) \end{bmatrix}. \tag{5}$$

This leads (Jouvet, 2016) to the following elliptic equations for the horizontal velocities (u, v)

$$-\nabla \cdot (2\mu \hat{\mathbf{D}}) = -\rho g \nabla_{xy} s, \tag{6}$$

where $\nabla_{xy} := [\partial_x, \partial_y]^\top$, and

$$\hat{\mathbf{D}} = \begin{bmatrix} 2u_x + v_y & \frac{1}{2}(u_y + v_x) & \frac{1}{2}u_z \\ \frac{1}{2}(u_y + v_x) & u_x + 2v_y & \frac{1}{2}v_z \end{bmatrix}. \tag{7}$$

such that the viscosity μ in Eq. (4) has the effective strain rate

$$D_e = \sqrt{u_x^2 + v_y^2 + u_x v_y + \frac{1}{4}(u_y + v_x)^2 + \frac{1}{4}u_z^2 + \frac{1}{4}v_z^2}.$$

MOLHO is derived from the weak form of the Blatter-Pattyn model Eq. (6), with the ansatz that the velocity can be expressed as

$$\mathbf{u}(x, y, z) = \mathbf{u}_b(x, y)\,\phi_b + \mathbf{u}_v(x, y)\,\phi_v\left(\frac{s - z}{H}\right), \quad \text{with } \phi_b = 1, \text{ and } \phi_v(\zeta) = 1 - \zeta^{\frac{1}{q}+1},$$

where the functions ϕ_b and ϕ_v are also used to define the test functions of the weak formulation of the MOLHO model. This ansatz allows the Blatter-Pattyn model to be simplified into a system of two two-dimensional partial differential equations (PDEs) for $\mathbf{u}_b$ and $\mathbf{u}_v$ — Dias dos Santos et al. (2022) give a detailed derivation — such that the thickness-averaged velocity satisfies $\bar{\mathbf{u}} = \mathbf{u}_b + \frac{(1+q)}{(1+2q)}\,\mathbf{u}_v$, where q is the same coefficient appearing in the viscosity definition Eq. (4).

We used the following boundary conditions when using MOLHO to simulate ice-flow

$$\begin{cases} 2\mu\hat{\mathbf{D}}\,\mathbf{n} = 0 & \text{on } \Gamma_s & \text{stress free, atmospheric pressure neglected} \\ 2\mu\hat{\mathbf{D}}\,\mathbf{n} = \psi\mathbf{n} & \text{on } \Gamma_m & \text{boundary condition at at ice margin} \\ \mathbf{u} = \mathbf{u}_d & \text{on } \Gamma_d & \text{Dirichlet condition at internal boundary (ice-flow divide)} \\ 2\mu\hat{\mathbf{D}}\,\mathbf{n} = \beta\mathbf{u}_{\parallel} & \text{on } \Gamma_g & \text{sliding condition} \\ 2\mu\hat{\mathbf{D}}\,\mathbf{n} = 0 & \text{on } \Gamma_f & \text{free slip under ice shelves.} \end{cases}$$

Additionally, we approximated the term $\psi = \rho g(s - z) + \rho_w\, g\, \min(z, 0)$ by its thickness-averaged value $\bar{\psi} = \frac{1}{2}gH(\rho - r^2\rho_w)$, where $r = \max\left(1 - \frac{s}{H}, 0\right)$ is the the submerged ratio.

2.1.3 Shallow Shelf Approximation (SSA)

The SSA model (Morland and Johnson, 1980) is a simplification of the Blatter-Pattyn model that assumes the ice-velocity is uniform in z, so $\mathbf{u} = \bar{\mathbf{u}}$ and thus $u_z = 0, v_z = 0$. This simplification yields

$$\mathbf{D} = \begin{bmatrix} u_x & \frac{1}{2}(u_y + v_x) & 0 \\ \frac{1}{2}(u_y + v_x) & v_y & 0 \\ 0 & 0 & -(u_x + v_y) \end{bmatrix}, \quad \hat{\mathbf{D}} = \begin{bmatrix} 2u_x + v_y & \frac{1}{2}(u_y + v_x) & 0 \\ \frac{1}{2}(u_y + v_x) & u_x + 2v_y & 0 \end{bmatrix}, \tag{8}$$

and $D_e = \sqrt{u_x^2 + v_y^2 + u_x v_y + \frac{1}{4}(u_y + v_x)^2}$. Consequently, the SSA is a single two-dimensional PDE in Σ

$$-\nabla \cdot \left(2\mu H \hat{\mathbf{D}}(\bar{\mathbf{u}})\right) + \beta \bar{\mathbf{u}} = -\rho g H \nabla_{xy} s, \quad \text{in } \Sigma,$$

where $\bar{\mu} = \frac{1}{2}\bar{A}(T)^{-q} D_e(\bar{\mathbf{u}})^{q-1}$, and $\bar{A}$ is the thickness-averaged flow factor. This study explored the use of SSA with the boundary conditions

$$\begin{cases} 2\mu \hat{\mathbf{D}}(\bar{\mathbf{u}})\,\mathbf{n} = \bar{\psi}\mathbf{n} & \text{on } \Gamma_m \quad \text{boundary condition at ice margin} \\ \bar{\mathbf{u}} = \bar{\mathbf{u}}_d & \text{on } \Gamma_d \quad \text{Dirichlet condition at internal boundary.} \end{cases}$$

With abuse of notation, here Γ_m and Γ_d denotes subsets of the boundary of Σ.

2.2 Numerical discretization

The ability to predict ice-sheet evolution accurately is dictated not only by the governing equations used, but also by the properties of the numerical methods used to solve the governing equations. In this study, we discretized the thickness and the velocity equations of the MOLHO and SSA models using the popular Galerkin-based finite element method with piecewise linear elements, which we implemented in `FEniCS` (Alnæs et al., 2015). The coupled thickness and velocity equations were solved in a semi-implicit fashion using a Backward Euler time discretization for the thickness and lagging the evaluation of the velocity. The thickness equation was stabilized using the streamline upwind method. Additionally, the advection term was integrated by part and the thickness was treated implicitly. The discretized problem was solved using the `PETSc` (Balay et al., 1998) SNES nonlinear solver. Using this time evolution process, we did not observe any numerical instabilities when using the time-step sizes adopted in this study.

Because the thickness H obtained from Eq. (1) is not guaranteed to be positive due to the forcing term f_H and that the discretization used is not positivity preserving, we adopted two different approaches to guarantee the positivity of the thickness computed by our finite element models. The first approach involved updating the thickness value at each node so that it was greater than or equal to a minimum thickness value $H_m = 1$ m. The second approach used an optimization-based approach (Bochev et al., 2020) to preserve the thickness constraint ($H \geq H_m$) and guarantee that the total mass change is always consistent with the forcing term in regions where the ice is present and with the boundary fluxes. The first approach is computationally cheaper than the second, but unlike the second method does not conserve mass.

In addition to mass conservation, the number of finite elements and the time-step size both affect the error in the finite element approximation of the governing equations of the MOLHO and SSA models. In this study we investigated the impact of the number of finite elements, which we also refer to as the spatial mesh resolution, and time-step size, on the precision of statistical estimates of mass-change. Specifically, the MOLHO and SSA models were both used to simulate ice-sheet evolution with four different finite element meshes and four different time-step sizes. More details on the spatial mesh and time-step sizes used are given in Section 5.1. Figure 2 compares the four different finite element meshes used to model the Humboldt Glacier. Due to the differences in the characteristic element size of each mesh, the computational domain of each mesh is different. However, we will show that this did not prevent the use of these meshes in our study.

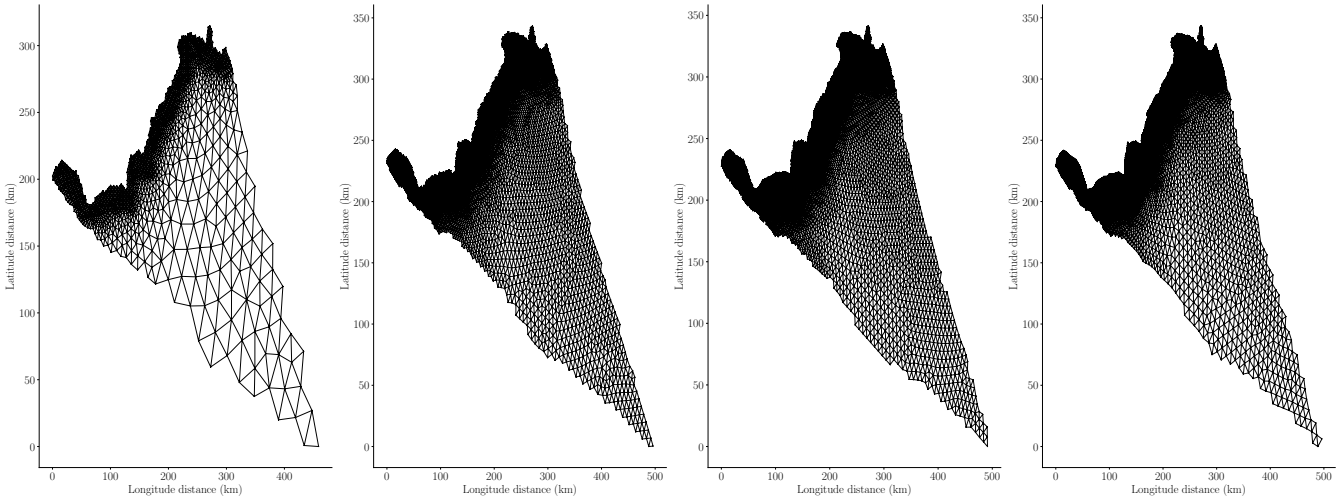


Figure 2. Comparison of the four finite element meshes used to model the Humboldt Glacier with characteristic finite-element sizes of 1km, 1.5km, 2km, and 3km, shown left to right.

2.3 Parameterization of uncertainty

Many factors introduce parametric uncertainty into the predictions of ice-sheet models, including atmospheric forcing, ice rheology, basal friction, ice temperature, calving, and submarine melting. While all these sources of parametric uncertainty may significantly impact predictions of mass change from ice sheets, this study focused on quantifying uncertainty due to unknown basal friction, which is considered one of the largest sources of prediction uncertainty after future environmental forcing (Nias et al., 2018; Joughin et al., 2019; Brondex et al., 2019; Åkesson et al., 2021; Hillebrand et al., 2022; Nias et al., 2023). This singular focus was made to improve our ability to assess whether MFSE is useful for quantifying uncertain in ice-sheet modeling with high-dimensional parameter uncertainty, which most existing UQ methods cannot tractably address. By doing so, we ensured that the conclusions drawn by our study can be plausibly extended to studies considering additional sources of uncertainty. This ensures that the conclusions drawn by our study can be plausibly extended to studies considering additional sources of uncertainty.

The uncertainty in basal friction β, which impacts the boundary conditions of the MOLHO and SSA models, can be parameterized in a number of ways. For example, a lumped approach would assign a single scalar random variable to the whole domain or a semi-distributed approach may use different constants in predefined subdomains, e.g. catchments, of the glacier. In this study, we adopted a fully distributed approach that treated the friction as a log-Gaussian random field that is $\boldsymbol{\theta} = \log(\boldsymbol{\beta}) \in \mathbb{R}^{N_\theta}$, with a Gaussian prior distribution $p(\boldsymbol{\theta}) \sim \mathcal{N}(\mu, \mathcal{C})$ with $\mu = 0$.

Following Isaac et al. (2015), we defined the prior covariance operator $\mathcal{C}$ to be a infinite-dimensional Laplacian squared operator. Specifically, we used a finite-dimensional discretization of the operator $\boldsymbol{\Sigma}_{\text{prior}} \approx \mathcal{C}$ with

$$\boldsymbol{\Sigma}_{\text{prior}}^{-1} = \mathbf{K}\mathbf{M}^{-1}\mathbf{K}, \tag{9}$$

where $\mathbf{K} \in \mathbb{R}^{N_\theta \times N_\theta}$ and $\mathbf{M} \in \mathbb{R}^{N_\theta \times N_\theta}$ are finite elements matrices for the elliptic and mass operators, defined as

$$K_{ij} = \gamma \int_{\Gamma_l} \nabla\phi_i(\boldsymbol{x}) \cdot \nabla\phi_j(\boldsymbol{x}) d\boldsymbol{x} + \delta \int_{\Gamma_l} \phi_i(\boldsymbol{x}) \cdot \phi_j(\boldsymbol{x}) d\boldsymbol{x} + \xi \int_{\partial\Gamma_l} \phi_i(\boldsymbol{x}) \cdot \phi_j(\boldsymbol{x}) d\boldsymbol{x}, \tag{10}$$

$$M_{ij} = \int_{\Gamma_l} \phi_i(\boldsymbol{x}) \cdot \phi_j(\boldsymbol{x}) d\boldsymbol{x}, \tag{11}$$

and ϕ_i are finite element basis functions and $\boldsymbol{x} = (x, y)$. The first term in the definition of $\mathbf{K}$ is the Laplacian operator, the second term is a mass operator representing a source term, and the last term is a boundary mass operator for Robin boundary conditions. The ratio of the coefficients γ and δ determines the correlation length $l = \sqrt{\frac{\gamma}{\delta}}$ of the covariance. In our simulations, we set $\gamma = 2000$ km, $\delta = 2$ km^{-1} and $\xi = 20$, hence $l \approx 31.6$ km. These values were found to balance the smoothness of realizations of the friction field with the ability to capture the fine scale friction features needed to produce an acceptable match between the model prediction of surface velocity and the observed values. Two random samples from the prior distribution of the log-friction are depicted in Figure 3.

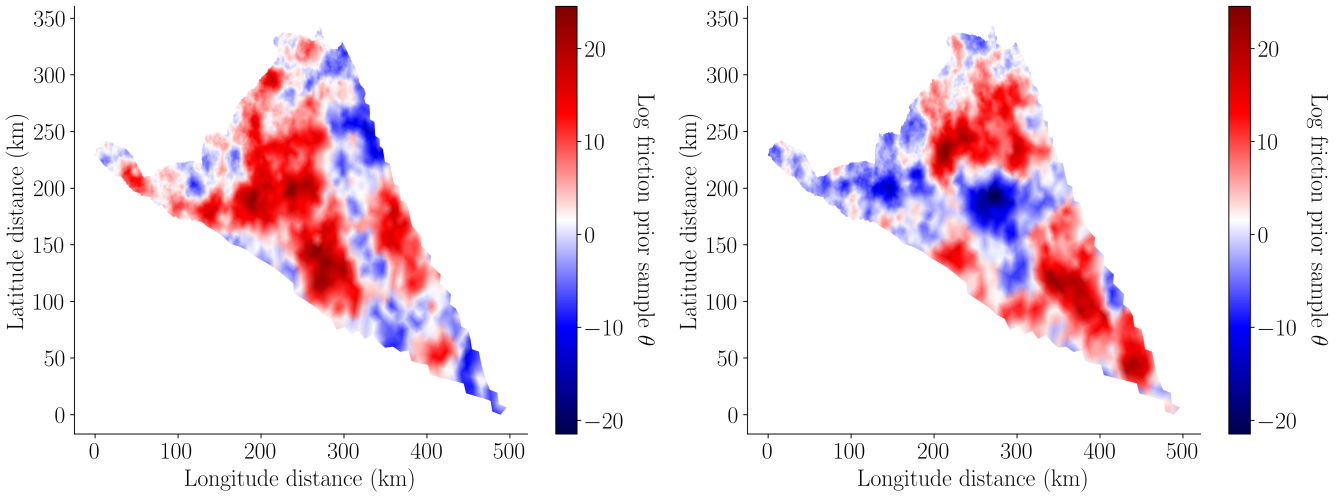


Figure 3. Two random samples from the prior distribution of the log-friction $p(\boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{prior}})$, where $\boldsymbol{\Sigma}_{\text{prior}}$ is defined in Eq. (9).

The parameterization of the prior in Eq. (9) has two main advantages. First, computationally efficient linear algebra can be used to draw samples from the prior distribution. In this study we drew samples from the prior using

$$\boldsymbol{\theta} = \boldsymbol{\mu}_{\text{prior}} + \mathbf{L}\boldsymbol{n}, \quad \boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{N_\theta})$$

with $\boldsymbol{\mu}_{\text{prior}} = \mathbf{0}$, $\mathbf{I}_{N_\theta}$ is the identity matrix with N_θ rows, $\mathbf{L} = \mathbf{K}^{-1}\mathbf{M}^{\frac{1}{2}}$, such that $\boldsymbol{\Sigma}_{\text{prior}} = \mathbf{L}\mathbf{L}^\top$, and we lump the mass matrix $\mathbf{M}$. The second advantage is that this Gaussian prior enables an efficient procedure for computing the posterior distribution of the friction field constrained by the observations, which we present in Section 3.

2.4 Additional model setup

Additional details regarding the model setup are as follows. First, the glacier's bed topography, ice surface elevation, and ice thickness were obtained from observations (refer to Hillebrand et al. (2022) for details) and interpolated onto the finite element mesh. Second, the MIROC5 climate forcing from the CMIP5 for the Representative Concentration Pathway (RCP) 2.6 scenario was used to generate the surface mass balance (difference between ice accumulation and ablation) f_H and drive the ice-sheet evolution from 2007 to 2100. This surface mass balance was provided by the Ice Sheet Model Intercomparison Project for CMIP6 (ISMIP6), which down-scaled output from Earth system models using the state-of-the-art regional climate model MAR (Nowicki et al., 2020). Finally, the ice front was kept fixed such that any ice that moved beyond the calving front is assumed to melt, and any explicit ocean forcing was ignored.

3 Calibration

The goal of this study was to investigate uncertainty in predictions of the future mass change of Humboldt Glacier. However, generating realistic predictions with a model requires calibrating that model to available data. Consequently, in this paper we calibrated the basal friction field of our numerical models to measurements of surface velocity of the ice sheet. We processed Humboldt Glacier geometry data and surface velocity observations for year 2007 as detailed in Hillebrand et al. (2022). The geometry was assumed to be error free and the ice sheet was assumed to be at thermal equilibrium. Thus, we calibrated the friction field by fitting the outputs of a high-resolution steady-state thermo-coupled flow model to the observational data.

Ice-sheet models are typically calibrated using deterministic optimization methods that find the values of the model parameters that lead to the best match between observations and the model prediction of the observations, e.g MacAyeal (1993); Morlighem et al. (2010); Petra et al. (2012); Perego et al. (2014); Goldberg et al. (2015). However, such approaches produce a single optimized parameter value to represent the uncertainty in the model parameters that arises from using a limited amount of noisy observational data.

Bayesian inference uses Bayes' Theorem to quantify the probability of the parameters conditioned on the data $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \in \mathbb{R}$, known as the posterior distribution, as proportional to the conditional probability of observing the data given the parameters $p(\boldsymbol{y} \mid \boldsymbol{\theta}) \in \mathbb{R}$, known as the likelihood, multiplied by the prior probability assigned to the parameters $p(\boldsymbol{\theta}) \in \mathbb{R}$,

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

In this work we assumed that the observational data (surface velocities $\boldsymbol{y} = \boldsymbol{u}_{\text{obs}} \in \mathbb{R}^{2N_{\text{obs}}}$), were corrupted by centered Gaussian noise $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{noise}}) \in \mathbb{R}^{2N_{\text{obs}}}$. Specifically, given a Blatter-Pattyn flow model $\boldsymbol{g}(\boldsymbol{\theta}) \in \mathbb{R}^{2N_{\text{obs}}}$ that maps the logarithm of the basal friction to the computed surface velocity, we assumed $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{\theta}) + \boldsymbol{\eta}$ such that the likelihood function was given by

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = (2\pi |\boldsymbol{\Sigma}_{\text{noise}}|)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))^{\top} \boldsymbol{\Sigma}_{\text{noise}}^{-1} (\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta})) \right).$$

Here, $\boldsymbol{g}(\boldsymbol{\theta})$ denotes the output of the steady-state ice-sheet model at the locations of the observations for a given realization of the model parameters. We were able to calibrate the model using only a steady model without time-stepping because we

assumed that the velocity data were collected over a short period of time over which the ice sheet state is approximately steady. We also assumed that the observations were uncorrelated and set

$$\boldsymbol{\Sigma}_{\text{noise}} = \frac{1}{\alpha}\begin{bmatrix} \mathbf{U}_{\text{err}}\mathbf{M}_s^{-1}\mathbf{U}_{\text{err}} & \\ & \mathbf{U}_{\text{err}}\mathbf{M}_s^{-1}\mathbf{U}_{\text{err}} \end{bmatrix} \in \mathbb{R}^{2N_{\text{obs}}\times 2N_{\text{obs}}}, \tag{12}$$

where $\mathbf{U}_{\text{err}} = \text{Diag}(u_{\text{err}})$ is the diagonal matrix containing the root mean square errors $u_{\text{err}} \in \mathbb{R}^{N_{\text{obs}}}$ of the surface velocity magnitudes, and $M_s \in \mathbb{R}^{N_{\text{obs}}}$ is the mass matrix computed on the upper surface Γ_s and α is a scaling term. We set $\alpha = 8\ \text{km}^{-2}$.

Before continuing, we wish to emphasize two important aspects of the calibration used in this study that mean our results must be viewed with some caution. First, we assumed the observational data to be uncorrelated, as assumed in most ice-sheet inference studies, including (Recinos et al., 2023; Isaac et al., 2015). Moreover, we also assumed our Gaussian error model to be exact. However, neither of these assumptions are likely to be perfect in reality. For example, Koziol et al. (2021) showed that, for an idealized problem, ignoring spatial correlation in the observational noise can lead to uncertainty being underestimated. Second, our optimization of the MAP point was constrained by the coupled velocity flow equations and steady-state enthalpy equation, which is equivalent to implicitly assume that the ice is at thermal equilibrium. Theoretically, this assumption could be avoided if the temperature tendencies were known, but they are not. Alternatively, transient optimization over long time periods, comparable to the temperature time scales, could be used. However, this approach would be computationally expensive and would require including time-varying temperature data (e.g., inferred by ice cores) which are very sparse.

Quantifying uncertainty in mass-change projections conditioned on observational data requires drawing samples from the posterior of $\log(\boldsymbol{\beta})$, evaluating the transient model at each sample and computing estimates of statistics summarizing the prediction uncertainty using those evaluations. Typically, samples are drawn using Markov Chain Monte Carlo (Hoffman and Gelman, 2014), however such methods can be computationally intractable for high-dimensional uncertain variables (Bui-Thanh et al., 2013), such as the variable we used to parameterize basal friction. Consequently, we used the two-step method presented in Bui-Thanh et al. (2013); Isaac et al. (2015) to construct a Laplace approximation of the posterior. Please note that, recently variational inference has been used to infer high-dimensional basal friction (Brinkerhoff, 2022), however we did not use such methods in out study.

First, we performed a PDE-constrained deterministic optimization to compute the maximum a posteriori (MAP) point $\boldsymbol{\theta}_{\text{MAP}}$

$$\boldsymbol{\theta}_{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmin}}\ \frac{1}{2}(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}_{\text{noise}}^{-1}(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{\theta})) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_{\text{prior}})^\top \boldsymbol{\Sigma}_{\text{prior}}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_{\text{prior}}), \tag{13}$$

which maximizes the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$. For linear models and Gaussian priors, the MAP point has close ties with the optimal solution obtained using Tikhonov regularization (Stuart, 2010). Specifically, the first term above minimizes the difference between the model predictions and the observations and the second term penalizes the deviation of the optimal point from the prior mean.

Second, we constructed a low-rank quadratic approximation of the log posterior, centered at the MAP point

$$\log(p(\boldsymbol{\theta} \mid \boldsymbol{y})) \approx C - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{MAP}})^\top \boldsymbol{\Sigma}_{\text{post}}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\text{MAP}}), \tag{14}$$

where

$$\boldsymbol{\Sigma}_{\text{post}}^{-1} = \mathbf{H}_{\text{MAP}} + \boldsymbol{\Sigma}_{\text{prior}}^{-1} \tag{15}$$

and $\mathbf{H}_{\text{MAP}} \in \mathbb{R}^{N_\theta \times N_\theta}$ is the Hessian of $\frac{1}{2}(\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}_{\text{noise}}^{-1}(\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{\theta}))$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{MAP}}$ and C is a constant independent of $\boldsymbol{\theta}$. This resulted in a Gaussian approximation of the posterior $p(\theta \mid \boldsymbol{y}) \approx \mathcal{N}(\boldsymbol{\theta}_{\text{MAP}}, \boldsymbol{\Sigma}_{\text{post}})$, also known as a Laplace approximation of the posterior. Naively computing the posterior covariance using the aforementioned formula for $\boldsymbol{\Sigma}_{\text{post}}$ is computationally intractable. That approach requires solving $2N_\theta$ linearized (adjoint) flow models to computing and inverting the large dense matrix $\mathbf{H}_{\text{MAP}}$, which require O($N_\theta{}^3$) operations. For reference, in this study we use N_θ =11,356 variables to parameterize the basal friction and the adjoints of the flow model had 227,120 unknowns. Consequently, we used a low-rank Laplace approximation which is detailed in Appendix A to efficiently draw random samples from the posterior distribution.

The posterior characterizes the balance between the prior uncertainty in the friction field and the model-observation mismatch, weighted by the observational noise. In the limit of infinite observational data, the posterior distribution will collapse to a single value. However, in practice when using a finite amount of data, the posterior will only change substantially from the prior in directions of the parameter space informed by the available data, which were captured by our low-rank approximation.

4 Uncertainty quantification

This study investigated the efficacy of using MFSE to compute the uncertainty in predictions of future mass change from Humboldt Glacier. We defined mass change to be the difference between the final mass[2] of the glacier at $t = 2100$ and its mass at $t = 2007$. While the mass change is a functional of the ice-sheet thickness H, for simplicity the following discussion simply refers to the mass change as a scalar function of only the model parameters, that is $f_\alpha(\theta) \in \mathbb{R}$, where α indexes the model fidelity that was used to simulate the ice sheet. Previous UQ studies computed statistics summarizing the uncertainty in ice-sheet predictions, such as mean and variance, using *single-fidelity* Monte Carlo (SFMC) quadrature, that is MC quadrature applied to a single physics model with a fixed numerical discretization α, for example Ritz et al. (2015); Schlegel et al. (2018). However, in this study we used MFSE to reduce the computational cost of quantifying uncertainty. Specific details on the MFSE methods investigated are given in the following subsections.

4.1 Single-fidelity Monte Carlo quadrature

SFMC quadrature is a highly versatile procedure that can be used to estimate a wide range of statistics for nearly any function regardless of the number of parameters involved. SFMC can be used used to compute the mean $Q^\mu \in \mathbb{R}$ and variance $Q^{\sigma^2} \in \mathbb{R}$ of the Humboldt glacier mass change predicted by a single model using a three-step procedure. The first step randomly samples N realizations of the the model parameters $\Theta = \{\theta^{(n)}\}_{n=1}^N$ from their posterior distribution. The second step simulates the model at each realization of the random variable (basal friction field) and computes the mass change at the final time $f_\alpha^{(n)} = f_\alpha(\theta^{(n)})$.

[2] In this work, we compute the mass of the glacier considering only the ice above flotation, which is what contributes to sea-level change

The third step approximates the mean and variance using the unbiased estimators

$$\mathbb{E}_\pi\left[f_\alpha\right] \approx Q_\alpha^\mu(\Theta) = N^{-1}\sum_{n=1}^{N} f_\alpha^{(n)} \tag{16}$$

$$\mathbb{V}_\pi\left[f_\alpha\right] \approx Q_\alpha^{\sigma^2}(\Theta) = (N-1)^{-1}\sum_{n=1}^{N}\left(f_\alpha^{(n)} - Q_\alpha^\mu(\Theta)\right)^2, \tag{17}$$

where we use the script π on the exact expectation $\mathbb{E}_\pi\left[f_\alpha\right]$ and variance $\mathbb{V}_\pi\left[f_\alpha\right]$ to make clear these statistics are always computed by sampling from the distribution of $\boldsymbol{\theta}$. In out study, we sampled from the posterior distribution of the basal friction parameters, i.e. $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathcal{M}, \boldsymbol{y})$.

MC estimators converge to the true mean and variance of f_α as the number of samples tends to infinity, but using a finite number of samples N introduces an error into the MC estimator that depends on the sample realizations used to compute the estimators. That is, two different realizations of N parameter samples Θ, and the associated QoI values, will produce two different mean and variance estimates (see Figure 4). Consequently, any MC estimator $Q_\alpha(\Theta)$ of an exact statistic Q, such as $Q_\alpha^\mu(\Theta)$ and $Q_\alpha^{\sigma^2}(\Theta)$, is a random variable.

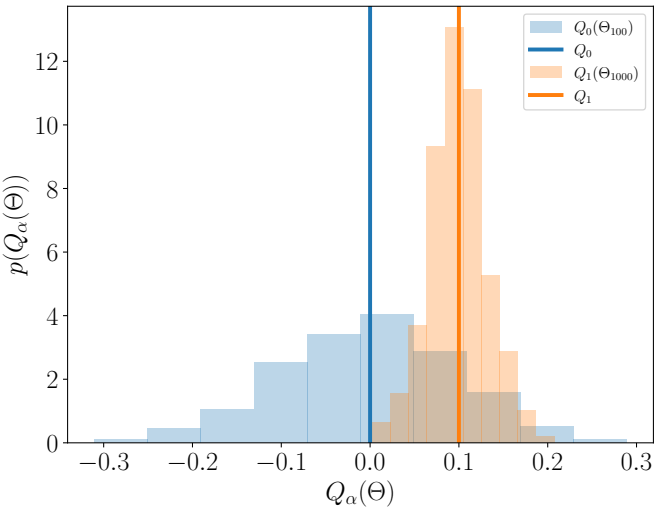


Figure 4. The bias-variance trade-off of MC estimators, stated of the same computational cost is depicted in blue and orange, The blue represents the true mean of a computationally expensive model f_0 and the orange line represents the mean of a model f_1 that is 10-times cheaper but less accurate. The models are only conceptual and not related to the ice-sheet models used in this study and were designed so that evaluating f_0 100 times took the same computational effort as evaluating f_1 1000 times. The histograms are constructed by computing the mean 1000 times using different realizations of the parameters set Θ_N, where N denotes the number of parameter samples in Θ_N. The blue and orange histograms were computed using the 1000 different estimates of the means of f_0 and f_1, respectively.

The mean-squared error (MSE) is typically used to quantify the error in a MC estimate of a statistic and is given by

$$\mathbb{E}_\Theta\left[(Q_\alpha(\Theta)-Q)^2\right]=\mathbb{E}_\Theta\left[(Q_\alpha(\Theta)-\mathbb{E}_\Theta\left[Q_\alpha(\Theta)\right]+\mathbb{E}_\Theta\left[Q_\alpha(\Theta)\right]-Q)^2\right]=\underbrace{\mathbb{V}_\Theta\left[Q_\alpha(\Theta)\right]}_{I}+\underbrace{(\mathbb{E}_\Theta\left[Q_\alpha(\Theta)\right]-Q)^2}_{II}, \tag{18}$$

where $\mathbb{E}_\Theta[\cdot]$ and $\mathbb{V}_\Theta[\cdot]$ respectively denote taking the expectation and the variance over different realizations of the set of parameter realizations Θ. The MSE of an MC estimator, Eq. (18), consists of two terms referred to as the estimator variance (I) and the estimator bias (II). The bias term of the MSE is caused by using a numerical model, with inadequacy and discretization errors, to compute the mass change. More specifically, letting Q_∞ denote the exact value of the statistic of a numerical model with zero discretization error but non-zero model inadequacy error, and Q_0 denote the highest-fidelity computationally tractable model approximation of Q_∞, then the bias can be decomposed into three terms

$$(\mathbb{E}\left[Q_\alpha(\Theta)\right]-Q)=(\mathbb{E}_\Theta\left[Q_\alpha(\Theta)\right]-Q_0+Q_0-Q_\infty+Q_\infty-Q)=(\mathbb{E}_\Theta\left[Q_\alpha(\Theta)\right]-Q_0)+(Q_0-Q_\infty)+(Q_\infty-Q) \tag{19}$$

The first term is caused by using a model f_α with numerical discretization that is inferior to that employed by the highest fidelity model f_0. The second term represents the error in the statistic introduced by the numerical discretization of the highest-fidelity model. The third term quantifies the model inadequacy error caused by the numerical model being an approximation of reality. The variance of an MC estimator comes from using a finite number of samples and decreases as the number of samples increases. For example, the variances of the estimators of mean and variance are, respectively,

$$\mathbb{V}_\Theta\left[Q^\mu_\alpha(\Theta)\right]=\frac{1}{N}\mathbb{V}_\pi\left[f_\alpha\right],\quad \mathbb{V}_\Theta\left[Q^{\sigma^2}_\alpha(\Theta)\right]=\frac{1}{N}\left(\frac{2}{(N-1)}\mathbb{V}_\pi\left[f_\alpha\right]^2+\mathbb{V}_\pi\left[(f_\alpha-\mathbb{E}_\pi\left[f_\alpha\right])^2\right]\right), \tag{20}$$

where the variances involving f_α are exact statistics of the model, which are typically unknown. A detailed derivation of the expression for $\mathbb{V}_\pi\left[Q^{\sigma^2}_\alpha(\Theta)\right]$ can be found in Dixon et al. (2023).

Constructing a SFMC estimator of statistics, such as the mean Eq. (16) or variance Eq. (17), with a small MSE ensures that the value of the estimator will be likely close to the true value, for any set of model parameters samples. However, when using numerical models approximating a physical system, constructing an unbiased estimator of Q is not possible. All models are approximations of reality and thus the model inadequacy contribution $Q_\infty - Q$ to the bias decomposition in Eq. (19) can never be driven to zero. Additionally, it is impractical to quantify the discretization error $Q_\infty - Q_0$. Consequently, SFUQ methods focus on producing unbiased estimators of Q_0, such that $\mathbb{E}_\Theta\left[Q_\alpha(\Theta)\right]=Q_0$.

Unfortunately, even when ignoring inadequacy and discretization errors, constructing a SFMC estimator with a small MSE, Eq. (18), using a computationally expensive high-fidelity model is computationally demanding. The cost is high because the variance term of the MSE of an estimator, Eq. (18), only decreases linearly with the number of samples. In contrast, N can be significantly increased if a cheaper, lower-fidelity model is used, but the corresponding decrease in the estimator variance will be offset by an increase in its bias. Consequently, the bias and variance of any estimator (see Figure 4) should be balanced, but most SFMC analyses do not consider this trade-off explicitly when choosing the fidelity of the model used. In the following section we detail how to use MFSE to substantially improve the precision of estimated statistics for a fixed computational cost.

4.2 Two-model multi-fidelity uncertainty quantification

MFSE leverages the correlation between models of varying cost and fidelity to reduce the computational cost of constructing MC estimators with a desired MSE. While various multi-fidelity estimators have been developed, this study used approximate control variate (ACV) estimators (implemented in PyApprox (Jakeman, 2023)), which include most existing estimators, including Multi-level Monte Carlo (MLMC) (Giles, 2015) and Multi-fidelity Monte Carlo (MFMC) (Peherstorfer et al., 2016), as special cases. [3] In this section, we describe how to construct an ACV estimate of the mean of a model using two models. We then introduce the ACV procedure we used to compute the mean and variance of our our highest-fidelity ice-sheet model using an ensemble of 13 models.

Using only high-fidelity model simulations to estimate a statistic with single-fidelity MC produces an unbiased estimator of Q_0. However, when the computational cost of running a high-fidelity model limits the number of model simulations that can be used, the variance and thus the MSE, of the MC estimator will be large. Fortunately, the MSE error of the estimator can be reduced by correcting the high-fidelity estimator with statistics computed using lower-fidelity models. For example, given a high-fidelity model $f_0(\theta)$ and a single low-fidelity model $f_1(\theta)$, an MFMC ACV estimator approximates the mean of the high-fidelity model using

$$Q^{\mu}_{\mathrm{ACV}}(\Theta_0,\Theta_1) = N_0^{-1}\sum_{n=1}^{N_0} f_0(\theta_0^{(n)}) + \eta\left(N_0^{-1}\sum_{n=1}^{N_0} f_1(\theta_0^{(n)}) - N_1^{-1}\sum_{j=1}^{N_1} f_1(\theta_1^{(j)})\right) \tag{21}$$

$$= Q^{\mu}_0(\Theta_0) + \eta(Q^{\mu}_1(\Theta_0) - Q^{\mu}_1(\Theta_1)) \approx \mathbb{E}_{\Theta}\left[f_0\right]. \tag{22}$$

The two-model ACV estimator in Eq.(21) uses a weighted combination of a high-fidelity MC estimator and two low-fidelity estimators. The high-fidelity model evaluations are used to ensure the ACV estimator is unbiased, i.e. $\mathbb{E}_{\Theta}\left[Q^{\mu}_{\mathrm{ACV}}(\Theta_0,\Theta_1)\right] = \mathbb{E}_{\pi}\left[f_0\right]$, while the low-fidelity evaluations are used to reduce the variance of the estimator. The estimator of the low-fidelity mean $Q^{\mu}_1(\Theta_0)$ is referred to as a control variate because it is a random variable, which is correlated with the random estimator $Q^{\mu}_0(\Theta_0)$, and can be used to control the variance of that high-fidelity estimator. The term $Q^{\mu}_1(\Theta_1) \approx Q^{\mu}_1$ is an approximation of the true low-fidelity statistic Q_1 that is used to ensure that the ACV estimator is unbiased, i.e. $\mathbb{E}_{\Theta}\left[Q^{\mu}_{\mathrm{ACV}}(\Theta_0,\Theta_1)\right] = \mathbb{E}_{\Theta}\left[Q^{\mu}_0(\Theta)\right] + \eta\left(\mathbb{E}_{\Theta}\left[Q^{\mu}_1(\Theta_0)\right] - \mathbb{E}_{\Theta}\left[Q^{\mu}_1(\Theta_1)\right]\right) = Q^{\mu}_0 + \eta\left(Q^{\mu}_1 - Q^{\mu}_1\right) = Q^{\mu}_0$. The weight η can either be fixed, e.g. MLMC sets $\eta = -1$, or optimized to minimize the MSE of the estimator. However, an ACV estimator will always be unbiased, with respect to Q_0, regardless of the value of η, because the expected values of the second and third terms will always cancel.

Computing the ACV estimate of the high-fidelity mean in Eq. (21) requires two different sets of model evaluations. These evaluations must be obtained by first drawing two sets of samples $\Theta_0 = \{\theta_0^{(n)}\}_{n=1}^{N_0}, \Theta_1 = \{\theta_1^{(n)}\}_{n=1}^{N_1}$ from the distribution of the random variables. In our study, we draw random samples from the posterior distribution of the log basal friction, i.e $p(\boldsymbol{\theta} \mid \mathcal{M}, \boldsymbol{y})$. The high-fidelity model must be evaluated on all the samples in Θ_0 and the low-fidelity model must be evaluated on both the sets Θ_0 and Θ_1. Typically $N_0 < N_1$. In most practical applications, such as this study, the model f_0 used with an

[3] Multilevel Best Linear Unbiased Estimators] Recently, multilevel best linear unbiased estimators (MLBLUEs) Schaden and Ullmann (2020) were developed as an alternative to ACV estimators to estimate the expectation of a high-fidelity model using an ensemble of models of varying cost and fidelity. However, we did not use MBLUEs in this study because they can only be used to estimate the mass-change mean and not its variance.

ACV estimate is chosen to be the highest-fidelity model that can be simulate $O(10)$ times, However, when a model utilizes a numerical discretization that can be refined indefinitely, MLMC can be used to adaptively set Q_0 such that the discretization error $Q_0 - Q_\infty$, in Eq. (18), is equal to the variance $\mathbb{V}_\Theta[Q_{\text{ACV}}]$ of the MLMC estimator.

The ACV estimator is an unbiased estimator of the mean high-fidelity model. So the MSE, Eq. (18) ignoring the model inadequacy and model discretization errors, is equal to the variance of the estimator, which when $\Theta_0 \subset \Theta_1$ is

$$\mathbb{V}_\Theta\left[Q^\mu_{\text{ACV}}(\Theta_0,\Theta_1)\right] = N_0^{-1}\mathbb{V}_\pi\left[f_0\right]\left(1-\frac{N_1-N_0}{N_1}\mathbb{C}\text{orr}_\pi[f_0,f_1]^2\right) \tag{23}$$

Thus, for fixed $N0$ and $N1$, if the high and low-fidelity models are highly correlated, the ACV estimator will be much more accurate than the SFMC estimator, see Eq. (20). Moreover, the values of N_0, N_1 can be optimized to minimize the error of an ACV estimator given a fixed computational budget. In the following section, we provide more details on how to construct ACV estimator using more than one-low fidelity model, including information on how to optimize η and the number of samples used to evaluate each model.

4.3 Many model multi-fidelity uncertainty quantification

Given an ensemble of $M+1$ models $\{f_\alpha(\theta)\}_{\alpha=0}^M$ an ACV MC estimator can be used to compute a vector-valued estimator $\boldsymbol{Q}_0 = [Q_{0,1},\cdots,Q_{0,K}]^\top \in \mathbb{R}^K$ of statistics of the highest fidelity model f_0; the specific instances of the ice-sheet models used by this study are presented in Section 5.1. The vector $\boldsymbol{Q}_0$ may be comprised of a single type of statistic computed for multiple quantities of interest (QoI), multiple statistics of a single QoI, or a combination of both. For example, in this study we computed the ACV estimator of the mean and variance of the mass change, that is $\boldsymbol{Q}_0 = [Q_0^\mu, Q_0^{\sigma^2}]^\top \in \mathbb{R}^2$.

Any ACV estimators $\boldsymbol{Q}_{\text{ACV}} = [Q^\mu_{\text{ACV}}, Q^{\sigma^2}_{\text{ACV}}]^\top \in \mathbb{R}^2$ of the vector-valued high-fidelity statistic $\boldsymbol{Q}_0$ can be expressed as

$$\boldsymbol{Q}_{\text{ACV}}(\Theta_0,\Theta_1^*,\Theta_1,\ldots,\Theta_M^*,\Theta_M) = \begin{bmatrix} Q_0^\mu \\ Q_0^{\sigma^2}\end{bmatrix} + \begin{bmatrix}\eta_{1,1} & \cdots & \eta_{1,2M}\\ \eta_{2,1} & \cdots & \eta_{2,2M}\end{bmatrix}\begin{bmatrix} Q_1^\mu(\Theta_1^*) - Q_1^\mu(\Theta_1)\\ Q_1^{\sigma^2}(\Theta_1^*) - Q_1^{\sigma^2}(\Theta_1)\\ \vdots \\ Q_M^\mu(\Theta_M^*) - Q_M^\mu(\Theta_M)\\ Q_M^{\sigma^2}(\Theta_M^*) - Q_M^{\sigma^2}(\Theta_M)\end{bmatrix},$$

where $Q_m^\mu(\Theta_m^*)$ and $Q_m^\mu(\Theta_m)$ are single model MC estimates of the mean, $\mathbb{E}_\Theta[f_m]$ Eq.(16), computed using the m-th model, $m=0,\ldots,M$, using different sample sets Θ_m^* and Θ_m. Similarly, $Q_m^{\sigma^2}(\Theta_m^*)$ and $Q_m^{\sigma^2}(\Theta_m)$ are estimates of the model variance, $\mathbb{V}_\Theta[f_m]$ Eq.(17), computed using the m-th model, In more compact notation

$$\boldsymbol{Q}_{\text{ACV}}(\Theta_{\text{ACV}}) = \boldsymbol{Q}_0(\Theta_0) + \boldsymbol{\eta}\boldsymbol{\Delta}(\Theta_\Delta), \tag{24}$$

where , $\Theta_\Delta = \{\Theta_1^*,\Theta_1,\ldots,\Theta_M^*,\Theta_M\}$, and $\Theta_{\text{ACV}} = \{\Theta_0,\Theta_\Delta\}$,

$$\boldsymbol{\Delta}(\Theta_\Delta) = \begin{bmatrix}\boldsymbol{\Delta}_1(\Theta_1^*,\Theta_1)\\ \vdots \\ \boldsymbol{\Delta}_M(\Theta_M^*,\Theta_M)\end{bmatrix} \in \mathbb{R}^{2M} \qquad \boldsymbol{\Delta}_m(\Theta_m^*,\Theta_m) = \begin{bmatrix} Q_m^\mu(\Theta_m^*) - Q_m^\mu(\Theta_m)\\ Q_m^{\sigma^2}(\Theta_M^*) - Q_m^{\sigma^2}(\Theta_m)\end{bmatrix} \in \mathbb{R}^2, \qquad m=1,\ldots,M \tag{25}$$

and the entries of $\boldsymbol{\eta} \in \mathbb{R}^{2\times 2M}$ are called control variate weights.

A multi-model ACV estimator is constructed by evaluating the highest-fidelity model at a single set of samples Θ_0 and evaluating each low-fidelity model at two sets of samples $\Theta_\alpha^* = \{\theta^{(n)}\}_{n=1}^{N_{\alpha^*}}$ and $\Theta_\alpha = \{\theta^{(n)}\}_{n=1}^{N_\alpha}$. Different ACV estimators can be produced by changing the way each sample set is structured. For example, MFMC estimators sample the uncertain parameters such that $\Theta_\alpha^* \subset \Theta_\alpha$ and $\Theta_\alpha^* = \Theta_{\alpha-1}$ and MLMC estimators sample such that $\Theta_\alpha^* \cap \Theta_\alpha = \emptyset$, and $\Theta_\alpha^* = \Theta_{\alpha-1}$.

By construction any ACV estimator is an unbiased estimator of $\boldsymbol{Q}_0$ because $\mathbb{E}_\Theta[\boldsymbol{\Delta}_\alpha] = 0, \alpha > 0$. Consequently, the MSE of the ACV estimator, Eq. (18), can be minimized by optimizing the determinant of the estimator covariance matrix. When estimating a single statistic ($K = 1$), this is equivalent to minimizing the variance of the estimator. Given sample sets Θ_{ACV}, the determinant of the covariance of an ACV estimator, $\mathbb{C}\text{ov}_\Theta[\boldsymbol{Q}_{\text{ACV}}, \boldsymbol{Q}_{\text{ACV}}]$ in Eq. (27), can be minimized using the optimal weights

$$\boldsymbol{\eta}(\Theta_{\text{ACV}}) = -\mathbb{C}\text{ov}_\Theta[\boldsymbol{Q}_0, \boldsymbol{\Delta}]\,\mathbb{C}\text{ov}_\Theta[\boldsymbol{\Delta}, \boldsymbol{\Delta}]^{-1}, \qquad \mathbb{C}\text{ov}_\Theta[\boldsymbol{Q}_0, \boldsymbol{\Delta}] \in \mathbb{R}^{2\times 2M},\ \mathbb{C}\text{ov}_\Theta[\boldsymbol{\Delta}, \boldsymbol{\Delta}] \in \mathbb{R}^{2M\times 2M}, \tag{26}$$

which produces an ACV estimator with covariance

$$\mathbb{C}\text{ov}_\Theta[\boldsymbol{Q}_{\text{ACV}}, \boldsymbol{Q}_{\text{ACV}}](\Theta_{\text{ACV}}) = \mathbb{C}\text{ov}_\Theta[\boldsymbol{Q}_0, \boldsymbol{Q}_0] - \mathbb{C}\text{ov}_\Theta[\boldsymbol{Q}_0, \boldsymbol{\Delta}]\,\mathbb{C}\text{ov}_\Theta[\boldsymbol{\Delta}, \boldsymbol{\Delta}]^{-1}\,\mathbb{C}\text{ov}_\Theta[\boldsymbol{Q}_0, \boldsymbol{\Delta}]^\top \in \mathbb{R}^{2\times 2}, \tag{27}$$

where the dependence of $\boldsymbol{\Delta}$ and $\boldsymbol{Q}_0$ on the sample sets Θ_Δ and Θ_0 was dropped to improve readability. Note that, in Eq. (26) and Eq. (27), and the remainder of this paper, we use $\mathbb{C}\text{ov}[\boldsymbol{v}, \boldsymbol{v}]$ as long hand for $\mathbb{V}[\boldsymbol{v}]$ to emphasize that the covariance is a matrix when the random variable $\boldsymbol{v}$ is a vector.

4.4 Computational considerations for multi-fidelity uncertainty quantification

The approximation of model statistics using ACV estimators is broken into two steps. The first step, referred to as the *pilot study* or *exploration phase*, involves collecting evaluations of each model on a common set of samples. These evaluations are used to compute the so called pilot statistics that are needed to evaluate Eq. (26) and Eq. (27). Subsequently, these pilot statistics are used to find the optimal sample allocation of the best estimator (see Algorithm 1). The second step, known as the *exploitation phase*, involves evaluating each model according to the optimal sample allocation and then computing the model statistics using Eq. (24). We will discuss the important computational aspects of these two phases in the following subsections.

4.4.1 Estimating pilot statistics.

Computing the covariance of an ACV estimator, $\mathbb{C}\text{ov}_\Theta[\boldsymbol{Q}_{\text{ACV}}, \boldsymbol{Q}_{\text{ACV}}]$ in Eq. (27), requires estimates of the covariance between the estimator discrepancies $\boldsymbol{\Delta}$, Eq. (25), with each other and the high-fidelity estimator and the covariance of the high-fidelity estimator, i.e. $\mathbb{C}\text{ov}_\Theta[\Delta, \Delta]$ and $\mathbb{C}\text{ov}_\Theta[Q_0, \Delta]$. In practice, these quantities, which we call *pilot statistics*, are unknown and must be estimated with a pilot study. Specifically, following standard practice (Peherstorfer and Willcox, 2016), we used MC quadrature with P, so-called, *pilot samples* $\Theta_{\text{pilot}} = \{\theta^{(p)}\}_{p=1}^{P}$ to compute the pilot statistics. This involves computing the high-fidelity and all the low-fidelity models at the same set of samples Θ_{pilot}. For example, we approximated $\mathbb{C}\text{ov}_\pi[f_\alpha, f_\beta]$,

which is needed to compute the quantities in Eq. (27), by

$$\mathbb{C}\mathrm{ov}_{\pi}\left[f_\alpha, f_\beta\right] \approx P^{-1}\sum_{p=1}^{P}\left(f_\alpha(\theta^{(p)}) - Q_\alpha^\mu(\Theta_{\text{pilot}})\right)\left(f_\beta(\theta^{(p)}) - Q_\beta^\mu(\Theta_{\text{pilot}})\right) \in \mathbb{R}^{2\times 2}, \tag{28}$$

Please refer to Dixon et al. (2023) to see the additional quantities needed to compute the covariance blocks of $\mathbb{C}\mathrm{ov}_\Theta\left[\boldsymbol{\Delta}_\alpha, \boldsymbol{\Delta}_\beta\right]$ and $\mathbb{C}\mathrm{ov}_\Theta\left[\boldsymbol{Q}_0, \boldsymbol{\Delta}_\alpha\right]$, which are required to compute the estimator covariance $\mathbb{C}\mathrm{ov}_\Theta\left[\boldsymbol{Q}_{\text{ACV}}, \boldsymbol{Q}_{\text{ACV}}\right]$ of a vector-valued statistic that consists of both the mean and variance of a model. Finally, we recorded the CPU time needed to simulate each model at all pilot samples and set the model costs $\boldsymbol{w}^\top = [w_0, w_1, \ldots, w_M] \in \mathbb{R}^{M+1}$ to be the the median simulation time of each model.

Unfortunately, using a finite P introduces sampling errors into $\mathbb{C}\mathrm{ov}_\Theta\left[\boldsymbol{\Delta}, \boldsymbol{\Delta}\right]$ and $\mathbb{C}\mathrm{ov}_\Theta\left[\boldsymbol{Q}_0, \boldsymbol{\Delta}\right]$, which in turn induces error in the ACV estimator covariance, Eq. (27). This error can be decreased by using a large P but this would require additional evaluations of expensive numerical models, which we were trying to avoid. Consequently, in this study we investigated the sensitivity of the number of pilot samples on the error in ACV MC estimators. Results of this study are presented in Section 5.

4.4.2 Optimal computational resource allocation.

The covariance of an ACV estimator, $\mathbb{C}\mathrm{ov}_\Theta\left[\boldsymbol{Q}_{\text{ACV}}, \boldsymbol{Q}_{\text{ACV}}\right]$ in Eq. (27), is dependent on how samples are allocated to the sets $\Theta_\alpha, \Theta_\alpha^*$, which we call the sample allocation $\mathcal{A}$. $\mathcal{A}$ uniquely defines the allocation strategy by listing the number of samples of each set Θ_α and Θ_α^* and their pairwise intersections. Namely, $\mathcal{A} = \{N_0, N_{\alpha\cap\beta}, N_{\alpha^*\cap\beta}, N_{\alpha\cap\beta^*}, N_{\alpha^*\cap\beta^*} \mid \alpha, \beta = 1, \ldots, M\}$, where $N_{\alpha\cap\beta} = |\Theta_\alpha \cap \Theta_\beta|$, $N_{\alpha^*\cap\beta} = |\Theta_\alpha^* \cap \Theta_\beta|$, $N_{\alpha\cap\beta^*} = |\Theta_\alpha \cap \Theta_\beta^*|$, $N_{\alpha^*\cap\beta^*} = |\Theta_\alpha^* \cap \Theta_\beta^*|$ denote the number of samples in the intersections of pairs of sets, and $N_{\alpha\cup\beta} = |\Theta_\alpha \cup \Theta_\beta|$, $N_{\alpha^*\cup\beta} = |\Theta_\alpha^* \cup \Theta_\beta|$, $N_{\alpha\cup\beta^*} = |\Theta_\alpha \cup \Theta_\beta^*|$, $N_{\alpha^*\cup\beta^*} = |\Theta_\alpha^* \cup \Theta_\beta^*|$ denote the number of samples in the union of pairs of sets . Thus, the best ACV estimator can be theoretically found by solving the constrained non-linear optimization problem

$$\min_{\mathcal{A}\in\mathbb{A}} \mathrm{Det}\left[\mathbb{C}\mathrm{ov}_\Theta\left[\boldsymbol{Q}_{\text{ACV}}, \boldsymbol{Q}_{\text{ACV}}\right](\mathcal{A})\right] \qquad \text{s.t.} \qquad W(\boldsymbol{w}, \mathcal{A}) \le W_{\max}. \tag{29}$$

In the above equation, $\mathbb{A}$ is the set of all possible sample allocations and the constraint ensures that the computational cost of computing the ACV estimator

$$W(\boldsymbol{w}, \mathcal{A}) = \sum_{\alpha=0}^{M} N_{\alpha^*\cup\alpha} w_\alpha$$

is smaller than a computational budget $W_{\max} \in \mathbb{R}$. The solution to this optimization problem is often called the optimal sample allocation.

Unfortunately, a tractable algorithm for solving Eq. (29) has not yet been developed, largely due to the extremely high number of possible sample allocations in the set $\mathbb{A}$. Consequently, various ACV estimators have been derived in the literature that simplify the optimization problem, by specifying what we call the sample structure $\mathcal{T}$, which restricts how samples are shared between the sets $\Theta_\alpha, \Theta_\alpha^*$. For example, optimizing the estimator variance, Eq. (23), of a two model MFMC (Peherstorfer

et al., 2016) mean estimator, Eq. (21), requires solving

$$\min_{N_0,N_1} N_0^{-1}\mathbb{V}[f_0]\left(1-\frac{N_1-N_0}{N_1}\mathbb{C}\text{orr}[f_0,f_1]^2\right)$$
$$\text{s.t.} \quad N_0w_0+N_1w_1\le W_{\max}, \qquad \mathcal{T}=\{N_{0\cap 1*}=N_0, N_{0\cup 1*}=N_0, N_{0\cap 1}=N_0, N_{0\cup 1}=N_1, N_{1*\cap 1}=N_0, N_{1*\cup 1}=N_1\}.$$

Alternatively, minimizing the estimator variance of the two model MLMC (Giles, 2015)[4] mean estimator requires solving

$$\min_{N_0,N_1} N_0^{-1}\mathbb{V}[f_1-f_0]+(N_1-N_0)^{-1}\mathbb{V}[f_1]$$
$$\text{s.t.} \quad N_0w_0+N_1w_1\le W_{\max}, \qquad \mathcal{T}=\{N_{0\cap 1*}=N_0, N_{0\cup 1*}=N_0, N_{0\cap 1}=0, N_{0\cup 1}=N_1, N_{1*\cap 1}=0, N_{1*\cup 1}=N_1\}.$$

MLMC and MFMC employ sample structures $\mathcal{T}$ that simplify the general expression for the estimator covariance $\mathbb{C}\text{ov}_\Theta[\boldsymbol{Q}_{\text{ACV}},\boldsymbol{Q}_{\text{ACV}}]$ in Eq. (27). These simplifications were used to derive analytically solutions of the sample allocation optimization problem in Eq. (29) when estimating the mean, $\mathbb{E}_\Theta[f_0]$ in Eq. (16), for a scalar-valued model. However, the optimal sample allocation of MLMC and MFMC must be computed numerically when estimating other statistics, such as variance $\mathbb{V}_\Theta[f_0]$ in Eq. (17). Similarly, numerical optimization must be used to optimize the estimator covariance, $\mathbb{C}\text{ov}_\Theta[\boldsymbol{Q}_{\text{ACV}},\boldsymbol{Q}_{\text{ACV}}]$ in Eq. (27), of most other ACV estimators, including the ACVMF and ACVIS (Gorodetsky et al., 2020), as well as their tunable generalizations (Bomarito et al., 2022).

Each existing ACV estimator was developed to exploit alternative sample structures $\mathcal{T}$ to improve the performance of ACV estimators in different settings. For example, a three model ACVMF estimator performs well when the low-fidelity models are conditionally independent of the high-fidelity model. Imposing this conditional independence is useful when knowing one-low-fidelity does not provide any additional information about the second low-fidelity model, given enough samples of the high-fidelity model. This situation can arise when the low-fidelity models use different physics simplifications of the high-fidelity model. In contrast, MLMC assumes that each model in the hierarchy is conditionally independent of all other models given the next highest fidelity model. This allows MLMC to perform well with with a set of models ordered in a hierarchy by bias relative to the exact solution of the governing equations.

The performance of different ACV estimators is problem dependent. Consequently, in this paper we investigated the use of a large number of different ACV estimators from the literature. For each estimator we used the general purpose numerical optimization algorithm proposed in Bomarito et al. (2022) to find the optimal sample allocations that minimize the determinant of the estimator covariance.[5]

[4]MLMC estimators set all the control variate weights in Eq. (24) to $\boldsymbol{\eta}=-1$. Refer to Gorodetsky et al. (2020) for more details on the connections between ACV and MLMC.

[5]The presentation of the optimization algorithms in (Bomarito et al., 2022) focuses on the estimation of a single statistic, but can be trivially be extended to the vector-valued QoI considered here.

4.4.3 Model and estimator selection.

Using data from all available models may produce an estimator with larger MSE than an estimator that is only constructed using a subset of the available models. This occurs when a subset of the low-fidelity models correlate much better with the high-fidelity model than the rest of the low-fidelity models. For instance, some low-fidelity models may fail to capture physical behaviours that are important to estimating the QoI. Consequently, it is difficult to determine the best estimator a priori. However, we can accurately predict the relative performance of any ACV estimator using only the model simulations run during the pilot study. Thus, in this study we enumerated a large set of estimator types encoded by the different sample structures $\mathcal{T}$ and model subsets.

Algorithm 1 summarizes the procedure we use to choose the best ACV estimator. Line 8 loops over all model subsets $\mathcal{S}$. In this study, we enumerated all permutations of the sets of models that contained the high-fidelity model and at most 3 low-fidelity models. Line 10 enumerates each parametrically defined estimator E. We enumerated the large sets of parametrically defined generalized multi-fidelity (GMF), generalized independent sample (GIS) and generalized recursive difference (GRD) ACV estimators introduced by Bomarito et al. (2022). These sets of estimators include ACVMF, MFMC, MLMC (with optimized control variate weights) as special cases. For each estimator E and model subset $\mathcal{S}$, line 12 was used to find the optimal sample allocation $\mathcal{A}_E$, using the pilot values $\{f_\alpha(\Theta_{\text{pilot}})\}_{\alpha\in\mathcal{S}}$ when minimizing Eq. (29). Lines 13-16 were used to record the best estimator found at each iteration of the outer-loops.

Whether a model is useful for reducing the MSE error of a multi-fidelity estimator depends on the correlations between that model, the high-fidelity, and the other low-fidelity models. For toy parameterized PDE problems, such as the diffusion equation with an uncertain diffusion coefficient, theoretical convergence rates and theoretical estimates of computational costs can be used to rank models. However, for the models we used in this study, and likely many other ice-sheet studies, ordering models hierarchically, that is, by bias or correlation relative to the highest-fidelity model, before evaluating them is challenging. Indeed, the best model ensemble for multi-fidelity UQ may not be hierarchical (see Gorodetsky, 2020). Yet, estimators such as MLMC and MFMC only work well on model hierarchies. Consequently, having a practical approach for learning the best model ensemble is needed. Yet, to date this issue has received little attention in the multi-fidelity literature. Section 5 provides a sorely needed discussion of the impact of the pilot study on model selection and the error a multi-fidelity estimator.

5 Results

This section presents the results of our MFSE study. First, we describe the ensemble of numerical models we used to solve the governing equations presented in Section 5.1. Second, we summarize the results of our Bayesian model calibration. Third, in Section 5.3 we present the results of our pilot study. Specifically, we compare the computational costs of each model and their SFMC-based estimates of the mean and variance of the mass change computed using the pilot samples. We also report the MSE of ACV estimators predicted using the pilot and note the subset of models they employed. Fourth, we detail the impact of increasing the number of pilot samples on the predicted MSE of the ACV estimators in Section 5.4. Finally, we quantify the

Algorithm 1 Estimator selection

Input

$\{f_\alpha(\Theta_{\text{pilot}})\}_{\alpha=0}^{M}$ Pilot evaluations of each model

Output

$\mathcal{A}_{\text{best}}$ Best estimator sample allocation

J_{best} Best estimator objective

$J_{\text{best}} \leftarrow \infty$

▷ Loop over all low-fidelity model subsets

for $\mathcal{S} \subseteq \{1, \ldots, M\}$ **do**

  ▷ Loop over all MF estimators, e.g MLMC, MFMF, ACVMF, etc

  for $E \in \mathcal{E}$ **do**

    ▷ Compute the optimal estimator objective J_E and sample allocation $\mathcal{A}_E$ for the current estimator and subset of models

    $J_E, \mathcal{A}_E \leftarrow$ Solve Eq. (29) using $\mathbb{A}_{E,\mathcal{S}}$ and $\{f_\alpha(\Theta_{\text{pilot}})\}_{\alpha\in\mathcal{S}}$

    if $J_E < J_{\text{best}}$ **then**

      ▷ Update the best estimator

      $\mathcal{A}_{\text{best}} \leftarrow \mathcal{A}_E$

      $J_{\text{best}} \leftarrow J_E$

    end if

  end for

end for

improvement in the precision of MFSE estimates of mass-change statistics relative to SFMC in Section 5.5. All results were generated with the `PyApprox` software package (Jakeman, 2023).

5.1 Multi-fidelity model ensemble

In this study we investigated the use of 13 different models of varying computational cost and fidelity to compute glacier mass change. Specifically, we used MFSE to estimate the mean and variance of a highly-resolved finite element model using an ensemble of 12 low-fidelity models. We compactly denote the fidelity of each model using the notation $\text{PHYSICSNAME}_{dx,dt}$, where PHYSICSNAME refers to the governing equations solved, dx denotes the size of the representative spatial element and dt the size of the time-step. The four different meshes we used are shown in Figure 2.

The highest-fidelity model we considered in this study was a MOLHO-based model denoted by $\text{MOLHO}^{*}_{1km,9days}$, where the star indicates that the model was modified to conserve mass. The low-fidelity model ensemble consisted of four MOLHO-based low-fidelity models, $\text{MOLHO}_{1km,36days}$, $\text{MOLHO}_{1.5km,36days}$, $\text{MOLHO}_{2km,36days}$, $\text{MOLHO}_{3km,36days}$, and eight SSA-based low-fidelity models $\text{SSA}_{1km,36days}$,$\text{SSA}_{1.5km,36days}$, $\text{SSA}_{2km,36days}$, $\text{SSA}_{3km,36days}$, $\text{SSA}_{1km,365days}$, $\text{SSA}_{1.5km,365days}$,

$\text{SSA}_{2km,365days}$, $\text{SSA}_{3km,365days}$. The number of elements associated with the four meshes with characteristic element sizes 1km, 1.5km, 2km and 3km, were 2611, 9238, 13744, 22334, respectively. The number of nodes for the same four meshes were 1422, 4846, 7154, 11536. Note, no low-fidelity model enforced the conservation of mass.

The models we used were all different numerical discretizations of two distinct physics models. However, in the future we could also use alternative classes of low-fidelity models, if they become available. For example, we could use linearizations of the parameter-to-QoI map, proposed by Recinos et al. (2023), if our MOLHO and/or SSA codes become capable of efficiently computing the gradient of the map. Such an approach would require only one non-linear forward transient solve of the governing equations followed by a linear solve of the corresponding backward adjoint. Once constructed, the linearized map could then be evaluated very cheaply and used to reduce the estimator variance, $\mathbb{C}\text{ov}_\Theta\left[\boldsymbol{Q}_{\text{ACV}}, \boldsymbol{Q}_{\text{ACV}}\right]$ in Eq. (27), of the MFSE estimators, provided the error introduced by the linearization is not substantial. Other types of surrogates could also be used in principle, however, the large number of parameters used pose significant challenges to traditional methods such as the Gaussian processes used in Jantre et al. (2024). Recently developed machine-learning surrogates (Jouvet et al., 2021; Brinkerhoff et al., 2021; He et al., 2023) could be competitive alternatives to the low fidelity models considered in this work.

Lastly, note that we used a different model, to the 13 described above, for the Bayesian calibration of the basal friction parameters. Specifically, we used the C++ code `MALI` (Hoffman et al., 2018), which can solve the Blatter-Pattyn equations (Pattyn, 2003; Dukowicz et al., 2010) and compute the action of the Hessian on a vector. `MALI` efficiently computed these Hessian-vector products, needed to compute our Laplace approximation of the posterior in Eq. (14), by solving the adjoint equations for the steady state Blatter-Pattyn equations. However, SSA equations (Section 2.1.3) are not currently implemented in `MALI` and the MOLHO (Section 2.1.2) equations have only recently been implemented (after the simulations for this work were performed). Consequently, we used `FEniCS` (Alnæs et al., 2015) to implement both MOLHO and SSA to ensure that the relative computational timings of these models would be consistent. Solving the Blatter-Pattyn model using the C++-based `MALI` code and solving MOLHO and SSA using the python based `FEniCS`, would have corrupted the MFSE results. Moreover, implementing SSA in `MALI` would be time consuming because it is currently only used to solve 3D models and not 2D models, such as SSA. Indeed, a partial motivation for this study was to to determine the utility of implementing the SSA equations in `MALI`.

5.2 Bayesian model calibration

In this study we used the `MALI` ice-sheet code (Hoffman et al., 2018; Tezaur et al., 2021) to calibrate the basal friction field on the finest mesh, as described in Section 3. The MAP point of the posterior, determined using Eq. (13), is depicted in the left panel of Figure 5. The pointwise variance of the Laplace approximation of the posterior of the log-friction (i.e. the diagonal of Eq. (15)) is depicted in the right-panel of Figure 5. When this figure is compared to the pointwise variance of the prior (i.e. the diagonal of Eq. (9)) depicted in the center panel of Figure 5, it is clear that conditioning the prior uncertainty on the surface velocity significantly reduced the uncertainty in the basal friction field. This conclusion is further corroborated by Figure 6, which compares a random sample from the prior and a random sample from the posterior. The minimum and maximum values of the posterior sample of the log-friction are much smaller than the same bounds of the prior sample. However, the posterior sample has much higher-frequency content because the data only informed the lower-frequency modes of the friction field.

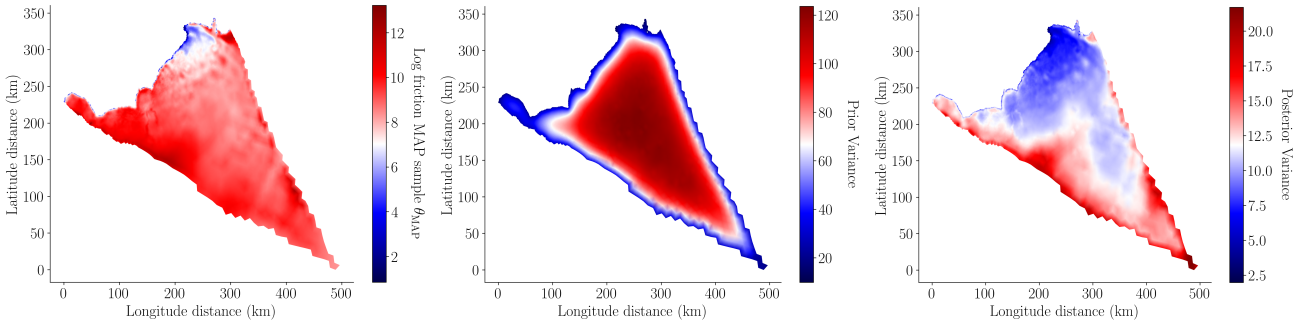


Figure 5. (*Left*) Log of the basal friction MAP point, $\boldsymbol{\theta}_{\text{MAP}}$ computed using Eq. (13). (*Center*) Pointwise prior variance, i.e. the diagonal entries of $\boldsymbol{\Sigma}_{\text{prior}}$, defined in Eq. (9). (*Right*) Pointwise posterior variance, i.e. the diagonal entries of, $\boldsymbol{\Sigma}_{\text{post}}$, defined in Eq. (15).

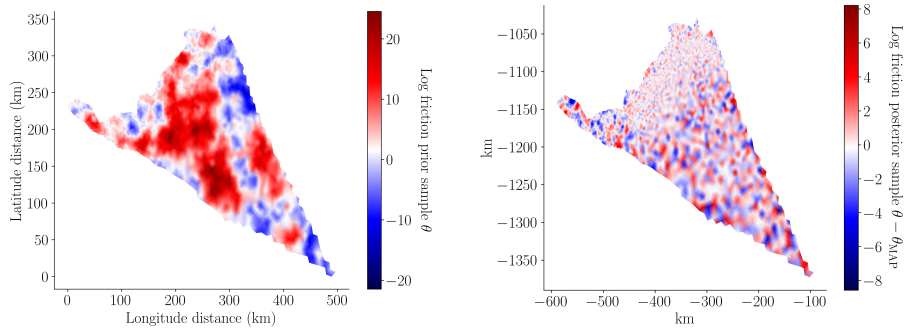


Figure 6. (*Left*) A random sample from the prior distribution of the log-friction $p(\boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{prior}})$, where $\boldsymbol{\Sigma}_{\text{prior}}$ is defined in Eq. (9). (*Right*) A random sample from the Laplace approximation of the posterior $p(\boldsymbol{\theta} \mid \mathcal{M}, \boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\theta}_{\text{MAP}}, \boldsymbol{\Sigma}_{\text{post}})$, defined in Eq. (14).

To demonstrate a projection to 2100 using a calibrated model, figure 7 depicts the difference between the final (year 2100) and initial (year 2007) ice thickness and the final surface velocity of Humboldt Glacier computed by the highest-fidelity model ($\text{MOLHO}^*_{1km,9days}$) for a random posterior realization of the basal friction field. The final ice thickness shown differs substantially from the initial thickness with thickness decreasing substantially at lower elevations of the glacier in the ablation zone where increasingly negative surface mass balance occurs through 2100. In general, the glacier speeds up as negative surface mass balance causes the surface to steepen near the terminus. The largest speedup occurs in the region of fast flow in the north where basal friction is small. Also note that the high-frequency differences present in the thickness difference was due to the high-frequency oscillations in the posterior sample, see Figure 6.

Remark 5.1 (Prior sensitivity)**.** *In this study we used our domain experience to determine the best values of the prior hyper-parameters γ, δ, η reported in Section 2.3 and the likelihood hyper-parameter α in Eq. (12). However, varying these hyper-parameters, would likely change the estimates of uncertainty in ice-sheet predictions produced by this study. Similar*

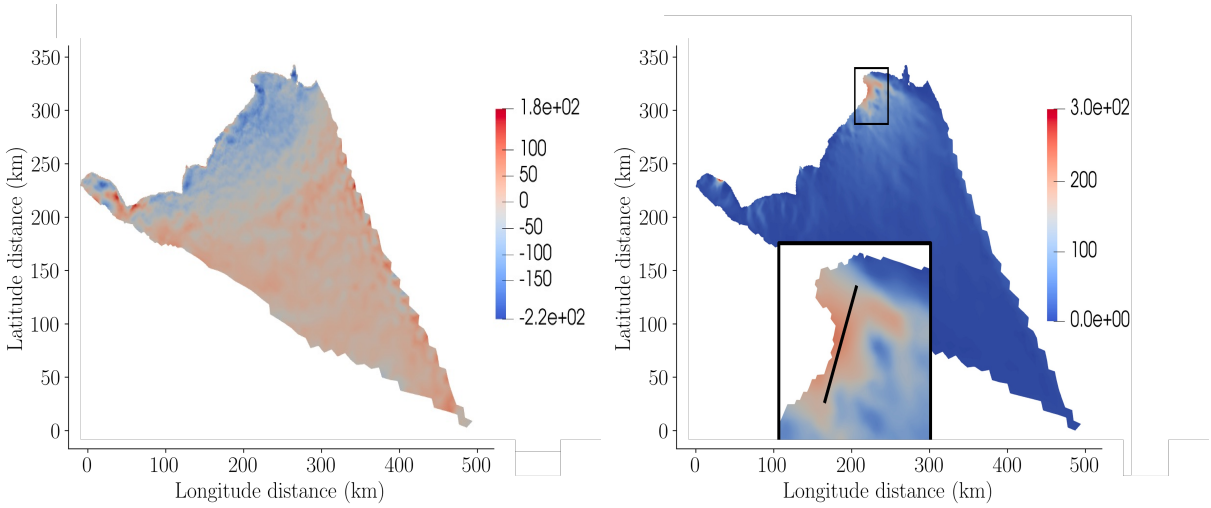


Figure 7. (*Left*) The difference between the final and initial ice thickness in meters and (*Right*) the surface velocity of Humboldt Glacier. The left box is a zoomed in picture of the top right tip of the glacier. The black line in the inset was used to plot cross-sections of the thickness and friction profiles at 2100 in a region with high velocities (see Figure 14). Both the left and right figures were generated using the highest-fidelity model $\text{MOLHO}^*_{1km,9days}$ evaluated at one random realization of the posterior of the basal friction field.

to previous studies (Isaac et al., 2015), we did not investigate these sensitivities extensively. We heuristically chose the prior hyper-parameters so that the prior samples would have a variance and spatial variability that we deemed inline with our domain experience. Further, we found that reducing α substantially from the value we ultimately used while keeping the prior hyper-parameter fixed, prevented the MAP point from capturing the high-frequency content of the basal friction field needed to accurately match the observed surface velocities. Future studies should investigate the sensitivity of mass change to the values of the hyper-parameters more rigorously using an approach such as the one developed by Recinos et al. (2023).

Remark 5.2 (Interpolating basal friction)**.** *In this study we drew samples from the posterior distribution of the friction parameters defined on the finest spatial mesh. However, a posterior sample defined on the fine mesh cannot be used to predict mass change with a low-fidelity model defined on a coarser mesh. Consequently, before using a low-fidelity model with a coarse mesh to predict mass change, we first interpolated each sample of the posterior distribution of the basal friction field defined on the finest mesh onto the mesh used by the low-fidelity model.*

5.3 Initial Pilot study

This section details the results of the pilot study that we used to obtain the computational cost, $\boldsymbol{w}$, of each model and the pilot statistics, e.g. Eq. (28), needed to construct ACV estimators. First, we evaluated each of our 13 models at the same 20 random pilot samples of the model parameters Θ_{pilot}, i.e. 20 different basal friction fields drawn from the Laplace approximation of the posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{M}, \boldsymbol{y})$, Eq. (14). Second we computed the median computational cost (wall-time), $\boldsymbol{w}$, required to

solve each model at one pilot sample. The median computational costs are plotted in the top panel of Figure 8 and the total cost of the evaluating all 13 models was approximately 144 hours. Third, using the pilot samples, we computed the SFMC estimators of the mean, Eq. (16), and standard deviation , using the square-root of Eq. (17), of the mass change predicted by each of the 13 models. The middle and lower panels of Figure 8 show that the means and standard deviations of each model differ. However in the next section, we show that despite the differences between the statistics computed using each model and the differences between the ice-evolution predicted by each model (see Figure 14), MFSE was able to effectively increase the precision of the mean and variance of the mass change, relative to SFMC.

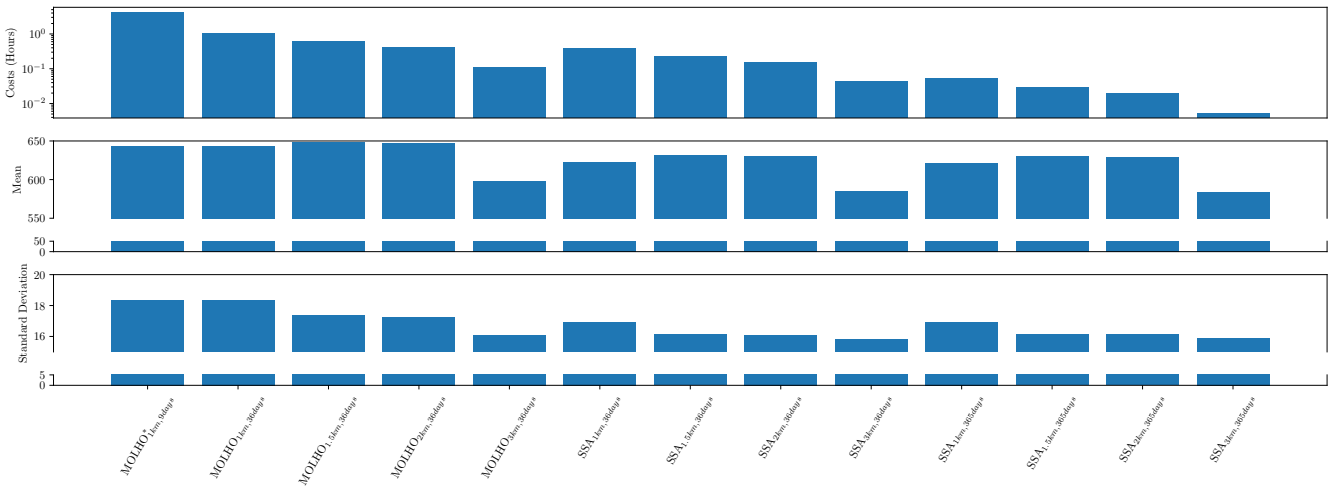


Figure 8. (*Top*) The median computational cost $\boldsymbol{w}$ (wall-time in hours) of simulating each model used in this study for one realization of the random parameters. (*Middle*) The mean mass loss – negative expected mass change – (gigatons) at 2100. (*Bottom*) The standard deviation of the mass change (gigatons) at 2100. Each quantity was computed using 20 pilot samples.

The exact gain in performance achieved by MFSE is dependent on the correlations between each model and the other pilot quantities needed to compute $\mathbb{C}\text{ov}_\Theta[\Delta, \Delta]$ and $\mathbb{C}\text{ov}_\Theta[Q_0, \Delta]$ in Eq. (27). Consequently, in Figure 9 we plot the entries of the correlation matrix, $\mathbb{C}\text{orr}_\pi[\boldsymbol{f}, \boldsymbol{f}]$ with $\boldsymbol{f} = [f_0, \ldots, f_M]^\top$. This figure shows that, despite the differences between each model's prediction of ice thickness and velocities at the final time (see Figure 14), as well as the variations in the SFMC estimate of the mean and variance computed using each model, the correlation between each model's prediction of the mass change is high.[6] However, inspecting the correlation between models can only qualitatively suggest the relative utility of each model for reducing the error of an MFSE estimator. Thus, to be more precise, we used Eq. (29), and our pilot statistics, to predict the determinant of the the ACV estimator covariance, $\text{Det}[\mathbb{C}\text{ov}_\Theta[\boldsymbol{Q}_{\text{ACV}}, \boldsymbol{Q}_{\text{ACV}}]]$. Specifically, we made these predictions assuming that a budget of 160 high-fidelity model evaluations would be allocated to the high-and low-fidelity models. Moreover, this cost was assumed additional to the computational cost of simulating each model at the pilot samples. We then computed the so

[6] The correlation between $\text{MOLHO}^*_{1km,9days}$ and $\text{MOLHO}_{1km,36days}$, reported in Figure 9, was not exactly 1. Each correlation was rounded to 4 significant digits.

called *variance reductions* of the ACV estimator

$$\mathcal{R}_\Theta[Q^\mu_{\text{ACV}}] = \mathbb{V}_\Theta\left[Q^\mu_0\right] / \mathbb{V}_\Theta\left[Q^\mu_{\text{ACV}}\right] \quad \text{and} \quad \mathcal{R}_\Theta[Q^{\sigma^2}_{\text{ACV}}] = \mathbb{V}_\Theta\left[Q^{\sigma^2}_0\right] / \mathbb{V}_\Theta\left[Q^{\sigma^2}_{\text{ACV}}\right] \tag{30}$$

by extracting the diagonal elements of the estimator covariance, $\mathbb{C}\text{ov}_\Theta\left[\boldsymbol{Q}_{\text{ACV}}, \boldsymbol{Q}_{\text{ACV}}\right]$ in Eq. (27). To ensure a fair comparison we compared the ACV estimator variance to the SFMC estimator variance obtained using a computational budget equivalent to 160 high-fidelity evaluations plus the computational cost of collecting the pilot model evaluations.

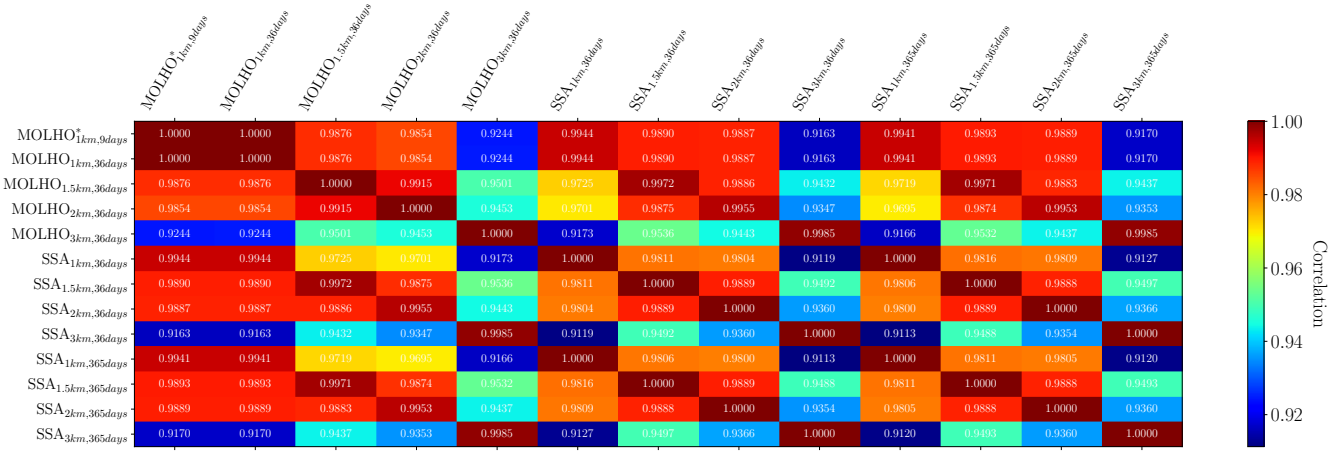
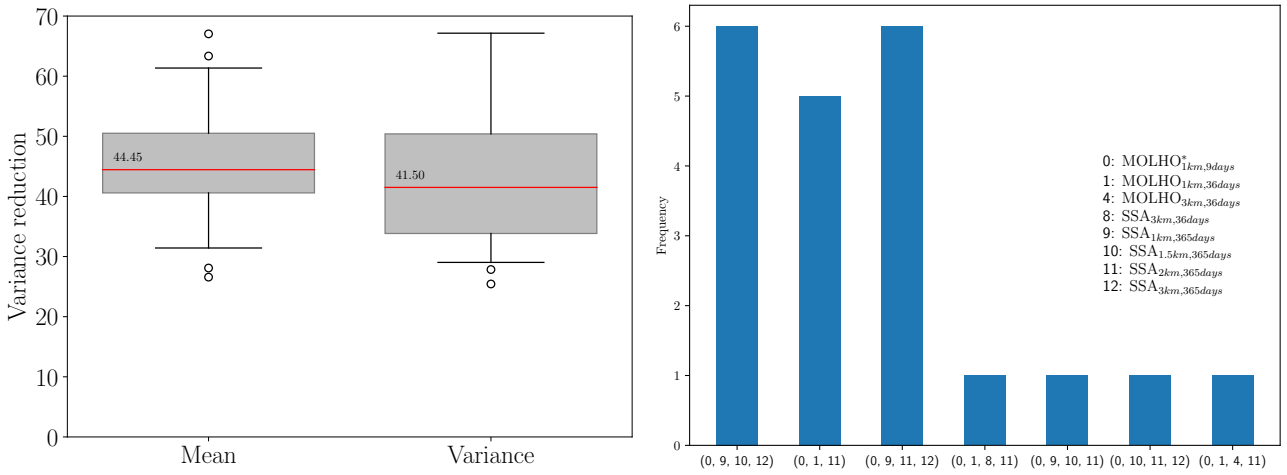


Figure 9. The correlations, , $\mathbb{C}\text{orr}_\pi[\boldsymbol{f}, \boldsymbol{f}]$ with $\boldsymbol{f} = [f_0, \ldots, f_M]^\top$, between the 13 ice-sheet models considered by this study using 20 pilot samples.

Existing literature assumes that the pilot statistics used with Eq. (29) are exact, however using a small number of pilot samples can introduce error into the estimator covariance $\mathbb{C}\text{ov}_\Theta\left[\boldsymbol{Q}_{\text{ACV}}, \boldsymbol{Q}_{\text{ACV}}\right]$. Moreover, we found that the error introduced by using a small number of pilot samples can be substantial, yet it is typically ignored in existing literature. Consequently, in Figure 10a we plot the variance reduction of the ACV estimators of the mean, Q^μ_{ACV}, and variance, $Q^{\sigma^2}_{\text{ACV}}$, of mass loss for 21 different bootstraps of the 20 pilot samples (1 bootstrap was just the original pilot data and each bootstrap set contained 20 samples). The plot is created by randomly sampling the model evaluations with replacement, computing the pilot statistics with those samples, and solving Eq. (29). Please note that, while we enumerate over numerous estimators, each with a different variance reduction, the variability in the plots is induced entirely by the bootstrapping procedure we employed. The box plots report the largest variance reduction, across all estimators, for each bootstrapped sample.

The median variance reduction was over 40 for the ACV estimators of both the mean and variance of the mass change. In other words, our initial pilot study predicted that using ACV estimators would reduce the cost of estimating uncertainty in projections of the mass change by over a factor of 40 when compared to SFMC estimators, which only use the highest-fidelity model. However, the box plots in Figure 10a highlight that using only 20 samples introduces a large degree of uncertainty into the estimated variance reduction. The 10% quantile of the variance reduction for both the mean and variance estimators were close to 30.

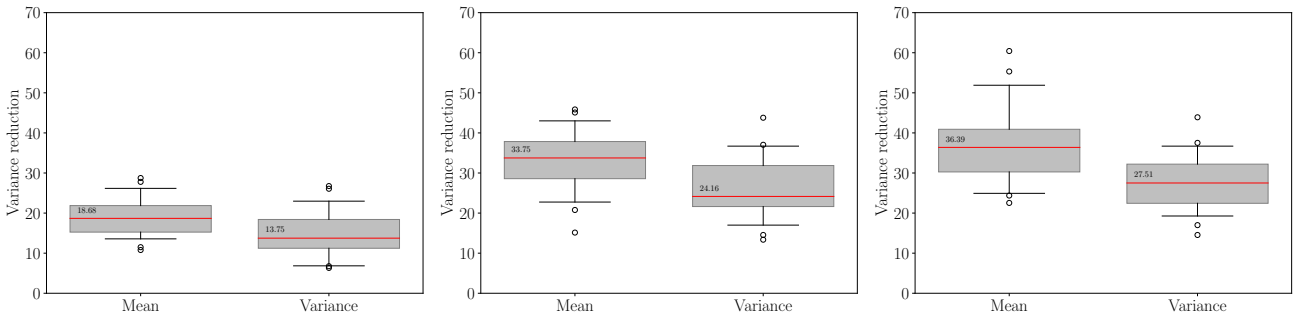


(a) Variance reductions

(b) Selected model subsets

Figure 10. (a) The predicted variance reductions $\mathcal{R}_\Theta[Q^\mu_{\text{ACV}}]$ (mean) and $\mathcal{R}_\Theta[Q^{\sigma^2}_{\text{ACV}}]$ (variance), see Eq. (30), obtained using bootstrapping of the initial 20 pilot samples. The red lines represent the median estimator variance reductions. The lower and upper whiskers represent the 10% and 90% quantiles. Note, two outliers, with values 73 and 125, do not appear on the plot for $\mathcal{R}[Q^{\sigma^2}_{\text{ACV}}]$. (b) The model subsets chosen by the bootstrapped estimators using the initial 20 pilot samples.

The estimators obtained by bootstrapping the initial 20 pilot samples not only had different estimator variances (see Figure 10a), they also predicted that different model subsets (combinations of models) are needed to minimize the estimator variance. Figure 10b plots the model subsets chosen by the bootstrapped estimators and the number of times (frequency) each subset was chosen; the set (0, 9, 10, 12) was chosen when the original 20 pilot samples were used (bootstrapping was not used). Moreover, bootstrapping the estimators also revealed that using all models simultaneously, to reduce the variance of the ACV estimator, was not as effective as using a smaller subset of models. Specifically, only eight out of the 13 models considered were chosen at least once by a bootstrapped estimator. The models $\text{MOLHO}_{1.5km,36days}$, $\text{MOLHO}_{2km,36days}$, $\text{SSA}_{1km,36days}$, $\text{SSA}_{1.5km,36days}$, $\text{SSA}_{2km,36days}$, were never selected by any of the bootstrapped estimators. Moreover, in some cases only two low-fidelity models were chosen and in other cases three low-fidelity models were chosen. Lastly, not only did the chosen model subsets vary between bootstrapped estimators, the type of estimator chosen also varied. In 7 cases, a hierarchical relationship was identified and in the other 14 cases a non-hierarchical relationship was identified; a non-hierarchical estimator was chosen using the original 20 pilot samples (the 21st estimator). Recall, a model ensemble is hierarchical if it can be ordered by bias or correlation relative to the highest-fidelity model and each low-fidelity is only used to reduce the variance of the estimator of the next highest-fidelity model in a recursive fashion.



(a) At most two models (b) At most three models (c) At most four models

Figure 11. The predicted variance reductions, $\mathcal{R}[Q^{\mu}_{\mathrm{ACV}}]$ (mean) and $\mathcal{R}[Q^{\sigma^2}_{\mathrm{ACV}}]$ (variance), see Eq. (30), of the best ACV estimators obtained by bootstrapping the final 30 pilot samples, while enforcing a limit on the number of models an estimator can use, including the highest-fidelity model. The red lines indicate the median estimator variance reductions. The lower and upper whiskers represent the 10% and 90% quantiles.

5.4 Secondary pilot study

Upon quantifying the impact of only using 20 pilot samples on the estimator covariance, $\mathbb{C}\mathrm{ov}_{\Theta}\left[\boldsymbol{Q}_{\mathrm{ACV}}, \boldsymbol{Q}_{\mathrm{ACV}}\right]$, and the model subsets, $\mathcal{S}$, chosen by Algorithm 1, we incremented the number of pilot samples we used to compute the performance of the ACV estimators. To avoid wasting computational resources in our secondary pilot study, we only evaluated the 8 models, selected by at least one bootstrapped estimator, on an additional 10 pilot samples. The combined cost of the initial and secondary pilot study was approximately 197 hours, which equated to the equivalent of approximately 47 simulations of the highest-fidelity model. Note that only the models included in the second pilot were simulated 30 times. The models only included in the first pilot were simulated 20 times.

Figure 11c plots the variance reductions of the mean and variance of mass loss , given by $\mathcal{R}[Q^{\mu}_{\mathrm{ACV}}]$ and $\mathcal{R}[Q^{\sigma^2}_{\mathrm{ACV}}]$, respectively, as the maximum number of models used by the ACV estimators is increased. Note, an estimator allowed to choose four models may still choose less than four models, which will happen when some of those models are not highly-informative. 21 different bootstraps of the final 30 pilot samples were used to quantify the error of the variance reductions caused by only using a small number of pilot samples. Comparing Figure 11c with Figure 10a, which plots variance reductions using only 20 pilot samples, we observed that increasing the number of pilot samples decreased the variability of the estimator variances. However, this increasing the number of pilot samples also increased the computational cost of the pilot study, which in turn reduced the reported median variance reduction. That is, the median variance reductions obtained using 30 pilot samples (Figure 11c) was lower than the median variance reductions reported using 20 pilot samples (Figure 10a).

The median variance reduction decreased because ACV estimators utilize the pilot samples solely to compute pilot statistics, such as variance, and do not reuse these samples for calculating the final statistics. In contrast, an equivalent SFMC estimator can leverage both the pilot and exploitation budgets to estimate the final statistics. In other words, the variance of an SFMC

estimator decreases linearly with the number of pilot samples, whereas the variance of an ACV estimator does not exhibit the same behavior. Specifically, the variance of an ACV estimator is only marginally affected by an increase in the number of pilot samples, as the sample allocation becomes more optimal.

While increasing the number of pilot samples decreased variability, we believed that the benefit of further increasing the number of pilot samples would be outweighed by the resulting drop in the variance reduction. Despite the remaining variability in the variance reduction, we were able to confidently conclude that the MSE of the final ACV estimator we would construct would be much smaller than the MSE of a SFMC estimator of the same cost because even the smallest variance reduction was greater than 14. Consequently, we used the unaltered 30 pilot samples to determine the ACV estimator and its optimal sample allocation, which we used to construct our final estimates of the mean and variance of the mass change. The best estimator chosen was an MFMC estimator that used the three models $\text{MOLHO}^*_{1km,9days}$, $\text{MOLHO}_{1km,36days}$, $\text{SSA}_{1.5km,365days}$.

5.5 Multi-fidelity sea-level rise projections

The cost of constructing our final estimator, $\boldsymbol{Q}_{\text{ACV}}$ in Eq. (24), was equal to the sum of the pilot cost (197.13 hours) and the exploitation cost (160×4.18) hours, which was approximately 36 days on a single CPU. The pilot cost was the sum of evaluating all 13 models on the initial 20 pilot samples and 8 models on an additional 10 pilot samples (see Section 5.4). The exploitation cost was fixed at the beginning of the study to the computational cost equivalent to evaluating the high-fidelity model 160 times, which takes a median time of 4.18 hour to simulate on a single realization of basal friction. The number of samples allocated to evaluating each model by the ACV estimator during the exploitation phase are shown Figure 12. Only two samples of the high-fidelity model were used. Yet, while running these simulations only accounted for approximately 1.25% of the total computational cost budget, these samples ensured the estimators were unbiased, with respect to the highest-fidelity model. In contrast, many more evaluations of the lower fidelity models were used. The lower computational costs of these models and their high-correlation with each other and the highest-fidelity model were effectively exploited to significantly reduce the MSE of the ACV estimator relative to the SFMC estimator.

We constructed our final estimator of the mean, Q^{μ}_{ACV}, and variance, $Q^{\sigma^2}_{\text{ACV}}$, of the mass change by evaluating each model at the number of samples determined by Figure 12. All models were evaluated at the same two samples, the two low-fidelity models were both evaluated at another 351 samples, and the $\text{SSA}_{1.5km,365days}$ model was evaluated at another 10130 samples. The small number of samples allocated to the highest-fidelity model was due to the extremely high-correlation between that model and the model $\text{MOLHO}_{1km,36days}$. This high-correlation suggests that the temporal discretization error of the highest-fidelity model is smaller than the spatial discretization error.

Note, the exact number of samples allocated to each model that we reported is determined by the properties of the MFMC estimator chosen, however, if another estimator, for example MLMC, was chosen to use the same models the way samples are shared between models would likely change.[7]

The mean and standard deviation computed using the best ACV estimator were -639.06 ± 0.23 and 17.68 ± 6.67, respectively. It is clear that with our budget we were able to confidently estimate the expected mass change at year 2100. However,

[7] 37 of the 10130 $\text{SSA}_{1.5km,365days}$ model simulations failed so an additional 37 simulations at new random realizations of the friction field were run.

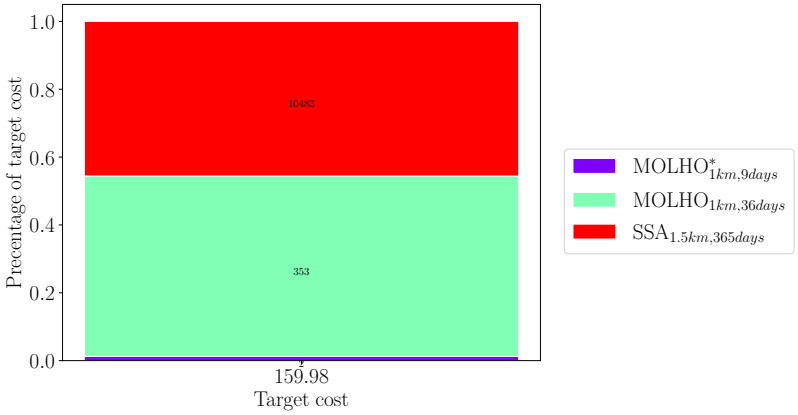


Figure 12. The optimal number of samples (number inside rectangles), computed using Eq. (29), required by the best ACV estimator to simulate each model.

our estimates of the standard deviation in the mass change were less precise. We could improve the precision of both estimated statistics by further increasing the exploitation budget; however, we choose not to do so, as our results emphasize that estimating statistics high-order statistics, such as variance, is more computationally demanding than estimating a mean. Moreover, the precision requirements of a UQ study should be determined by the stakeholders, which will use the uncertainty estimates to make decisions.

The left panel of Figure 13 plots the time evolution of mass loss predicted by the three models selected by our final ACV estimator. The right panel plots the distribution of mass loss at the final year, 2100, computed using the $\text{SSA}_{1.5km,365days}$ model. The bias of the $\text{SSA}_{1.5km,365days}$ is clear in both plots, for example, in the right panel the mean of the blue distribution is not close to the mean computed by the ACV estimator. However, we must emphasize that, by construction, the ACV estimate of the mean mass loss, and its variance, is unbiased with respect to the highest-fidelity model $\text{MOLHO}_{1km,9days}$. We also point out that while our Laplace approximation of the posterior is a Gaussian, the push-forward of this distribution through the $\text{SSA}_{1.5km,365days}$ model model is nonlinear. Specifically, the push-forward of a Gaussian through a linear model remains Gaussian; however, in this case, the right tail of the push-forward density is longer than the left tail, indicating that it is not Gaussian. This suggests that the mapping from the basal friction parameters to the quantity of interest is nonlinear. We were unable to compute reasonable push-forward densities with the simulations obtained from the other two models used to construct the ACV estimator due to an insufficient number of simulations. However, we believe it is reasonable to assume that the parameter-to-QoI map if these models is also non-linear.

6 Discussion

The cost of constructing our final estimator was equal to the pilot cost and the exploitation cost, totalling $197.13 + (160 \times 4.18)$ hours, or approximately 36 days. Additionally, the median variance reduction obtained by the bootstrapped estimators

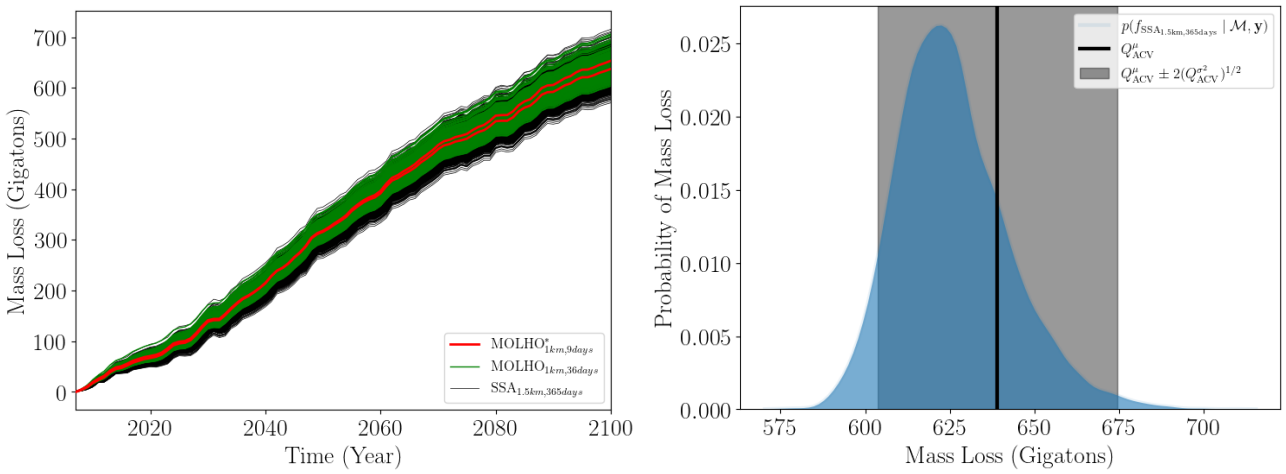


Figure 13. (*Left*) The evolution of mass loss predicted by the three models we used in our final ACV estimator, corresponding to each of the simulations used to construct the estimator. (*Right*) The probability of mass loss computed using the $\mathrm{SSA}_{1.5km,365days}$ model. The black vertical line represents the ACV estimate of the mean, while the gray shaded region represents plus and minus 2 standard deviations, again computed by the ACV estimator.

was $\mathbb{V}_\Theta\left[Q_0^\mu\right]/\mathbb{V}_\Theta\left[Q^\mu_{\mathrm{ACV}}\right] = 38.24$ for estimating the mean and $\mathbb{V}_\Theta\left[Q_0^\mu\right]/\mathbb{V}_\Theta\left[Q^{\sigma^2}_{\mathrm{ACV}}\right] = 28.91$ for estimating the variance of the mass change. Achieving the same precision with SFMC estimators using only the highest fidelity model would require approximately $28.91 \times 160 \times 4.18$ hours $= 805$ days. This calculation used the smallest variance reduction motivated by the observation that high-fidelity simulation data can be used to compute both mean and variance. Thus, MFSE reduced the cost of estimating uncertainty from over two and a half years of CPU time to just over a month, assuming the models are evaluated in serial. Note that while applying MFSE to the Humboldt Glacier took over a month of serial computations, the clock time needed for MFSE can be substantially reduced because MFSE is embarrassingly parallel. Each simulation run in the pilot stage can be run in parallel without communication between. Similarly, for the exploitation phase. Moreover, each simulation ca can be computed in parallel. Consequently, while using MFSE for continental scale UQ studies may require years of serial CPU time, distributed computing could substantially reduce this cost, potentially one to two orders of magnitude. The exact reduction would depend on the number of CPUs used and the scalability of the computational models.

While the highest-fidelity model, MOLHO, was capable of capturing ice-sheet dynamics that the SSA model was not – that is vertical changes in the horizontal velocities (Figure 14 shows the different ice-thickness predicted at the final time by the MOLHO and SSA model) – the best ACV estimator was still able to use the simplified physics of SSA to reduce the MSE of the best ACV estimator. Moreover, the best ACV estimator also used evaluations of the SSA model on a coarse mesh, which failed to resolve all the local features of the friction and ice-sheet flow-field (see Figure 14) and did not conserve mass, unlike the highest fidelity model. This result demonstrates that, provided there is high correlation between the model predictions of a QoI, MFSE can be effective when there is high correlation between the model predictions of a QoI, even when the model states

vary differently across time and space for a single realization of the random model parameters. Moreover, future MFSE studies may benefit from not only using low-fidelity models derived from different physics assumptions and numerical discretizations but also those based on data driven models, such as machine learning operators (He et al., 2023; Lowery et al., 2024) or adjoint-based linearizations (Recinos et al., 2023). However, if such models are used, the computational cost of constructing them must also be considered (Peherstorfer, 2019), just as we accounted for the pilot cost in this study.

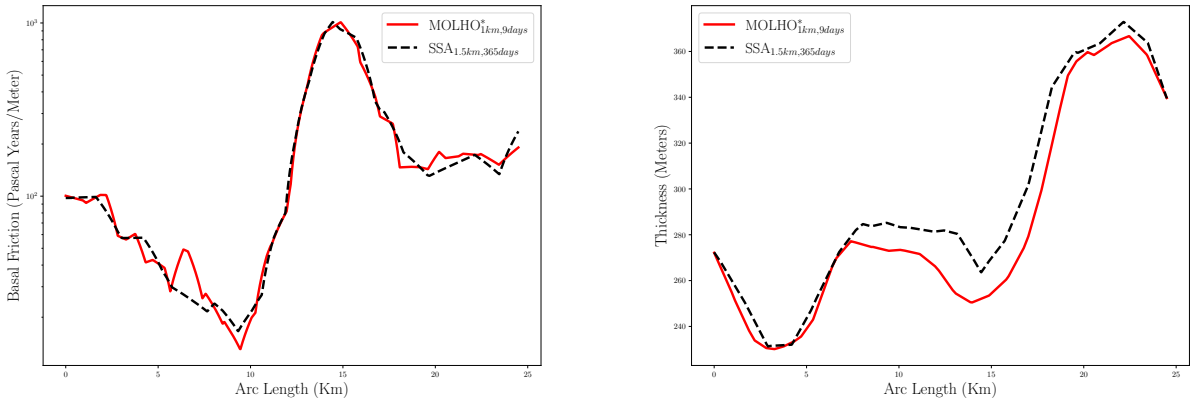


Figure 14. (*Left*) The basal friction, β, along the cross section (black line) depicted in the right panel of Figure 7. (*Right*) The difference between the thickness fields simulated by the $\mathrm{MOLHO}^*_{1km,9days}$ model and the $\mathrm{SSA}_{1.5km,365days}$ along the same cross section.

Our study used a high-dimensional representation of the basal friction field capable of capturing high-frequency modes. However, previous studies have commonly used lower-dimensional parameterizations (Nias et al., 2023; Ritz et al., 2015; Schlegel et al., 2018; Jantre et al., 2024). Consequently, we investigated the impact of using a low-frequency, low-dimensional representation of the friction field on the efficiency of ACV estimators applied to ice-sheet models. Specifically, we estimated the mean and variance of the mass change using a 10-dimensional Karhunen Loeve expansion (KLE) to represent the posterior uncertainty of the basal friction field (complete details are presented in Appendix B). We found that using the low-dimensional KLE smoothed realizations of the basal friction, which in turn drastically improved the variance reduction of MFSE to over a factor of 200. However, only using 10 modes to represent the basal friction caused the variance of the mass change to be substantially underestimated. (Recinos et al., 2023) also demonstrated that lower-dimensional parameterizations can result in misleading estimates. Consequently, while low-dimensional representations of friction enable faster UQ, the results may be misleading. Thus, future research is needed to balance the increased bias introduced by the low-dimensional parameterization with the improved variance reduction properties of an ACV estimator.

This study emphasizes that the relative effectiveness of ACV estmators – such as MLMC, MFMC, and ACVMF – is problem dependent. Although each MFSE algorithm in the literature has its own theoretical advantages and disadvantages, it is often difficult to determine which will be the most effective at the onset of a study. Indeed, several types of estimators enu-

merated by this study yielded estimates of the mean and variance of the mass change with similar precision. For example, Figures 11a, 11b, 11c show that while using three models is clearly better than using two, there is little, if any, marginal benefit in moving from three to four models, as indicated by the size of the box plots. Moreover, it is difficult to determine *a priori* the numerical discretizations and model physics needed by a model ensemble to produce an ACV estimator with the smallest MSE. Consequently, we used a small pilot sample to compute the correlation between model outputs and then use the analytical properties of ACV estimators to predict the MSE of each estimator produced by popular MFSE algorithms.

While pilot studies are required for ACV methods, our results suggest that using a small number of pilot samples can introduce non-trivial variability into the optimal sample allocation used by ACV estimators. Consequently, we introduced a novel two-step bootstrapping procedure to quantify the impact of a small number of pilot samples. While our two step procedure was able to down select from a large set of possible models, further research is needed to develop algorithms that can efficiently conduct pilot studies involving a large numbers of models. Furthermore, it is essential that new algorithms balance the computational cost of computing the correlation between models with the impact the error in the estimated correlations when determining the optimal MSE of an ACV estimator.

Our study predicted the mean and standard deviation of mass change (in Gigatons) from Humboldt Glacier to be -639.06 and 17.68 respectively. However, the exact values of these statistics were impacted by our modeling choices. First, we only quantified uncertainty due to unknown basal friction which ignores other contributions to mass-loss variability arising from uncertain climate and ice-sheet processes such as iceberg calving, subglacial hydrology, and submarine melting. Including these processes would have likely affected both the mean and variance of the mass change. Indeed, our predicted mass loss is significantly less than in two recent studies of Humboldt Glacier (Hillebrand et al., 2022; Carr et al., 2024) due to our use of a low-emissions climate scenario and our neglect of ocean forcing. Moreover, introducing more complicated physics in the highest-fidelity model, such as calving, could degrade the performance of MFSE. For example, ice melt at the boundary can induce strong dynamical responses in a marine-terminating glacier, which could potentially reduce the correlation between models that do not capture this phenomenon. However, despite our imperfect description of uncertainty, we believe our study reflects the challenges of a more comprehensive study while still facilitating a computationally feasible investigation of MFSE methods.

This study focused on investigating the efficacy of using MFSE to accelerate the quantification of parametric uncertainty using deterministic ice-sheet models. We did not quantify the uncertainty arising from model inadequacy. Recently Verjans et al. (2022), attempted to quantify model uncertainty by developing stochastic ice-sheet models designed to simulate the impact of glaciological processes that exhibit variability that cannot be captured by the spatiotemporal resolution typically employed by ice-sheet models, such as calving and subglacial hydrology. The MFSE algorithms presented in this paper can be applied to such stochastic models, by sampling the model parameters and treating the stochasticity of model as noise. However, the noise typically reduces the correlation between models and thus the efficiency of MFSE (Reuter et al., 2024). Moreover, this study only focused on estimating the mean and variance of mass change. Consequently, the efficacy of MFSE may change when estimating statistics – such as probability of failure, entropic risk, and average value at risk (Rockafellar and Uryasev, 2013; Jakeman et al., 2022) – to quantify the impact of rare instabilities and feedback mechanisms in the system. We anticipate

that larger number of pilot samples than the amount used in this study will be needed to estimate such tail statistics, potentially reducing the efficiency of MFSE.

Many recent studies have conducted formal uncertainty quantification of projections of ice-sheet change considering numerous sources of uncertainty, such as climate forcing, iceberg calving, basal friction parameters, and ice viscosity. Although, these generally deal with scalar parameters, such as a single calving threshold stress (Aschwanden and Brinkerhoff, 2022; Jantre et al., 2024) or scalar adjustment factors to basal friction and ice viscosity fields (Nias et al., 2023; Felikson et al., 2023; Jantre et al., 2024). However, recently automatic differentiation was used to linearize the parameter-to-QoI map of an SSA model to facilitate computationally efficiently quantify the uncertainty caused by high-dimensional parameterizations of basal friction and ice-stiffness (Recinos et al., 2023), Additionally, other UQ studies have primarily relied on a large number of simulations from a single low fidelity model (e.g., Nias et al., 2019; Bevan et al., 2023), sometimes with informal validation using a small number of higher-fidelity simulations (e.g., Nias et al., 2023), or on the construction of surrogate models to sufficiently sample the parameter space (e.g., Bulthuis et al., 2019; Berdahl et al., 2021; DeConto et al., 2021; Hill et al., 2021; Aschwanden and Brinkerhoff, 2022; Jantre et al., 2024). Furthermore, another set of studies quantified the uncertainty associated with the use of many different numerical models — termed an "ensemble of opportunity" — which includes a wide range of modeling choices that sample parameter values and model fidelity in an unsystematic manner (Edwards et al., 2021; Seroussi et al., 2023; Van Katwyk et al., 2023; Yoo et al., 2024). While this study is limited in scope, because it focus on solely estimating parametric uncertainty induced by basal friction variability, our results demonstrate that even when low-fidelity ice-sheet models do not capture the flow features predicted by higher-fidelity models, they can still be effectively utilized by MFSE methods to reduce the cost of quantifying high-dimensional parametric uncertainty in ice-sheet model predictions. Consequently, low-fidelity models, when used with MFSE methods, may be able to substantially reduce the computational cost of future efforts to quantify uncertainty in the projection of the mass change from the entire Greenland and Antarctic ice sheets.

7 Conclusions

Mass loss from ice sheets is anticipated to contribute $O(10)$ cm to sea-level rise in the next century under all but the lowest emission scenarios (Edwards et al., 2021). However, projections of sea-level rise due to ice-sheet mass change are inherently uncertain, and quantifying the impact of this uncertainty is essential for making these projections useful to policy makers and planners. Unfortunately, accurately estimating uncertainty is challenging because it requires numerous simulations of a computationally expensive numerical model. Consequently, we evaluated the efficacy of MFSE for reducing the computational cost of quantifying uncertainty in projections of mass loss from Humboldt Glacier, Greenland.

This study used MFSE to estimate the mean and the variance of uncertain mass-change projections caused by uncertainty in glacier basal friction using 13 different models of varying computational cost and fidelity. While ice sheets are subject to other sources of uncertainty, focus was given to basal friction because its inherent high-dimensionality typically make quantifying its impact on the uncertainty in model predictions challenging. Yet, despite this challenge, we found that for a fixed computational

budget, MFSE was able to reduce the MSE in our estimates of the mean and variance of the mass change by over an order-of-magnitude compared to a SFMC based approach that just used simulations from the highest fidelity model.

In out study, we were able to use MFSE to substantially reduce the MSE error in the statistics by exploiting the correlation between the predictions of the mass change produced by each model. However, it was not necessary to using simulations from all of the models to reduce the MSE. Indeed, the MFSE algorithm determined that only three models (including the highest-fidelity model) were needed to minimize the MSE in the statistics given our computational budget. The low-fidelity models selected used: 1) simplifications of the high-fidelity model physics, 2) were solved on coarser resolution spatial and temporal meshes, and 3) were solved without the requirement of mass conservation. These simplifications result in significant computational cost savings relative to use of the high-fidelity model alone. This result demonstrated that MFSE can be effective even when the lower-fidelity models are incapable of capturing the local features of the ice flow fields predicted by the high-fidelity model. Moreover, while the utility of the lower-fidelity models ultimately chosen for MFSE were not clear at the onset of the study, we were still able to estimate uncertainty at a fraction of the cost of single fidelity MC. This was achieved despite the need to conduct a pilot study that evaluated all models a small number of times.

Finally, this study demonstrated that MFSE can be used to reduce the computational cost of quantifying parametric uncertainty in projections of a single glacier, which suggests that MFSE could plausibly be used for continental-scale studies of ice-sheet evolution in Greenland and Antarctica. Future research should increase the complexity of this study in two directions. First, future studies should include additional sources of ice-sheet uncertainty beyond the basal friction field studied here, for example uncertain surface mass balance and ocean forcing. Second, future studies should include the use of model fidelities that capture additional physical processes such as calving, fracture, and ocean-forced melting. Consequently, while our findings should be interpreted with caution given the aforementioned limitations they encourage future studies to utilize MFSE for reducing the cost of computing probabilistic projections of sea-level rise due to ice-sheet mass change.

Code availability. The code used to construct ACV estimators has been released in the open-source Python package PyApprox https://github.com/sandialabs/pyapprox.

Appendix A: Low-rank Laplace approximation

Following Bui-Thanh et al. (2013); Isaac et al. (2015) we computed the covariance of the Laplace approximation of the posterior distribution of the friction parameters, Eq. (14), using

$$\boldsymbol{\Sigma}_{\text{post}} = \left(\mathbf{H}_{\text{MAP}} + \boldsymbol{\Sigma}_{\text{prior}}^{-1}\right)^{-1} = \mathbf{L}\left(\mathbf{L}^{\top}\mathbf{H}_{\text{MAP}}\mathbf{L} + \mathbf{I}\right)^{-1}\mathbf{L}^{\top},$$

where $\mathbf{H}_{\text{MAP}}$ is the Hessian of $\frac{1}{2}(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{\theta}))^{\top}\boldsymbol{\Sigma}_{\text{noise}}^{-1}(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{\theta}))$ at $\boldsymbol{\theta}=\boldsymbol{\theta}_{\text{MAP}}$, $\mathbf{L}=\mathbf{K}^{-1}\mathbf{M}^{\frac{1}{2}}$ and the entries of $\boldsymbol{K}$ and $\boldsymbol{M}$ are defined in Eq. (10) and Eq. (11), respectively.

Drawing samples from this Gaussian posterior is computationally challenging because the posterior covariance $\boldsymbol{\Sigma}_{\text{post}}$ depends on the Hessian $\mathbf{H}_{\text{MAP}}$ which is a high-dimensional dense matrix. Consequently, following Bui-Thanh et al. (2013) and

Isaac et al. (2015) we constructed a low-rank approximation of the prior-preconditioned Hessian $\mathbf{L}^\top \mathbf{H}_{\text{MAP}} \mathbf{L}$ using matrix-free randomized methods that requires only multiplications of the Hessian with random vectors. Specifically, computing a spectral decomposition of

$$\mathbf{L}^\top \mathbf{H}_{\text{MAP}} \mathbf{L} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top, \tag{A1}$$

with $\mathbf{U}$ orthogonal and $\mathbf{\Lambda}$ diagonal matrices and noting

$$\mathbf{\Sigma}_{\text{post}} = \mathbf{L} \left(\mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top + \mathbf{I} \right)^{-1} \mathbf{L}^\top = \mathbf{L} \left(\mathbf{U} (\mathbf{\Lambda} + \mathbf{I}) \mathbf{U}^\top \right)^{-1} \mathbf{L}^\top = \mathbf{L} \mathbf{U} (\mathbf{\Lambda} + \mathbf{I})^{-1} \mathbf{U}^\top \mathbf{L}^\top$$

we factorized $\mathbf{\Sigma}_{\text{post}}$ as

$$\mathbf{\Sigma}_{\text{post}} = \mathbf{T} \mathbf{T}^\top, \qquad \mathbf{T} = \mathbf{L} \mathbf{U} (\mathbf{\Lambda} + \mathbf{I})^{-\frac{1}{2}} \mathbf{U}^\top = \mathbf{L} \mathbf{U} \left((\mathbf{\Lambda} + \mathbf{I})^{-\frac{1}{2}} - \mathbf{I} \right) \mathbf{U}^\top + \mathbf{L}.$$

In order to perform a low-rank approximation of the matrix $\mathbf{T}$ we truncated the spectral decomposition of $\mathbf{W} = \mathbf{U} \left((\mathbf{\Lambda} + \mathbf{I})^{-\frac{1}{2}} - \mathbf{I} \right) \mathbf{U}^\top$ by discarding the eigenvalues λ_i such that $\left| 1 - \frac{1}{\sqrt{\lambda_i + 1}} \right| \ll 1$. This ensured that the low-rank approximation of $\mathbf{T}$ well approximated $\mathbf{T}$ in the spectral norm sense. The eigenvalues and two eigenvectors of the spectral decomposition we computed are depicted in Figure A1.

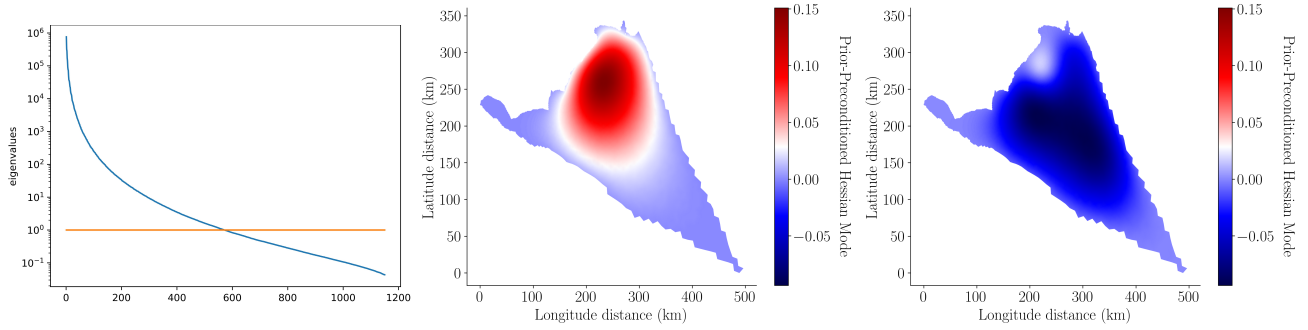


Figure A1. *(Left)* Eigenvalues λ_i of the prior-preconditioned Hessian, computed by solving Eq. (A1), and *(Center)* its eigenvectors associated to the largest and *(Right)* third largest eigenvalue. Note that similarly to Isaac et al. (2015), we plot the eigenvectors $\mathbf{V}_i = \mathbf{L}\mathbf{U}_i$ that are orthonormal with respect to the prior-induced dot-product, that is, $\mathbf{V}_i^\top \mathbf{\Sigma}_{\text{prior}}^{-1} \mathbf{V}_j = \delta_{ij}$.

We computed the truncated spectral decomposition using randomized algorithms (see Hartland et al. (2023); Halko et al. (2011)) implemented in PyAlbany, see Liegeois et al. (2023). The algorithms used were matrix-free and only required the multiplication of $\mathbf{L}^\top \mathbf{H}_{\text{MAP}} \mathbf{L}$ with vectors. Moreover, as described in Hartland et al. (2023); Isaac et al. (2015), the multiplication of the Hessian with a vector required solving two adjoint systems of the flow model. Similarly, the multiplication of the matrix $\mathbf{L}$ with a vector required the solution of the two-dimensional linear elliptic system with matrix $\boldsymbol{K}$, defined in Eq. (10). Consequently, we were able to efficiently draw samples from the posterior distribution of the friction parameters using

$$\boldsymbol{\theta}_{\text{post}} = \boldsymbol{\theta}_{\text{MAP}} + \mathbf{T} \boldsymbol{n}, \quad \boldsymbol{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Appendix B: Low-dimensional representation of basal friction using a Karhunen-Loeve expansion

In our main study we found that when using a high-dimensional representation of the uncertainty in the basal friction field, bootstrapped ACV estimators rarely chose to use models that had coarse spatial meshes relative to the mesh used by the high-fidelity model. This was likely due to the fact that our high-dimensional representation of the friction uncertainty was constructed on the high-fidelity mesh and interpolated onto coarser meshes. To verify this hypothesis we investigated using a lower-dimensional representation of the friction field based on a Karhunen Loeve expansion (KLE) of the friction field that smoothed out the high-frequency variations in the posterior samples of the friction field we used in our main study.

Construction of the KLE

In our investigations we used a KLE

$$\boldsymbol{\theta} = \boldsymbol{\theta}_{MAP} + \sum_{i=1}^{D} \sqrt{\lambda_i}\psi_i \eta_i, \qquad \eta_i \sim \mathcal{N}(0,1) \tag{B1}$$

to provide a low-dimension representation of the Laplace approximation of the posterior of the log basal friction field. We computed the eigenvalues λ_i and the orthonormal eigenvectors ψ_i by solving the eigenvalue problem

$$\boldsymbol{\Sigma}_{\text{post}}\psi_i = \lambda_i \psi_i, \tag{B2}$$

using the randomized matrix-free methods Hartland et al. (2023); Halko et al. (2011).

While a KLE basis could have been constructed on any of the four meshes we considered, in this study we solved the discretized eigenvalue problem using the finest mesh. The 1st, 2nd, and 10th modes of the KLE used in this study are depicted in Figure B1. Note, unlike what is typically seen when constructing a KLE of a field with a pointwise variance that is constant across the domain, the low-frequency KLE modes constructed here are localized where the posterior uncertainty is highest. The finite element basis on the finest mesh was then used to interpolate the KLE basis from the fine mesh onto the coarser meshes. This procedure ensured that varying the coefficients of the KLE basis (the random parameters to the model) would affect each model similarly regardless of the mesh discretization employed. Similarly to the KLE basis, the mean of the log KLE field (taken to be the mean of the Laplace approximation) was computed on the finest mesh.

Figure B2 compares a realization of the log of the basal friction perturbation (mean zero) drawn from the Laplace approximation of the posterior and a random random realization of the log of the basal friction perturbation computed using the KLE. It is clear that the KLE smooths out much of the high-frequency content present in the realization drawn from the Laplace approximation of the posterior.

Pilot study

In this section, we detail the pilot study we undertook to investigate the impact of using a low-dimensional KLE to represent friction when using MFSE to estimate statistics of mass change. We did not move beyond the pilot study to compute the values of the statistics to limit the computational cost of this supplementary study.

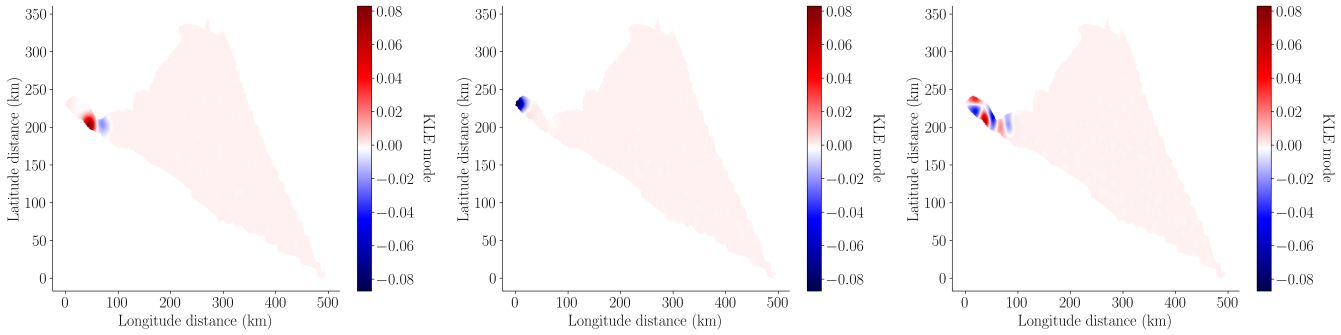


Figure B1. From left to right, the 1st, 2nd, and 10th mode of the KLE, Eq.(B1), used in this study, computed using Eq.(B2).

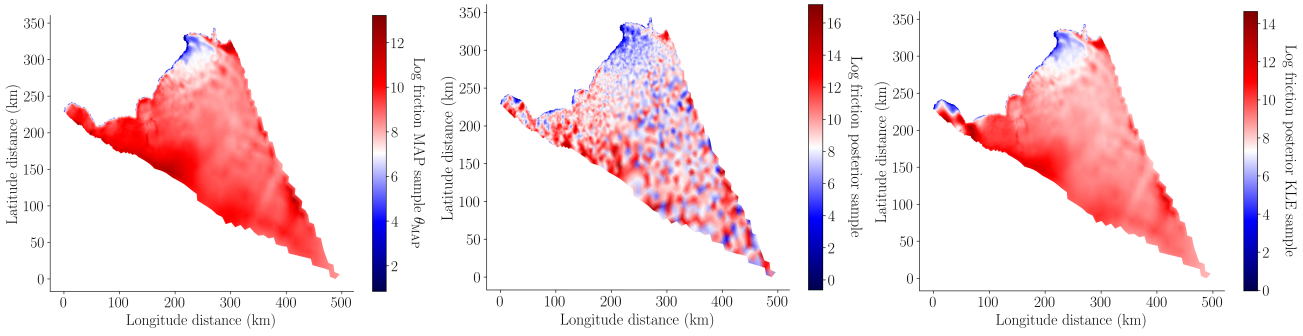


Figure B2. (*Left*) The mean of the log of the basal friction, $\boldsymbol{\theta}_{\text{MAP}}$ in Eq. (13). (*Center*) A random realization of the log of the basal friction drawn from the Laplace approximation of the posterior $p(\boldsymbol{\theta} \mid \mathcal{M}, \boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\theta}_{\text{MAP}}, \boldsymbol{\Sigma}_{\text{post}})$. (*Right*) A random realization of the log of the 10-dimensional basal friction computed using the KLE approximation, Eq. (B1), of the posterior.

First, we evaluated each of our 13 models at 20 random pilot samples of the KLE. Second we computed the pilot statistics needed to find the best ACV estimator. Third we bootstrapped the pilot samples to estimate the median and confidence intervals on the variance reduction obtained by the best ACV estimator.

The mean and variance bootstrapped variance reduction are depicted in Figure B4. The variance reductions reported are almost an order of magnitude larger than those reported for MFSE based on the Laplace approximation of the posterior. This improved performance is because correlations between the models (Figure B3) are significantly higher than the correlations obtained when sampling from the Laplace approximation of the posterior (Figure 9). However, the KLE representation underestimates the uncertainty in the predicted mass change at 2100. Specifically, the standard deviation of the mass change computed using 20 pilot samples of the highest-fidelity model using the Laplace approximation of the posterior is significantly higher than the standard deviation computed using the KLE. Please refer to Figure B5.

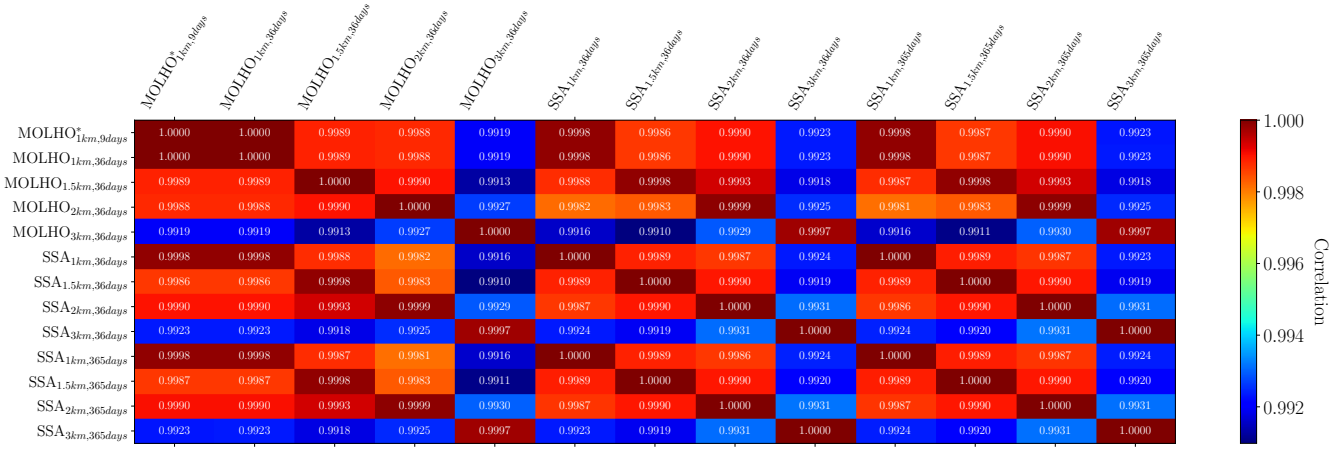


Figure B3. The correlations, $\mathbb{C}\mathrm{orr}_\pi[\boldsymbol{f}, \boldsymbol{f}]$ with $\boldsymbol{f} = [f_0, \ldots, f_M]^\top$, between the 13 ice-sheet models considered by this study using 20 pilot samples of the KLE using Eq.(B1).

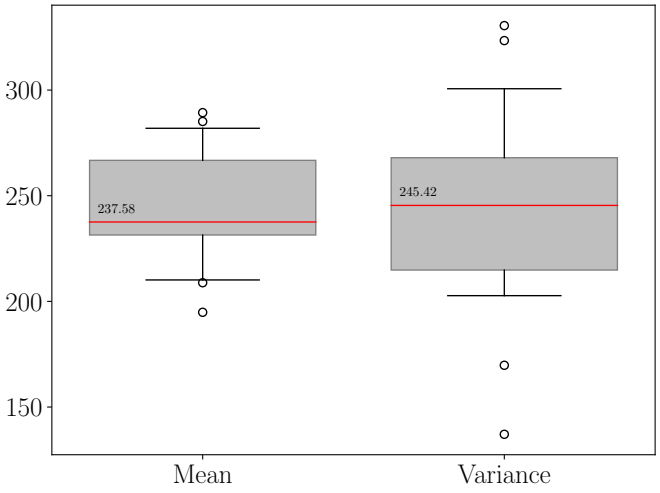


Figure B4. The predicted variance reductions $\mathcal{R}_\Theta[Q^\mu_{\mathrm{ACV}}]$ (mean) and $\mathcal{R}_\Theta[Q^{\sigma^2}_{\mathrm{ACV}}]$ (variance), see Eq. (30), obtained using bootstrapping of the 20 pilot samples of the KLE, Eq.(B1). The red lines represent the median estimator variance reductions. The lower and upper whiskers represent the 10% and 90% quantiles.

Author contributions. John Jakeman was responsible for formulation of the overarching research goals and aims (conceptualization), data curation, application of the statistical techniques used to analyze the data (formal analysis), conducting the computer experiments (investigation), developing the methodology , implementing and maintaining the software used (software), oversight and leadership for the research planning and execution (supervision), writing the original draft. Mauro Perego was responsible for investigation, methodology, software and writing the original draft. Tom Seidl was responsible for data curation, investigation, software, and writing the original draft. Tucker Hartland

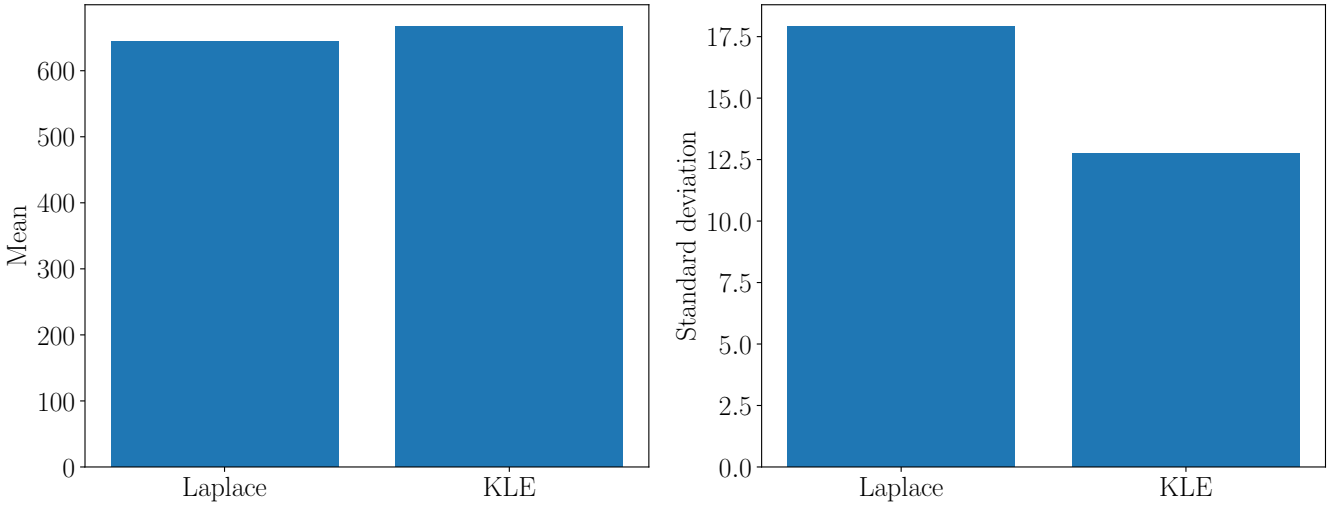


Figure B5. (*Left)* The mean, $Q_0^{\mu}(\Theta_{\text{pilot}})$, and (*Right)* standard deviation, $\sqrt{Q_0^{\sigma^2}(\Theta_{\text{pilot}})}$, computed using 20 pilot samples from the Laplace approximation of the posterior and the KLE.

was responsible for software and writing the original draft. Trevor Hillebrand was responsible for developing the data curation, methodology, and writing the original draft. Matthew Hoffman was responsible for developing the conceptualization, methodology, supervision, funding acquisition, and writing the original draft. Stephen Price was responsible for developing the conceptualization, funding acquisition, and writing the original draft.

Competing interests. The authors have no competing interests.

Acknowledgements. This work was was sponsored by the US Department of Energy's Office of Science Advanced Scientific Computing Research, Biological, and Environmental Research Scientific Discovery through Advanced Computing (SciDAC) programs.

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC (NTESS), a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration (DOE/NNSA) under contract DE-NA0003525. This written work is authored by an employee of NTESS. The employee, not NTESS, owns the right, title and interest in and to the written work and is responsible for its contents. Any subjective views or opinions that might be expressed in the written work do not necessarily represent the views of the U.S. Government. The publisher acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this written work or allow others to do so, for U.S. Government purposes. The DOE will provide public access to results of federally sponsored research in accordance with the DOE Public Access Plan.

The authors would like to thank Michael Eldred for his helpful discussions on how to practically assess the impact of small number of pilot samples on the predicted variance of ACV estimators. The authors would also like to thank Thomas Dixon and Alex Gorodetsky for their

useful discussions on constructing ACV estimators for multiple statistics. Finally, the authors would like to thank Kenneth Chad Sockwell for developing the FEniCS ice-sheet code.

The National Snow and Ice Data Center QGreenland package https://qgis.org/ was used to produce the image of Greenland in Figure 1.

References

Åkesson, H., Morlighem, M., O'Regan, M., and Jakobsson, M.: Future projections of Petermann Glacier under ocean warming depend strongly on friction law, Journal of Geophysical Research: Earth Surface, 126, e2020JF005 921, 2021.

Alnæs, M., Blechta, J., Hake, J., Johansson, A., Kehlet, B., Logg, A., Richardson, C., Ring, J., Rognes, M. E., and Wells, G. N.: The FEniCS project version 1.5, Archive of Numerical Software, 3, 2015.

Aschwanden, A. and Brinkerhoff, D. J.: Calibrated Mass Loss Predictions for the Greenland Ice Sheet, Geophysical Research Letters, 49, e2022GL099 058, https://doi.org/10.1029/2022GL099058, 2022.

Bakker, A. M. R., Wong, T. E., Ruckert, K. L., and Keller, K.: Sea-level projections representing the deeply uncertain contribution of the West Antarctic ice sheet, Scientific Reports, 7, 3880, https://doi.org/10.1038/s41598-017-04134-5, 2017.

Balay, S., Gropp, W., McInnes, L. C., and Smith, B. F.: PETSc, the portable, extensible toolkit for scientific computation, Argonne National Laboratory, 2, 1998.

Barnes, J. M., Dias dos Santos, T., Goldberg, D., Gudmundsson, G. H., Morlighem, M., and De Rydt, J.: The transferability of adjoint inversion products between different ice flow models, The Cryosphere, 15, 1975–2000, https://doi.org/10.5194/tc-15-1975-2021, 2021.

Berdahl, M., Leguy, G., Lipscomb, W. H., and Urban, N. M.: Statistical emulation of a perturbed basal melt ensemble of an ice sheet model to better quantify Antarctic sea level rise uncertainties, The Cryosphere, 15, 2683–2699, https://doi.org/10.5194/tc-15-2683-2021, 2021.

Bevan, S., Cornford, S., Gilbert, L., Otosaka, I., Martin, D., and Surawy-Stepney, T.: Amundsen Sea Embayment Ice-Sheet Mass-Loss Predictions to 2050 Calibrated Using Observations of Velocity and Elevation Change, Journal of Glaciology, pp. 1–11, https://doi.org/10.1017/jog.2023.57, 2023.

Bochev, P., Ridzal, D., D'Elia, M., Perego, M., and Peterson, K.: Optimization-based, property-preserving finite element methods for scalar advection equations and their connection to Algebraic Flux Correction, Computer Methods in Applied Mechanics and Engineering, 367, 112 982, https://doi.org/https://doi.org/10.1016/j.cma.2020.112982, 2020.

Bomarito, G., Leser, P., Warner, J., and Leser, W.: On the optimization of approximate control variates with parametrically defined estimators, Journal of Computational Physics, 451, 110 882, https://doi.org/10.1016/j.jcp.2021.110882, 2022.

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

Brinkerhoff, D. J.: Variational inference at glacier scale, Journal of Computational Physics, 459, 111 095, https://doi.org/https://doi.org/10.1016/j.jcp.2022.111095, 2022.

Brondex, J., Gillet-Chaulet, F., and Gagliardini, O.: Sensitivity of centennial mass loss projections of the Amundsen basin to the friction law, The Cryosphere, 13, 177–195, 2019.

Bui-Thanh, T., Ghattas, O., Martin, J., and Stadler, G.: A Computational Framework for Infinite-Dimensional Bayesian Inverse Problems Part I: The Linearized Case, with Application to Global Seismic Inversion, SIAM Journal on Scientific Computing, 35, A2494–A2523, https://doi.org/10.1137/12089586X, 2013.

Bulthuis, K., Arnst, M., Sun, S., and Pattyn, F.: Uncertainty quantification of the multi-centennial response of the Antarctic ice sheet to climate change, The Cryosphere, 13, 1349–1380, https://doi.org/10.5194/tc-13-1349-2019, 2019.

Carr, J. R., Hill, E. A., and Gudmundsson, G. H.: Sensitivity to forecast surface mass balance outweighs sensitivity to basal sliding descriptions for 21st century mass loss from three major Greenland outlet glaciers, The Cryosphere, 18, 2719–2737, 2024.

Cornford, S. L., Martin, D. F., Graves, D. T., Ranken, D. F., Le Brocq, A. M., Gladstone, R. M., Payne, A. J., Ng, E. G., and Lipscomb, W. H.: Adaptive mesh, finite volume modeling of marine ice sheets, Journal of Computational Physics, 232, 529–549, https://doi.org/https://doi.org/10.1016/j.jcp.2012.08.037, 2013.

Cuffey, K. and Paterson, W.: The Physics of Glaciers, Butterworth-Heinneman, Amsterdam, 4th edn., 2010.

DeConto, R. M., Pollard, D., Alley, R. B., Velicogna, I., Gasson, E., Gomez, N., Sadai, S., Condron, A., Gilford, D. M., Ashe, E. L., et al.: The Paris Climate Agreement and future sea-level rise from Antarctica, Nature, 593, 83–89, 2021.

Dias dos Santos, T., Morlighem, M., and Brinkerhoff, D.: A new vertically integrated MOno-Layer Higher-Order (MOLHO) ice flow model, The Cryosphere, 16, 179–195, https://doi.org/10.5194/tc-16-179-2022, 2022.

Dixon, T., Warner, J., Bomarito, G., and Gorodetsky, A.: Multi-Fidelity Estimation for Multi-Output Systems using Approximate Control Variates, ARXIV, 2023.

Dukowicz, J. K., Price, S. F., and Lipscomb, W. H.: Consistent approximations and boundary conditions for ice-sheet dynamics from a principle of least action, Journal of Glaciology, 56, 480–496, https://doi.org/10.3189/002214310792447851, 2010.

Durand, G., Gagliardini, O., Zwinger, T., Meur, E. L., and Hindmarsh, R. C.: Full Stokes modeling of marine ice sheets: influence of the grid size, Annals of Glaciology, 50, 109–114, https://doi.org/10.3189/172756409789624283, 2009.

Edwards, T. L., Brandon, M. A., Durand, G., Edwards, N. R., Golledge, N. R., Holden, P. B., Nias, I. J., Payne, A. J., Ritz, C., and Wernecke, A.: Revisiting Antarctic ice loss due to marine ice-cliff instability, Nature, 566, 58–64, https://doi.org/10.1038/s41586-019-0901-4, 2019.

Edwards, T. L., Nowicki, S., Marzeion, B., Hock, R., Goelzer, H., Seroussi, H., Jourdain, N. C., Slater, D. A., Turner, F. E., Smith, C. J., McKenna, C. M., Simon, E., Abe-Ouchi, A., Gregory, J. M., Larour, E., Lipscomb, W. H., Payne, A. J., Shepherd, A., Agosta, C., Alexander, P., Albrecht, T., Anderson, B., Asay-Davis, X., Aschwanden, A., Barthel, A., Bliss, A., Calov, R., Chambers, C., Champollion, N., Choi, Y., Cullather, R., Cuzzone, J., Dumas, C., Felikson, D., Fettweis, X., Fujita, K., Galton-Fenzi, B. K., Gladstone, R., Golledge, N. R., Greve, R., Hattermann, T., Hoffman, M. J., Humbert, A., Huss, M., Huybrechts, P., Immerzeel, W., Kleiner, T., Kraaijenbrink, P., Le clec'h, S., Lee, V., Leguy, G. R., Little, C. M., Lowry, D. P., Malles, J.-H., Martin, D. F., Maussion, F., Morlighem, M., O'Neill, J. F., Nias, I., Pattyn, F., Pelle, T., Price, S. F., Quiquet, A., Radić, V., Reese, R., Rounce, D. R., Rückamp, M., Sakai, A., Shafer, C., Schlegel, N.-J., Shannon, S., Smith, R. S., Straneo, F., Sun, S., Tarasov, L., Trusel, L. D., Van Breedam, J., van de Wal, R., van den Broeke, M., Winkelmann, R., Zekollari, H., Zhao, C., Zhang, T., and Zwinger, T.: Projected land ice contributions to twenty-first-century sea level rise, Nature, 593, 74–82, https://doi.org/10.1038/s41586-021-03302-y, 2021.

Felikson, D., Nowicki, S., Nias, I., Csatho, B., Schenk, A., Croteau, M. J., and Loomis, B.: Choice of observation type affects Bayesian calibration of Greenland Ice Sheet model simulations, The Cryosphere, 17, 4661–4673, https://doi.org/10.5194/tc-17-4661-2023, 2023.

Giles, M. B.: Multilevel Monte Carlo methods, Acta Numerica, 24, 259–328, https://doi.org/10.1017/S096249291500001X, 2015.

Goelzer, H., Nowicki, S., Edwards, T., Beckley, M., Abe-Ouchi, A., Aschwanden, A., Calov, R., Gagliardini, O., Gillet-Chaulet, F., Golledge, N. R., Gregory, J., Greve, R., Humbert, A., Huybrechts, P., Kennedy, J. H., Larour, E., Lipscomb, W. H., Le clec'h, S., Lee, V., Morlighem, M., Pattyn, F., Payne, A. J., Rodehacke, C., Rückamp, M., Saito, F., Schlegel, N., Seroussi, H., Shepherd, A., Sun, S., van de Wal, R., and Ziemen, F. A.: Design and results of the ice sheet model initialisation experiments initMIP-Greenland: an ISMIP6 intercomparison, The Cryosphere, 12, 1433–1460, https://doi.org/10.5194/tc-12-1433-2018, 2018.

Goldberg, D. N., Heimbach, P., Joughin, I., and Smith, B.: Committed retreat of Smith, Pope, and Kohler Glaciers over the next 30 years inferred by transient model calibration, The Cryosphere, 9, 2429–2446, https://doi.org/10.5194/tc-9-2429-2015, 2015.

Gorodetsky, A., Geraci, G., Eldred, M., and Jakeman, J.: A generalized approximate control variate framework for multifidelity uncertainty quantification, Journal of Computational Physics, 408, 109 257, https://doi.org/10.1016/j.jcp.2020.109257, 2020.

Gruber, A., Gunzburger, M., Ju, L., Lan, R., and Wang, Z.: Multifidelity Monte Carlo Estimation for Efficient Uncertainty Quantification in Climate-Related Modeling, EGUsphere, 2022, 1–27, https://doi.org/10.5194/egusphere-2022-797, 2022.

Halko, N., Martinsson, P. G., and Tropp, J. A.: Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions, SIAM Review, 53, 217–288, https://doi.org/10.1137/090771806, 2011.

Hartland, T., Stadler, G., Perego, M., Liegeois, K., and Petra, N.: Hierarchical off-diagonal low-rank approximation of Hessians in inverse problems, with application to ice sheet model initialization, Inverse Problems, 39, 085 006, https://doi.org/10.1088/1361-6420/acd719, 2023.

He, Q., Perego, M., Howard, A. A., Karniadakis, G. E., and Stinis, P.: A hybrid deep neural operator/finite element method for ice-sheet modeling, Journal of Computational Physics, 492, 112 428, https://doi.org/https://doi.org/10.1016/j.jcp.2023.112428, 2023.

Hill, E. A., Rosier, S. H. R., Gudmundsson, G. H., and Collins, M.: Quantifying the Potential Future Contribution to Global Mean Sea Level from the Filchner–Ronne Basin, Antarctica, The Cryosphere, 15, 4675–4702, https://doi.org/10.5194/tc-15-4675-2021, 2021.

Hillebrand, T. R., Hoffman, M. J., Perego, M., Price, S. F., and Howat, I. M.: The contribution of Humboldt Glacier, northern Greenland, to sea-level rise through 2100 constrained by recent observations of speedup and retreat, The Cryosphere, 16, 4679–4700, https://doi.org/10.5194/tc-16-4679-2022, 2022.

Hoffman, M. D. and Gelman, A.: The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, Journal of Machine Learning Research, 15, 1593–1623, http://jmlr.org/papers/v15/hoffman14a.html, 2014.

Hoffman, M. J., Perego, M., Price, S. F., Lipscomb, W. H., Zhang, T., Jacobsen, D., Tezaur, I., Salinger, A. G., Tuminaro, R., and Bertagna, L.: MPAS-Albany Land Ice (MALI): a variable-resolution ice sheet model for Earth system modeling using Voronoi grids, Geoscientific Model Development, 11, 3747–3780, 2018.

Isaac, T., Petra, N., Stadler, G., and Ghattas, O.: 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, 296, 348–368, https://doi.org/https://doi.org/10.1016/j.jcp.2015.04.047, 2015.

Jakeman, J.: PyApprox: A software package for sensitivity analysis, Bayesian inference, optimal experimental design, and multi-fidelity uncertainty quantification and surrogate modeling, Environmental Modelling & Software, 170, 105 825, https://doi.org/https://doi.org/10.1016/j.envsoft.2023.105825, 2023.

Jakeman, J. D., Kouri, D. P., and Huerta, J. G.: Surrogate modeling for efficiently, accurately and conservatively estimating measures of risk, Reliability Engineering & System Safety, p. 108280, https://doi.org/https://doi.org/10.1016/j.ress.2021.108280, 2022.

Jantre, S., Hoffman, M. J., Urban, N. M., Hillebrand, T., Perego, M., Price, S., and Jakeman, J. D.: Probabilistic projections of the Amery Ice Shelf catchment, Antarctica, under high ice-shelf basal melt conditions, EGUsphere, 2024, 1–45, https://doi.org/10.5194/egusphere-2024-1677, 2024.

Johnson, A., Aschwanden, A., Albrecht, T., and Hock, R.: Range of 21st century ice mass changes in the Filchner-Ronne region of Antarctica, Journal of Glaciology, p. 1–11, https://doi.org/10.1017/jog.2023.10, 2023.

Joughin, I., Smith, B. E., and Schoof, C. G.: Regularized Coulomb friction laws for ice sheet sliding: Application to Pine Island Glacier, Antarctica, Geophysical research letters, 46, 4764–4771, 2019.

Jouvet, G.: Mechanical error estimators for shallow ice flow models, Journal of Fluid Mechanics, 807, 40–61, https://doi.org/10.1017/jfm.2016.593, 2016.

Jouvet, G., Cordonnier, G., Kim, B., Lüthi, M., Vieli, A., and Aschwanden, A.: Deep learning speeds up ice flow modelling by several orders of magnitude, Journal of Glaciology, p. 1–14, https://doi.org/10.1017/jog.2021.120, 2021.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in Combining Projections from Multiple Climate Models, Journal of Climate, 23, 2739 – 2758, https://doi.org/https://doi.org/10.1175/2009JCLI3361.1, 2010.

Koziol, C. P., Todd, J. A., Goldberg, D. N., and Maddison, J. R.: fenics_ice 1.0: a framework for quantifying initialization uncertainty for time-dependent ice sheet models, Geoscientific Model Development, 14, 5843–5861, https://doi.org/10.5194/gmd-14-5843-2021, 2021.

Liegeois, K., Perego, M., and Hartland, T.: PyAlbany: A Python interface to the C++ multiphysics solver Albany, Journal of Computational and Applied Mathematics, 425, 115 037, https://doi.org/https://doi.org/10.1016/j.cam.2022.115037, 2023.

Lowery, M., Turnage, J., Morrow, Z., Jakeman, J., Narayan, A., Zhe, S., and Shankar, V.: Kernel Neural Operators (KNOs) for Scalable, Memory-efficient, Geometrically-flexible Operator Learning, https://arxiv.org/abs/2407.00809, 2024.

MacAyeal, D. R.: A tutorial on the use of control methods in ice-sheet modeling, Journal of Glaciology, 39, 91–98, https://doi.org/10.3189/S0022143000015744, 1993.

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Pean, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., eds.: Ocean, cryosphere and sea level change., p. 1211–1362, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA,, https://doi.org/10.1017/9781009157896.011, 2021.

Morland, L. W. and Johnson, I. R.: Steady Motion of Ice Sheets, Journal of Glaciology, 25, 229–246, https://doi.org/10.3189/S0022143000010467, 1980.

Morlighem, M., Rignot, E., Seroussi, H., Larour, E., Ben Dhia, H., and Aubry, D.: Spatial patterns of basal drag inferred using control methods from a full-Stokes and simpler models for Pine Island Glacier, West Antarctica, Geophysical Research Letters, 37, https://doi.org/https://doi.org/10.1029/2010GL043853, 2010.

Nias, I., Cornford, S., and Payne, A.: New mass-conserving bedrock topography for Pine Island Glacier impacts simulated decadal rates of mass loss, Geophysical Research Letters, 45, 3173–3181, 2018.

Nias, I. J., Cornford, S. L., Edwards, T. L., Gourmelen, N., and Payne, A. J.: Assessing Uncertainty in the Dynamical Ice Response to Ocean Warming in the Amundsen Sea Embayment, West Antarctica, Geophysical Research Letters, 46, 11 253–11 260, https://doi.org/10.1029/2019GL084941, 2019.

Nias, I. J., Nowicki, S., Felikson, D., and Loomis, B.: Modeling the Greenland Ice Sheet's Committed Contribution to Sea Level During the 21st Century, Journal of Geophysical Research: Earth Surface, 128, e2022JF006 914, https://doi.org/https://doi.org/10.1029/2022JF006914, e2022JF006914 2022JF006914, 2023.

Nowicki, S., Goelzer, H., Seroussi, H., Payne, A. J., Lipscomb, W. H., Abe-Ouchi, A., Agosta, C., Alexander, P., Asay-Davis, X. S., Barthel, A., Bracegirdle, T. J., Cullather, R., Felikson, D., Fettweis, X., Gregory, J. M., Hattermann, T., Jourdain, N. C., Kuipers Munneke, P., Larour, E., Little, C. M., Morlighem, M., Nias, I., Shepherd, A., Simon, E., Slater, D., Smith, R. S., Straneo, F., Trusel, L. D., van den Broeke, M. R., and van de Wal, R.: Experimental protocol for sea level projections from ISMIP6 stand-alone ice sheet models, The Cryosphere, 14, 2331–2368, https://doi.org/10.5194/tc-14-2331-2020, 2020.

Pattyn, F.: A new three-dimensional higher-order thermomechanical ice sheet model: Basic sensitivity, ice stream development, and ice flow across subglacial lakes, Journal of Geophysical Research: Solid Earth, 108, https://doi.org/https://doi.org/10.1029/2002JB002329, 2003.

Peherstorfer, B.: Multifidelity Monte Carlo Estimation with Adaptive Low-Fidelity Models, SIAM/ASA Journal on Uncertainty Quantification, 7, 579–603, https://doi.org/10.1137/17M1159208, 2019.

Peherstorfer, B. and Willcox, K.: Data-driven operator inference for nonintrusive projection-based model reduction, Computer Methods in Applied Mechanics and Engineering, 306, 196–215, https://doi.org/10.1016/j.cma.2016.03.025, 2016.

Peherstorfer, B., Willcox, K., and Gunzburger, M.: Optimal Model Management for Multifidelity Monte Carlo Estimation, SIAM Journal on Scientific Computing, 38, A3163–A3194, https://doi.org/10.1137/15M1046472, 2016.

Perego, M., Price, S., and Stadler, G.: Optimal initial conditions for coupling ice sheet models to Earth system models, Journal of Geophysical Research: Earth Surface, 119, 1894–1917, https://doi.org/https://doi.org/10.1002/2014JF003181, 2014.

Petra, N., Zhu, H., Stadler, G., Hughes, T. J., and Ghattas, O.: An inexact Gauss-Newton method for inversion of basal sliding and rheology parameters in a nonlinear Stokes ice sheet model, Journal of Glaciology, 58, 889–903, https://doi.org/10.3189/2012JoG11J182, 2012.

Recinos, B., Goldberg, D., Maddison, J. R., and Todd, J.: A framework for time-dependent ice sheet uncertainty quantification, applied to three West Antarctic ice streams, The Cryosphere, 17, 4241–4266, https://doi.org/10.5194/tc-17-4241-2023, 2023.

Reuter, B. W., Geraci, G., and Wildey, T.: ANALYSIS OF THE CHALLENGES IN DEVELOPING SAMPLE-BASED MULTIFIDELITY ESTIMATORS FOR NONDETERMINISTIC MODELS, International Journal for Uncertainty Quantification, 14, 1–30, 2024.

Ritz, C., Edwards, T. L., Durand, G., Payne, A. J., Peyaud, V., and Hindmarsh, R. C. A.: Potential sea-level rise from Antarctic ice-sheet instability constrained by observations, Nature, 528, 115–118, https://doi.org/10.1038/nature16147, 2015.

Rockafellar, R. T. and Uryasev, S.: The fundamental risk quadrangle in risk management, optimization and statistical estimation, Surveys in Operations Research and Management Science, 18, 33–53, https://doi.org/https://doi.org/10.1016/j.sorms.2013.03.001, 2013.

Schaden, D. and Ullmann, E.: On Multilevel Best Linear Unbiased Estimators, SIAM/ASA Journal on Uncertainty Quantification, 8, 601–635, https://doi.org/10.1137/19M1263534, 2020.

Schlegel, N.-J., Seroussi, H., Schodlok, M. P., Larour, E. Y., Boening, C., Limonadi, D., Watkins, M. M., Morlighem, M., and van den Broeke, M. R.: Exploration of Antarctic Ice Sheet 100-year contribution to sea level rise and associated model uncertainties using the ISSM framework, The Cryosphere, 12, 3511–3534, https://doi.org/10.5194/tc-12-3511-2018, 2018.

Seroussi, H., Verjans, V., Nowicki, S., Payne, A. J., Goelzer, H., Lipscomb, W. H., Abe-Ouchi, A., Agosta, C., Albrecht, T., Asay-Davis, X., et al.: Insights into the vulnerability of Antarctic glaciers from the ISMIP6 ice sheet model ensemble and associated uncertainty, The Cryosphere, 17, 5197–5217, https://doi.org/10.5194/tc-17-5197-2023, 2023.

Stuart, A. M.: Inverse problems: A Bayesian perspective, Acta Numerica, 19, 451–559, https://doi.org/10.1017/S0962492910000061, 2010.

Tezaur, I., Peterson, K., Powell, A., Jakeman, J., and Roesler, E.: Global sensitivity analysis of ultra-low resolution Energy Exascale Earth System Model (E3SM), Journal of Advances in Modeling Earth Systems, submitted, 2021.

Van Katwyk, P., Fox-Kemper, B., Seroussi, H., Nowicki, S., and Bergen, K. J.: A Variational LSTM Emulator of Sea Level Contribution From the Antarctic Ice Sheet, Journal of Advances in Modeling Earth Systems, 15, e2023MS003 899, https://doi.org/10.1029/2023ms003899, 2023.

Verjans, V., Robel, A. A., Seroussi, H., Ultee, L., and Thompson, A. F.: The Stochastic Ice-Sheet and Sea-Level System Model v1.0 (StISSM v1.0), Geoscientific Model Development, 15, 8269–8293, https://doi.org/10.5194/gmd-15-8269-2022, 2022.

Weis, M., Greve, R., and Hutter, K.: Theory of shallow ice shelves, Continuum Mechanics and Thermodynamics, 11, 15–50, https://doi.org/10.1007/s001610050102, 1999.

Yoo, M., Gopalan, G., Hoffman, M. J., Coulson, S., Han, H. K., Wikle, C. K., and Hillebrand, T.: Uncertainty-enabled machine learning for emulation of regional sea-level change caused by the Antarctic Ice Sheet, https://arxiv.org/abs/2406.17729, 2024.