Thank you again to the reviewers for their additional comments. Below we respond to each comment. Changes made while responding to the reviewers remarks are made in red in the revised document.

1 Reviewer 1 (Douglas Brinkerhoff)

No further comments were provided.

2 Reviewer 2 (Vincent Verjans)

2.1 General Comment

"This is my second review of this manuscript. Firstly, I want to commend the authors for their thorough work on the revisions. I find this second version of the manuscript much improved compared to the first submission. The methodology is much better explained, and the presentation quality is also greatly improved. At this stage, I have no major reservation concerning the manuscript. However, I still have plenty of minor technical comments, mostly focused on presentation details. Once these comments are addressed, as well as possible comments from other reviewers, I believe that this study will be a valuable contribution to the field of uncertainty quantification in ice sheet modeling."

2.2 Specific comments

- *"L72 "Basal friction field" should not be capitalized."*

  Fixed.

- *"L78 I believe that "mean and mass" should be mean and variance."*

  Fixed.

- *"L83 Typo: "defined a realistic physical domains"."*

  Fixed.

- *"L85 Typo: "a ice sheet model"."*

  Fixed.

- *"L86 I suggest specifying: two uncertain scalar parameters. Otherwise, one could think that the current study investigates only a single parameter (the basal friction field), while Gruber et al. (2023) quantifies uncertainty from two parameters."*

  Fixed.

- *"L120 For this summary sentence to be useful, it should include information about the Stokes model as well."*

  We removed the sentence completely, as it has caused confusion in multiple reviews and is captured in the previous discussion.

- *"L158 Why is there an extra term $\rho g(s - z)$ in the boundary condition on Γ_m here compared to the Stokes model?"*

  The Stokes boundary condition on Γ_m is

  $$\sigma \mathbf{n} = \rho_w \, g \, \min(z, 0)\mathbf{n},$$

or, equivalently,

$$2\mu D\mathbf{n} = p\mathbf{n} + \rho_w \, g \, \min(z,0)\mathbf{n}.$$

The Blatter-Pattyn and MOLHO model use the approximated condition

$$2\mu\hat{\mathbf{D}}\,\mathbf{n} = \rho g(s-z)\mathbf{n} + \rho_w \, g \, \min(z,0)\mathbf{n},$$

where $\hat{\mathbf{D}}$ is an approximation of $\mathbf{D}$, and $\rho g(s-z)$ is the hydrostatic approximation of p. Therefore the term $\rho g(s-z)$ is not an "extra" term, but just the approximation of the ice pressure p, which is included in the Stokes boundary condition.

- *"L172 "integrated by part": part should be plural."*

  Fixed.

- *"L193 (...) quantifying uncertainty in modeled ice mass change subject to high-dimensional parameter uncertainty (...)."*

  Fixed

- *"L206 Typo: "a infinite-dimensional"."*

  Fixed.

- *"L240 Please use the wording in thermal equilibrium."*

  Fixed.

- *"L251 Please specify two-dimensional surface velocities, so that the dimensionality $2N_{obs}$ is immediately clear to the readers."*

  Fixed.

- *"L257 is approximately in steady-state."*

  Fixed.

- *"Equation 12 I apologize for this request, but I consider it bad practice to use a same symbol for different notions. In this manuscript, α is used to denote the likelihood hyper-parameter as well as for model indexing. Please consider using another symbol here."*

  We now use the symbol ξ for the scaling term in Eq. 12.

- *"L263 I believe that it is the errors in the observational data that are assumed to be uncorrelated, not the data themselves."*

  Fixed.

- *"L268 Please use the wording in thermal equilibrium."* Fixed.

- *"L270 Please cite the work of Adalgeirsdottir et al. (2014) here."*

  The citation was added.

- *"L321 Typo: "In out study"."*

  Fixed.

- *"L331-332 Please italicize I and II to be consistent with Eq. (18)."*

  Fixed.

- *"L349 Although equivalent, this should be $Q_0 - Q_\infty$."*

  Fixed.

- *"L368 "MSE error" should simply be MSE."*

  Fixed.

- *"L388 Typo: "can be simulate"."*

  Fixed.

- *"L388 Typo: end sentence before "However"."*

  Fixed.

- *"L388-390 This statement is not clear to me: (i) Even though it is assumed that the discretization can be refined indefinitely, in practice, how can Q_∞ be evaluated? (ii) What is the advantage of having a numerical discretization error equal to $V_\Theta[Q_{ACV}]$? Finally, note that the reference to Eq. (18) should be Eq. (19) here."*

  We added the statement: Balancing these two errors ensures that computational effort is not wasted on resolving one source of error more than the other. However, in practice, the geometric complexity of many spatial domains makes generating large numbers of meshes impractical, and estimating the discretization error using techniques like posterior error estimation can be challenging. The correct equation is now referenced.

- *"L394 Typo: 0 and 1 should be subscripts."*

  Fixed.

- *"L413 Please add a sentence to explain intuitively the advantage of having $\eta_{1,2}, \eta_{1,4}, \ldots, \eta_{1,2M}$ and $\eta_{2,1}, \eta_{2,3}, \ldots, \eta_{2,2M-1}$ potentially different from 0. That is, what is the advantage of including the terms capturing differences in σ^2 affecting Q^{μ}_{ACV} , and the terms capturing differences in μ affecting $Q^{\sigma^2}_{ACV}$?"*

  We added the statement Formulating the control variate weights as a matrix enables the ACV estimator to exploit the correlation between the statistics Q^μ and Q^{σ^2}, producing estimates of these individual statistics with lower mean squared error (MSE) than would be possible if the two statistics were estimated independently.

- *"L445 These equations were included in the previous iteration of the manuscript. Instead of referring the reader to Dixon et al. (2024), I strongly recommend to include the equations for these quantities into an Appendix C."*

  We included only some of the equations needed in our initial submission. At the suggestion of another reviewer, we removed these equations from our first revision because including all equations would require many pages and additional notation. We still believe that adding these equations would overly complicate the paper when they can be easily found in the cited paper and in tutorials provided by PyApprox which was used to generate the results.

- *"L452 This should be: to the number of pilot samples of the error in ACV MC estimators."*

  Fixed.

- *"L470 two-model."*

  Fixed.

- *"L474 two-model."*

  Fixed.

- *"L490 I believe that "next highest-fidelity" should be next higher-fidelity."*

  Fixed.

- *"L513 "MSE error" should simply be MSE."*

  Fixed.

- *"L520 Please remove "sorely needed"."*

  Fixed.

- *"L520 Typo: "and the error a multi-fidelity estimator"."*

  Fixed.

- *"L540-542 There is an error here. The mesh with lowest characteristic element size should have the most elements and nodes."*

  Fixed.

- *"L560 "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": from this paragraph, I understand that this is exactly what the authors did. Sorry if I misunderstand this, but I think that more clarity would be beneficial."*

  Sorry for the typo. We meant to say MOLHO instead of Blatter-Pattyn. We now state Solving the MOLHO model using the C++-based `MALI` code and solving and SSA using the python based `FEniCS`, would have corrupted the MFSE results.

- *"Figures 5 and 6 Please mention explicitly in the captions that the color scales of different sub-panels span different ranges."*

  We added the following statement to both plots: Note, the color scales of each plot span different ranges so that the variability in the quantities plotted is visible.

- *"L588 "hyper-parameter" should be plural."*

  Fixed.

- *"Remark 5.2 In my previous review, I made a comment about the potential sensitivity of the low-resolution models to the interpolation method used for the basal friction field. Essentially, I argue that the results from the low-fidelity models are likely sensitive to the interpolation method used, because the authors demonstrate (Figure 7) that the simulated thickness change is sensitive to high-frequency variability in the basal friction field. I believe that my comment has not been addressed appropriately. In particular, I believe that the authors should at least (i) state clearly which interpolation method is used (linear, nearest neighbor, polynomial, other), (ii) point out that the basal friction field is too smooth due to the interpolation, as suggested by Figure 14, and (iii) explicitly state that using interpolation methods preserving more of the high-frequency variability in basal friction would likely improve the ability of the low-resolution models to reproduce f_0. Points (ii) and (iii) are mentioned in the first paragraph of Appendix B, and should be incorporated in the main text."*

We added the following statement to the remark Specifically, we used the linear finite element basis of the fine mesh to interpolate onto the coarser meshes. This procedure ensured that varying the basal friction field (the random parameters to the model) would affect each model similarly, regardless of the mesh discretization employed. However, the linear interpolation we used may have overly smooth the friction on coarse meshes relative to alternative higher-order interpoolation methods. Consequently, using alternative interpolation methods may increase the correlation between the mass loss predicted by the coarse meshes and that predicted using the finest mesh. However, we did not explore the use of alternative iterpolation schemes because our results demonstrate that linear interpolation still produces models that can be used produce a computationally efficient MFSE.

- *"L603 Typo: "of the evaluating all 13 models"."*

  Fixed.

- *"Figure 9 Because of the colormap chosen and the white text font, some correlation entries are impossible to read."*

  Fixed.

- *"Figure 9 caption Typo: two commas."*

  Fixed.

- *"L623 If the authors are referring to Peherstorfer et al. (2016), then they should cite it here."*

  Most papers, if not all papers, make this assumption.

- *"L642 I believe that the language here could be slightly confusing. The authors never use "all models simultaneously" because they impose maximum 4 models per sample. I know this is not what is meant here, but more precise language would be beneficial."*

  We now state: Moreover, bootstrapping the estimators also revealed that not all models are equally useful when reducing the variance of the ACV estimator. In some cases, using 3 models was more effective than using four models.

- *"L651 I believe that "next highest-fidelity" should be next higher-fidelity."*

  Fixed.

- *"Figure 11 In my previous review, I asked the authors to specify the number of samples for each case (2, 3, and 4 models). The authors responded that no one number can be provided. I do not understand this. If the authors can make a boxplot, it means that they have a list of values (in this case, values of variance reduction). As such, they can provide the number of values from which each single boxplot was made."*

  We apologize for the confusion, However, we are still not sure what number of samples you are referring to. As stated on L663 of the first revision *" 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. "* Thus, 21 different samples of the covariance matrix were used to generate the box plots for the 2, 3, and 4 model cases. Each bootstrap randomly selected 30 samples, with replacement, from the evaluations of each model at the 30 unique pilot samples. In our last response, the number of samples we referred to when stating "no one number can be provided" was the optimal number of samples allocated to each model for each bootstrap.

- *"Figure 11 Please increase the font size of the median values."*

  Fixed.

- *"L666 Typo: "this increasing"."*

  Fixed

- *"L666 Here, "which" refers to the increased computational cost of the pilot study. However, the computational cost does not affect the variance reduction value. Instead, as explained in the following paragraph, the lower median variance reduction is caused by having more pilot samples of the highest-fidelity model for the SFMC estimator. Please avoid this confusing wording."*

  The increased computational cost of the pilot study does affect the variance reduction value. We have added the statement: This fact can be explained by recalling the SFMC estimator variance, e.g. $\mathbb{V}_\Theta\left[Q_0^\mu\right]$, was obtained using a computational budget equivalent to 160 high-fidelity evaluations plus the computational cost of collecting the pilot model evaluations. In contrast, the ACV estimator variance, e.g, $\mathbb{V}_\Theta\left[Q_{\text{ACV}}^\mu\right]$, does not depend on the number of pilot samples. Therefore, while increasing the number of pilot samples decreases the SFMC estimator variance, it does not decrease the the ACV estimator variance. Consequently, increasing the pilot cost reduces the variance reduction achieved by the ACV estimator.

- *"L687 Please rephrase to make totally clear that it is each one of the 160 model evaluations that takes 4.18 hour, and not "evaluating the high-fidelity model 160 times" that takes 4.18 hour."*

  We now state: The exploitation cost was fixed at the beginning of the study to the computational cost equivalent to evaluating the high-fidelity model 160 times, with each simulation taking a median time of 4.18 hours to complete for a single realization of basal friction.

- *"L698-699 Please specify that this statement about error from spatial versus temporal discretizations is valid within the ranges of discretizations tested here."*

  Fixed.

- *"L700-703 Please break this sentence in two."* Fixed.

- *"L704 Add units."*

  Fixed.

- *"L716 Typo: "model model"."*

  Fixed.

- *"L720 Typo: "if" should be of."*

  Fixed.

- *"L732-733 Needs rephrasing: "Similarly, for the exploitation phase. Moreover, each simulation ca can be computed in parallel."."*

  We now state Each simulation run in the pilot stage can be executed in parallel without communication between them. Similarly, in the exploitation phase, each simulation can also be computed in parallel.

- *"L756 Typo: citation format."*

  Fixed.

- *"L760 Typo: "estmators"."*

  Fixed.

- *"L784 Change: correlation with models."*

  Fixed.

- *"L805 Rephrase: "facilitate computationally efficiently quantify"."*

  Fixed.

- *"L832 Typo: "In out study"."*

  Fixed.

- *"L850 Add comma after "limitations"."*

  Fixed.

- *"L919 Change "Please refer to Figure B5" to simply putting (see Fig. B5). Sorry if my comment in my previous review was unclear."*

  Fixed.

- *"Caption of Figure B3 Add space between Eq. and (B1)."*

  Fixed.

- *"Caption of Figure B4 Add space between Eq. and (B1)."*

  Fixed.

3 Reviewer 3 (Dan Goldberg)

3.1 General comments

"I have read in detail the responses of the authors to my comments, and appreciate they have worked extensively both to improve the manuscript and address my concerns. I think if the other reviewers are satisfied then [this was cutoff in the user portal]

*However, my one remaining concern is about the strength of the prior, in particular the autocorrelation scale of 30 km, which i still am confused about as it seems large, and wonder if the authors can give some physical or observational justification. is it based on velocity or bed semivariograms? I fully accept this is not a variable the authors can tractably explore, but i feel that it should be defensible, given the accepted knowledge that bed conditions and water pressure (which determines bed conditions) vary on quite short length scales. And if not, then the authors should caveat – **very clearly** – that these scales are very difficult to know a priori, and that overly informative prior densities could significantly impact posterior mass loss estimates. (they make clear in their response that the QoI uncertainty is not the main thrust of the paper so doing so should not affect their main results)*

I may seem I am being pedantic, but this is a paper on ice-sheet uncertainty quantification, of which there are very few, which is exactly why such things should be brought to light. "

We now state the following in the final paragraph of the conclusion section of the paper.

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. However, the predicted mean mass loss from Humboldt Glacier that we reported, should be viewed with caution, as the length scales of the prior distribution we employed for the uncertain basal friction field were not finely tuned, e.g. as done by Recinos et al. (2023), due to the computational expense of such a procedure. Since the length scales of basal friction are very difficult to determine a priori, the potentially over-informative length scales used in this paper could substantially impact the posterior mass loss estimates. Future research should address this issue and increase the complexity of this study in two further directions.