Review of “An evaluation of multi-fidelity methods for quantifying uncertainty in projections of ice-sheet mass-change” by Jakeman et al.
Reviewer: Vincent Verjans

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 understand 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 review 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.

I) Equations should be better explained and without errors.
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.

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 $\mathrm{Var}(Q^{\sigma^2})$ formula, so please provide a reference and/or a detailed derivation in the response.
In L298, to be valid, this equation requires some normalizing terms $(1/N_0, 1/N_0, 1/N_1)$.
In the first part of Eq. 16, one term should be $\mathrm{Cov}[\Delta_\alpha^{\sigma^2}, \Delta_\beta^{\mu}]$.
In Eq. 11, both Q_α and Q are referred to as MC estimators (in Eq. (10) and on L271, respectively). It would be nice define the MC estimator precisely, as well as the quantity that it is estimating.
Please be consistent in the notation. For example, Q_{ACV} is bold in Eq. 15, but not in Eq. 20.
Please use a symbol to represent only one single variable or parameter. For example, the letter η and the letter n are both used to represent different things in the manuscript.
Please use equation numbers for all equations.
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.

II) 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.

- 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.
- 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.
- 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.
- L298: Why is this valid regardless of how truthfully f_1 approximates f_0?
- L322: What do the control variates represent?
- 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?
- Etc.

Line-by-line comments

- 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.

- 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.

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.
L4
Replace ice sheet by glacier.
L5
The problem size is not "representative" of continental scale studies. Please use more careful wording.
L11
prediction should be plural.
L15
Add report after IPCC.
L15
Ice sheets are all land-based.
L26
Throughout the manuscript, affect should be used as a verb instead of effect.
L28
Replace inadequacy by uncertainty.
L31
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.
L39
There are also methods that have been developed to reduce the problem dimensionality. Please cite Brinkerhoff (2022).
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 expectation 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.
L50
I think that the authors might not be aware of the study of Bulthuis et al. (2020). Please consider referring to it.
L61
I find the changes between past and present tense somewhat confusing. I recommend consistently using a single tense.
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.
L102-104
In this sentence, the summary of the Stokes and MOLHO models sound identical to me.
L113
Replace exorbitant by impractical.
L121
Provide a reference for $q = 1/3$.
L125
Define the $||$ notation here.
L139
The ψ term is already multiplied by $\mathbf{n}$ 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?
L145
$\partial\Sigma$ is not defined.
Figure 2
Show the meshes side-by-side (+ all comments from General (2)).
L172
The statement "one of the largest sources of prediction uncertainty" should be quantified and referenced with a citation.
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.
L183
There is a switch from C to Σ without mentioning it. Specify that Σ_{prior} is a covariance.
L185
Why is the source term only integrated over Γ_g and not over Γ_f? I would assume that snow accumulation and surface melting should also be computed over the floating parts of the domain.
L199
Please specify "this Gaussian prior".
L199
Replace "on" by "constrained with".
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.
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.
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).
L222
"the likelihood distribution": the likelihood is a function, not a distribution.

L232
Please add a justification for this choice of α.
L233
Please specify "samples from the posterior of $\log(\beta)$".
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 ($H_{\mathrm{MAP}} + \Sigma_{\mathrm{prior}}^{-1}$). Thus, a brief sentence to explain why this is intractable would be beneficial.
L254 and 255
Replace ice sheet by glacier.
L263
What do the authors mean by "robust"?
L264
"three-step"
L265
The m superscript should be n (which would preferably be another letter than n, see Major comment).
L266
Specify "basal friction field".
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.
L287
Typo estimated.
L296
Two-model
L316
QoI is not defined.
L324
Concerning $\Theta_{\alpha}^{*} \cup \Theta_{\beta} \neq \emptyset$, (i) I believe that $\cup$ should be $\cap$, (ii) I believe that "for $\alpha \neq \beta$" should be specified.
Eq. (16)
Is $\mathrm{Cov}[\mathbf{Q}_0, \mathbf{\Delta}_0](\Theta_{\mathrm{ACV}})$ 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.
L351
"following standard practice": provide citation.
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}.
L360
Specify "introduce sampling errors".
L367
I believe that the same should be specified for $\alpha \cup \beta^{*}$ and $\alpha \cap \beta^{*}$.

L382
There is no verb in this sentence.
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.
L394
I think this should be estimator types.
L394
"model models subsets" is either a typo, or very confusing language.
L402
was should be were.
L413
ice-sheet should be glacier.
L415
Typo an an.
L424
Specify MALI ice-sheet code with the Blatter-Pattyn flow model.
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.
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.
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.
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.
L460
The meaning of accuracy is not clear here (see comment on L4).
L476
Provide a citation to support this statement.
L477

Please quantify “the error introduced”.
L477
“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.
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.
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.
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?
L489
Specify subsets of model combinations.
L491
the original 20 pilot samples are used.
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.)
L496-499
I could not understand the end of this paragraph. It would be helpful if the authors defined the notion of hierarchical relationship.
Figure 11
Please specify the number of samples for each case (2, 3, and 4 models).
L520
Again, the meaning of “accurate” is not well-defined.
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.
L522
Replace that by which (with a comma, see General comment (1)).
L523
The three models listed do not include $\text{MOLHO}^{*}{}_{1\text{km},9\text{days}}$. 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.

L525
Please remind the readers where these numbers come from.
L527
I do not see any right or left panel.
L535
Please specify another estimator (i.e., MLMC or ACVMF).
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.
L567
Appendix B.
Figure 13 (1)
Changing the color scale here is absolutely necessary.
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.
L589
"variance" should be standard deviation here, since Gigaton units are specified.
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.
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.
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.
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 $\mathcal{O}(1-10)$ 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.
L618
"substantially": please quantify and provide a citation.

L638
I do not understand the underlying meaning of this sentence. Please expand or remove it.
L641
Antarctica and Greenland.
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.
L668
Please define K here as well. Otherwise, the reader needs to go back to the main text.
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.
L685
Typo: in this study
L686
Typo: modes
L700
I believe that MF estimator should be ACV estimator
L701
I believe that MFUQ estimator should be ACV estimator
L702
This should be: The mean and variance bootstrapped (...).
L706
This should be: the uncertainty in the mean mass change (...)
L706-708
Please refer to Figure B5.

References

Douglas J Brinkerhoff. Variational inference at glacier scale. *Journal of Computational Physics*, 459:111095, 2022.

Kevin Bulthuis, Frank Pattyn, and Maarten Arnst. A multifidelity quantile-based approach for confidence sets of random excursion sets with application to ice-sheet dynamics. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):860–890, 2020.

DN Goldberg. Committed retreat of smith, pope, and kohler glaciers over the next 30 years inferred by transient model calibration. *The Cryosphere*, 9(6):2429–2446, 2015.

G Hilmar Gudmundsson and Melanie Raymond. On the limit to resolution and information on basal properties obtainable from surface data on ice streams. *The Cryosphere*, 2(2):167–178, 2008.

Vincent Verjans, Alexander A Robel, Helene Seroussi, Lizz Ultee, and Andrew F Thompson. The stochastic ice-sheet and sea-level system model v1. 0 (stissm v1. 0). *Geoscientific Model Development*, 15(22):8269–8293, 2022.