Review of "Drawing lessons for multi-model ensemble design from emulator experiments: application to future sea level contribution of the Greenland ice sheet" by Rohmer et al.

The paper *"Drawing lessons for multi-model ensemble design from emulator experiments: application to future sea level contribution of the Greenland ice sheet"* investigates the impact of different choices in constructing an emulator capable of predicting the future contribution of the Greenland Ice Sheet to sea-level rise. Specifically, the authors examine how the selection of models and parameters associated with different physical processes and scales (climate scenarios, regional climate models, ice-flow models, and ice-flow parameters) influences the emulator in terms of its fidelity, the estimated contribution to sea-level rise, and the associated uncertainty.

This study represents a valuable contribution to our understanding of multi-model ensemble approaches and emulator design. The numerical experiments are clearly described, and the results are noteworthy. Overall, the paper provides an interesting addition to the scientific literature. Nonetheless, I have a series of comments and questions that I would like the authors to address. On the one hand, in terms of the way the manuscript is written, some sections could be streamlined for conciseness (see my first general comment below, as well as specific comments). On the other hand, I have a series of more fundamental general comments on key aspects of the manuscript. There are three main comments, described herefater, which are followed by a series of specific comments addressing more detailed points. Once these concerns have been satisfactorily addressed, I will be happy to recommend the paper for publication in the Special Issue: *Improving the contribution of the land cryosphere to sea level rise projections*.

General comment 1

This paper lies at the intersection of two fields: glaciological modeling and statistical methods. Such interdisciplinary studies are particularly valuable, as the glaciology community may not be fully familiar with statistical techniques, while statisticians may not always be aware of the challenges involved in estimating future sea-level rise. Furthermore, comparative studies have gained increasing importance in glaciology, as they help assess the robustness of different modeling approaches. Given this, it is crucial to also investigate how these comparisons are constructed in the first place. In this context, the present study is highly relevant.

That being said, I believe the paper could be more explicit about its practical conclusions, specifically conclusions (1) and (2) mentioned in the abstract. I would also suggest highlighting the conclusions related to the (different) impact of the RCM/ISM choice in the abstract and in the introduction (see lines 313–315). Additionally, some technical details could be either removed or moved to an appendix or supplementary materials. The reasoning behind these suggestions is that most readers of this journal are likely geoscientists, who will primarily be interested in the study's results. By streamlining technical details, the paper's key findings –which are noteworthy– could be emphasized more effectively, improving its accessibility.

I also wonder whether subsections 2.4 and 3.2 could be simplified by focusing on just one of the two indicators (D_S and D_h), e.g. D_S. While some variability has been observed in the results, these indicators are strongly correlated (see figure 6). Furthermore, from a practical point-of-view, what really matters is the number of runs to be made (i.e., of D_S). Similarly, figure captions could be streamlined by omitting details that may not be particularly useful (e.g., the exact method used for kernel density estimation; see specific comments below). The technical lines 196–219 could also be relegated to supplementary materials.

General comment 2

Regarding the paper's methodology, I have some questions about how the way ensembles are introduced. Specifically, lines 34–36 state: *"Each member of a MME should evenly span a representative and exhaustive set of plausible realizations of the combined sources of uncertainty, (...), equally represented by a single model run"*. This suggests that each member of the ensemble should have the same weight. However, I find this somewhat misleading, and I believe the authors could elaborate further on this choice. The assumption that all runs should have equal weight holds only if our current knowledge suggests they are equally probable. From a Bayesian perspective, this would correspond to assuming a uniform prior distribution. However, this assumption may not always be justified. For example:

(i) Runs from lower-resolution models might be considered less reliable than those from higher-resolution models as they might not capture relevant small-scale processes.

(ii) Some values of the uncertain parameters might be less probable if they lead to results that deviate significantly from current observations.

Formally, these concerns can be addressed by updating the weights of each run based on their likelihood given observational data (e.g., Aschwanden and Brinkerhoff, 2022; Nias et al., 2023).

I understand that the authors did not include such a step in their analysis, as their focus was on assessing the emulator's capabilities. However, it seems to me that this point should be discussed in the methodology or discussion section for two reasons. First, to clarify for the reader that the choice of equal probability stems from an assumption about our current knowledge and that alternative approaches are possible. Second, because weighting model runs based on observational constraints is an emerging direction in the field, and this should be discussed in the context of future ISMIP ensemble designs. This could also be mentioned as a perspective for future work, as it would be interesting to see whether the conclusions remain valid when runs are weighted as a result of a calibration.

General comment 3

My third general remark concerns the parameter κ associated with the calving rate. It is unclear why this particular parameter was chosen over others. From reading the paper, the rationale behind this choice is not obvious—perhaps it is based on modeling considerations or supported by previous studies? If so, it seems to me that the authors should provide a stronger justification for including this specific parameter.

More fundamentally, I wonder whether comparing the effect of κ to that of RCP scenarios or RCMs is entirely meaningful. This comparison contrasts the impact of a single parameter of an ice-flow model with that of an entire climate scenario or a regional climate model, which incorporate numerous physical parameters. Given this, it is perhaps unsurprising that the effect of κ appears quite limited.

To illustrate this, one could consider a similar comparison in the opposite direction: assessing the impact of choosing a glaciological model of varying complexity (e.g., full Stokes, BP, or SIA) against a single parameter from a RCM. This would likely lead to the conclusion that the specific parameter from the RCM has a minimal influence. Therefore, I wonder whether including κ as an isolated parameter in this study is fully justified. Could the authors maybe clarify its relevance within the broader context of the study's objectives?

Specific Comments

(1) [Line 12] ‘projection and the quantification of its uncertainty’ → ‘projections and the quantification of their uncertainties’.

(2) [Lines 15, 17, and 65] You use ‘experiments’ for two distinct concepts: numerical simulations (e.g., line 16) and numerical tests (e.g., line 15). Consider using separate words to avoid any confusion.

(3) [Line 16] ‘(Regional Climate Model RCM, or Ice Sheet Model ISM)’ → ‘(Regional Climate Model; RCM, or Ice Sheet Model; ISM)’.

(4) [Line 19] Consider removing ‘utmost’ as it might be overly formal.

(5) [Line 25] ‘projection and the quantification of its uncertainty’ → ‘projections and the quantification of their uncertainties’.

(6) [Line 26] ‘co-ordinated sets of numerical experiments’ → ‘sets of numerical experiments’.

(7) [Line 47] Consider adding references related to (machine-learning-based) emulators.

(8) [Line 47] I might be a bit picky here, but I would argue that the key advantage of statistical emulators is their low computational cost; being able to predict the model response at untried input values is only useful if it can be done at a reasonable cost.

(9) [Line 63] Please be consistent with your use of acronyms: either your define what you mean by RCM and GCM, or you use directly the corresponding acronyms. Also, a table of acronyms would be useful in the paper.

(10) [Line 63] Avoid using ‘validation tests’ as this can lead to confusion when it comes to glaciological modeling (for which ‘validation’ has another meaning).

(11) [Line 76] ‘(Goelzer et al. (2020): in particular (...))’ → ‘(Goelzer et al., 2020; in particular (...))’.

(12) [Line 79] Consider adding a schematic displaying the modeling chain and indicating where modeling choices (MME inputs) are introduced. This could be very useful to effectively obtain an overview of the context.

(13) [Line 80] Here you define again what a RCM is. If you have already defined it before that is not necessary.

(14) [Table 1] Ensure consistent formatting of symbols (italics vs. non-italics).

(15) [Table 1] Consider renaming ‘Symbol’ to ‘Symbol/Acronym’ or simply ‘Name’ to clarify that most entries are acronyms.

(16) [Table 1] ‘Sliding basal law’ → ‘Sliding law’ or ‘Basal friction law’.

(17) [Line 101] Consider defining ‘input setting’ explicitly, e.g. as a particular combination of inputs.

(18) [Line 102] ‘Minimal resolution’ is ambiguous. If referring to spatial resolution, clarify: ‘e.g., for κ, or for the minimum spatial resolution’.

(19) [Figure 1] Consider changing the y-label to ‘Probability distribution’, ‘Probability distribution function’, or ‘PDF’.

(20) [Figure 1] ‘Greenland ice-sheet’ → ‘Greenland ice sheet’.

(21) [Figure 1] ‘raw MME’ → ‘original MME’.

(22) [Figure 1] Consider simplifying the caption by just stating that the black line is based on kernel density estimation, removing extraneous details.

(23) [Line 114] ‘have the highest importance’ is not clear; consider replacing it with something more precise, e.g. ‘contributes to most of the uncertainty of the slc’.

(24) [Line 118] Replace ‘distributions’ to avoid confusion with PDFs.

(25) [Figure 2] The orientation of the plots is a bit unusual. Consider switching the x and y axes, with the count number in the y direction.

(26) [Figure 2] ‘to the different inputs’ → ‘to different inputs’.

(27) [Figure 3] The orientation of the plots is a bit unusual. Consider switching the x and y axes, with the count number in the y direction.

(28) [Figure 3] ‘to the different inputs’ → ‘to different inputs’.

(29) [Line 134] It seems that this study only concerns sea-level contribution at 2100. If so, avoid referring to ‘a given time t’.

(30) [Table 2] Note that your question regarding the SSP-RCP scenario only corresponds to the ‘woSSP245’ case. Maybe make it a bit more general by asking whether removing one of the scenarios (SSP1-2.6, SSP2-4.5, SSP5-8.5) actually leads to an improvement of the results.

(31) [Table 2] Is there a reason why you did not consider a ‘woMAR’ case (by contrast with the ‘CISM’/‘woCISM’ cases for the ISM)?

(32) [Line 175] Avoid using ‘validation test’ as this can lead to confusion when it comes to glaciological modeling (for which ‘validation’ has another meaning).

(33) [Line 180] I wonder if it would not be simpler to represent the probability distribution of the normalized error $|e/\text{slc}|$ by treating the input setting as an uncertain parameter (i.e. by considering the full sample space of the original MME). By doing so, you would avoid depending on additional numerical parameters ($n_{\text{test}} = 55$ and 25 test samples) for your statistics.

(34) [Figure 5] Consider changing the y-label to ‘Probability distribution’, ‘Probability distribution function’, or ‘PDF’.

(35) [Figure 5] Consider simplifying the caption by removing details about density estimation. Stating that the PDF was estimated using a Monte-Carlo procedure does not seem to be a relevant information here.

(36) [Line 249] Avoid using 'logically'. If the rejected members are not representative of the input space, their removal may not necessarily improve the predictability of the random-forest surrogate.

(37) [Line 262] I wonder whether a table that contains for each test both the median RAE and the value of D_{S} could be an efficient way to summarize these results. You could then obtain a measure associated with their ratios, i.e., that corresponds to the decrease in the number of ensemble members (D_{S}) and the 'cost' to pay (the RAE).

(38) [Table 3] Note that your question regarding the SSP-RCP scenario only corresponds to the 'woSSP245' case. Maybe make it a bit more general by asking whether removing one of the scenarios (SSP1-2.6, SSP2-4.5, SSP5-8.5) actually leads to an improvement of the results.

(39) [Line 311] 'long numerical simulations' → ' numerical simulations' as you never discussed the length of numerical simulations previously.

(40) [Lines 328–332] Consider removing this sentence as it is not a novelty of this study.

(41) [Line 342] Did you previously introduce the acronym 'GWL'?

(42) [Line 356] Consider removing 'utmost' as it might be overly formal.

(43) [Figure B1] The orientation of the plot is a bit unusual. Consider switching the x and y axes, with the p-value in the y direction.

References

Aschwanden, A. and Brinkerhoff, D. J. (2022). Calibrated Mass Loss Predictions for the Greenland Ice Sheet. *Geophysical Research Letters*, 49(19).

Nias, I. J., Nowicki, S., Felikson, D., and Loomis, B. (2023). Modeling the Greenland Ice Sheet's Committed Contribution to Sea Level During the 21st Century. *Journal of Geophysical Research: Earth Surface*, 128(2).