

Replies to Referee #1's comments on "Drawing lessons for multi-model ensemble design from emulator experiments: application to future sea level contribution of the Greenland ice sheet" (egusphere-2025-52)

We would like to thank Referee #1 for the constructive comments. We agree with most of the suggestions and, therefore, we will modify the manuscript to take on board the comments and suggestions. In this document, we recall the reviews and we reply to each of the comments in turn (outlined in blue). Following the journal's reviewing procedure, the revised manuscript will be provided after the interactive review process, in a second phase.

Additional changes

Since the submission of this manuscript, we had the opportunity to include HIRHAM in the MME as well. In addition to the modifications suggested, we will thus include this third RCM model in the analysis. We expect that this modification will bring new insights and strengthen our conclusions which are, in the original version of the manuscript, based on two RCMs only. The results discussed in Sect. 4 will thus be modified.

Referee #1:

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 hereafter, 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.

We thank Referee #1 for the positive analysis. We will take the comments and suggestions into account. In what follows, we describe in detail the corrections that will be made in the revised version of the manuscript.

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.

We thank Referee #1 for these suggestions. We totally agree that the results described in our manuscript should be transferred to a wide readership that is not necessarily specialists of the methods.

To do so, the following modifications will be made:

- The conclusions of the RCM/ISM will be more highlighted in the abstract and in the conclusions;
- The technical details of the emulator implementation will be placed in Appendix A. We will also consider adding a new Appendix to detail the methods for the performance analysis suggested by Referee #2. Finally, the specific comments of Referee #2 will also be integrated in Appendix to decrease the level of technicality of the core text;
- In order to improve the transferability of our message, some choices made for the representation of the results are modified to be more consistent with IPCC standards, i.e. with more largely shared practices:
 - o We will slightly change the definition of GSAT by computing the difference between the temperature at the considered year and the mean temperature over the period 1995-2014;
 - o To ensure consistency, this is also done for *slc*, which is computed relative to the mean *slc* over the same period;
 - o We will analyse the changes in the likely range, i.e. 17th and 83rd percentile instead of the 5th and 95th percentile.

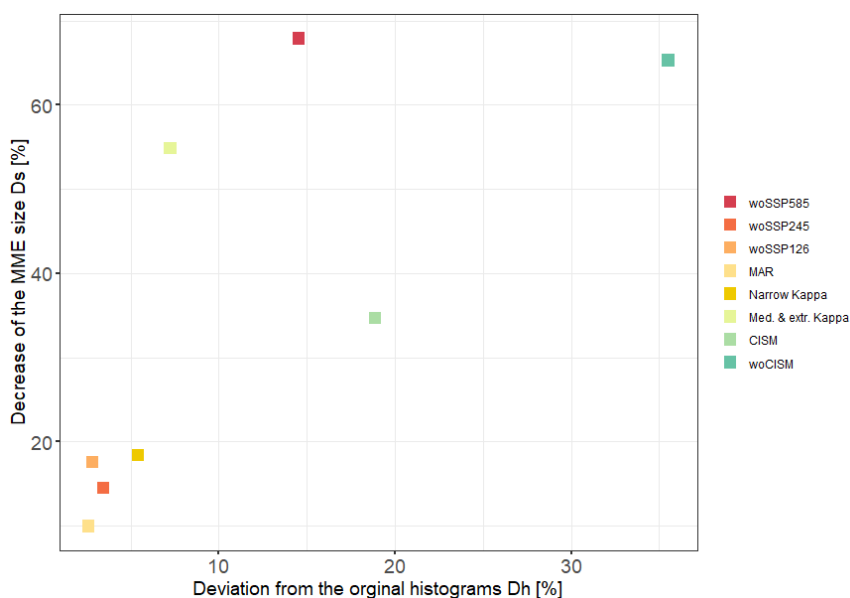
I also wonder whether subsections 2.4 and 3.2 could be simplified by focusing on just one of the two indicators (DS and Dh), e.g. DS. While some variability has been observed in the results, these indicators are strongly correlated (see figure 6). Furthermore, from a practical point-of-view, what really matters is the number of runs to be made (i.e., of DS). 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.

We agree with Referee #1's comment: the number of members is a key criterion, and the evolution of the emulator's predictive capability clearly highlights this aspect. Our primary idea for proposing a second criterion was to be able to reflect how much the distribution of the members is modified in addition to the size. Typically, we intend to highlight different

situations for a given decrease in size: (1) the decrease affects all levels of the considered variables, for instance all ISMs; (2) the decrease specifically affects only one level, e.g. CISM and not the other ISMs. The two different situations can lead to different behaviors, especially regarding the probabilistic projections.

As rightly indicated by Referee #1, our proposed criterion D_h fails, however, to highlight these situations because of the too strong correlation with D_s . This goes also in the same line with Referee #2's comment (minor comment 2.1).

Therefore, we propose to rely on a well-established, and more widely used, criterion for comparing different probability distributions, namely the Kolomogorov-Smirnov (*KS*) criterion, instead of a criterion constructed from 'scratch'. In the revised version of the manuscript, we will discuss in more details the links between these groups highlighted in the following figure and the implications with respect to the emulator's performance.



New Figure 6. Position of the emulator's experiment in a (D_h, D_s) diagram where D_s measures the relative decrease in the MME size after applying the experiment, and D_h measures the deviation of the histograms from the original ones.

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.

*We agree with Referee #1 that this point merits further discussion, particularly in view of the forthcoming ISMIP7. The primary goal our study was to study the influence of different factors for a given MME, i.e. discovering the influence of groups of members. Thus, the implicit assumption of this procedure is that we do not assign any weight to the members *a priori*. We will clarify this aspect in the core text; this was also Referee #2's suggestion (Minor comment 4(b)).*

*This clarification is more particularly useful in Sect. 4 regarding the meaning of the emulator-based probabilistic projections. These projections do neither represent calibrated uncertainty accounting for model-observation misfits nor the *slc* probability distribution from the MME, because the uniform distribution over the input space is not representative of the MME itself (e.g., the minimum spatial resolution is clearly not uniform between 1 and 40 km, see Fig. 3).*

In addition, the alternative option based on weighting either through expertise (as illustrated by Referee #1 with the resolution) or based on model-observation misfits will be discussed in Sect; 5. We propose to highlight the benefits of the weighting approach, but also the challenges to do it, namely: (1) the need for good quality data; (2) the need for data over a large period in the past; (3) the need for some types of ISMs to adjust or adapt their implementation. For this latter aspect, it should be noted that constraining some ISMs on past observations like surface mass balance may be hard due to the procedure chosen for the initialisation; this is the case for ISSM-type models based on velocity assimilation that represent almost 40% of the models involved in ISMIP6 (Goelzer et al., 2020: table A1).

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.

We thank Referee #1 for this suggestion. We feel that some clarification about κ should here be given.

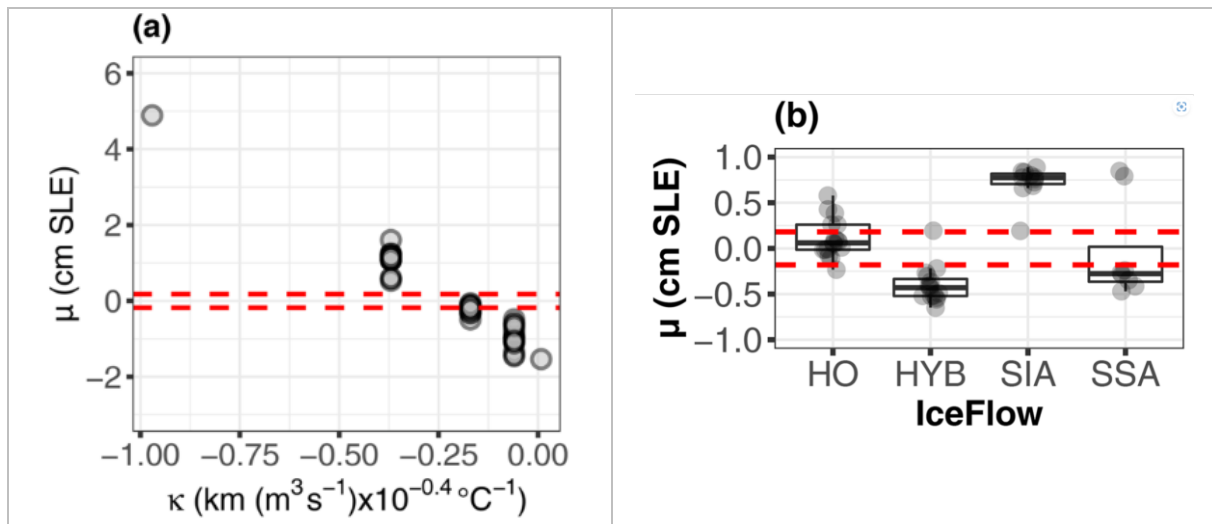
In this study, we rely on a standard approach for integrating ocean forcing, i.e. based on an empirically derived retreat parameterization for tidewater glaciers (Slater et al., 2019, 2020) that is forced by a RCM-based run-off and ocean temperature changes in seven drainage basins around Greenland. In this implementation, retreat and advance of marine-terminating outlet glaciers in the ISMs are prescribed as a yearly series of maximum ice front positions (Nowicki et al., 2020).

Here, κ is not thought as a parameter of the ice-flow model, it rather represents the sensitivity of the ocean forcing as a whole. It may be thought of as defining the sensitivity of the downscaling from global model to local ice sheet scale, similar to the combined parameter choices in RCMs for downscaling climate conditions. In the studied MME, we have different RCMs, which have different sensitivities and produce different melt for the same global forcing. Since we have only one approach to ‘downscale’ the ocean forcing, κ is sampling that uncertainty in a similar way.

We recognize that the κ -based approach remains a strong simplification of the complex interaction between marine-terminating outlet glaciers and the ocean, for which physically based solutions are in development but not available for all models. However, it should be underlined that the advantage of this retreat parameterization is to be applicable in the wide variety of models under consideration. Furthermore, it is currently the most widely used approach for producing large ensemble for sea level projections, as done for instance by Edwards et al. (2021).

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?

In response to to the specific comment of Referee #1, our previous study (Rohmer et al., 2022) highlighted the high importance of κ compared to other uncertainties, in particular some of them related to the complexity of the numerical model as suggested by Referee #1, i.e. the choice in the numerical method (Finite Difference, Finite Element), the grid resolution, and the ice flow formulation (approximation, higher order, hybrid). To illustrate, the following figure (adapted from Fig. 7 and Fig. 8 of Rohmer et al., (2022) based on ISMIP6 MME) shows the sensitivity index (denoted μ) that measures the contribution, in terms of sea level equivalent SLE, depending on the value of κ (panel a) or of the choice in the ice flow formulation (panel b). The influence measured by μ for κ is on the order of 1-2 cm (at most 5cm) whereas it remains on the order of 0.5-1cm for the ice flow method, hence confirming a large importance of this parameter.



References

- Edwards, T. L., Nowicki, S., Marzeion, B., Hock, R., Goelzer, H., Seroussi, H., et al. (2021). Projected land ice contributions to twenty-first-century sea level rise. *Nature*, 593(7857), 74–82.
- 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
- Rohmer, J., Thieblemont, R., Le Cozannet, G., Goelzer, H., and Durand, G.: Improving interpretation of sea-level projections through a machine-learning-based local explanation approach, *Cryosphere*, 16, 4637–4657, <https://doi.org/10.5194/tc-16-4637-2022>, 2022.
- Slater, D. A., Felikson, D., Straneo, F., Goelzer, H., Little, C. M., Morlighem, M., Fettweis, X., and Nowicki, S.: Twentyfirst century ocean forcing of the Greenland ice sheet for modelling of sea level contribution , *The Cryosphere*, 14, 985–1008, <https://doi.org/10.5194/tc-14-985-2020>, 2020.
- Slater, D. A., Straneo, F., Felikson, D., Little, C. M., Goelzer, H., Fettweis, X., and Holte, J.: Estimating Greenland tidewater glacier retreat driven by submarine melting, *The Cryosphere*, 13, 2489–2509, <https://doi.org/10.5194/tc-13-2489-2019>, 2019.

Specific Comments

We thank Referee #1 for the specific comments. The revised version of the manuscript will incorporate all of them. Following the journal’s reviewing procedure, the revised manuscript will be provided after the interactive review process, in a second phase.

(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 you 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.

Orleans,
May 5th, 2025

J. Rohmer¹ on behalf of the co-authors

¹ BRGM, 3 av. C. Guillemin - 45060 Orléans Cedex 2 – France