

Review of “Lessons for multi-model ensemble design drawn from emulator experiments: application to a large ensemble for 2100 sea level contributions of the Greenland ice sheet” by Rohmer et al.  
Reviewer: Vincent Verjans

This is my third review of this manuscript, following the second round of revisions. I am very pleased with the revisions made since the last round. The most notable improvements include:

- The analyses and conclusions that were previously not well supported by quantitative results have been thoroughly re-worded and/or removed. The arguments about the findings are now presented in a more methodological and substantiated manner.
- The evaluation procedure has been clarified. The revisions of Sects. 2.4.1 and 2.4.2 allow the reader to fully understand the predictive performance of the emulator.

More generally, I commend the authors for having very substantially improved the quality and scientific robustness of the manuscript compared to the very first submission. I also thank the authors for their efforts in addressing the concerns that I previously raised. In this review, I only raise one Minor General Comment, and some Specific Comments. My comments focus on some remaining minor issues, and on improving the clarity and readability of the manuscript. I believe that once these final points, as well as potential comments from other reviewers, are addressed, this study will be a valuable contribution to The Cryosphere. Line numbers refer to the main manuscript without tracked changes.

**Minor General Comment: Adding an explicit statement about ensemble size importance**

One of the most important driver of emulator performance is the size of the training set. I believe that the reduction in the MME size drives a large part of the strong performance decrease in the woMAR, woCISM, and woSSP585 experiments. It is not fortuitous that these 3 experiments lead to the largest performance decreases (Fig. 7), and are the ones with most restricted number of available members (Table 2). This is a strong argument for a key conclusion: the availability of large ensembles of ISM simulation outputs is the most important factor for developing accurate and reliable emulators.

In the current manuscript, this is alluded to (e.g., L429-430). But I believe that a clear and explicit statement about the critical importance of the MME size should be added in both the Abstract and in Sect. 5 “Concluding remarks and further work”. This would emphasize to the glaciological community that large participation to projects such as ISMIP6 are needed for designing useful emulators, with as many simulations from as many groups as possible.

**Specific comments**

L19-20. “for low and high levels of warning”: this sentence somewhat hides that the predictive performance is not satisfactory for intermediate levels of warming. Please specify this explicitly in the abstract.

L20. Typo: “warning”.

L175. “(...) are used to rank the different emulator experiments in terms of influence.” I see what the authors mean here. However, I think that the wording could be misinterpreted. It is not clear what the “influence” of an experiment refers to. Maybe rephrase this sentence, focusing more explicitly on the different impacts on emulator performance of different MME restriction experiments.

L247. Throughout the manuscript, I find the notations  $CA^\alpha$  and  $PI^\alpha$  potentially confusing. In statistical terminology,  $\alpha$  typically denotes the significance level, so  $\alpha = 0.1$  corresponds to the 90% confidence level, for example. However, the authors use expressions such as “ $CA$  at level 90%” (caption of Figure 5) and  $CA^{90}$  (e.g., L300). This is inconsistent with standard statistical conventions. Please revise the notation and associated wording to clarify whether superscripts refer to the  $\alpha$  level or to the confidence range. For example, I would recommend writing  $CA^{1-\alpha}$ , which would be consistent with, for example, writing  $CA^{90}$ .

L248. Typo: “fall” should be falls.

L257. Typo: “scenario” should be scenarios.

L290. The word “predictability” is misused here, since this refers to an intrinsic characteristic of a system. Please change this to predictive capacity or something similar.

Caption of Figure 6. “Note these probability density functions are derived using the conditional mean of the RF emulator (Appendix A) and do not include uncertainty arising from the emulator itself”. This seems to contradict the explanations provided in Sect. 2.4 (L269-270). Please verify if the emulator uncertainty is included or not.

L316. Please specify here relative to what the “relative differences” are computed. I believe that it is relative to the performance metrics computed from the validation test applied without leaving experiments out, but this should be 100% clear.

L330. Change “goes along” to: goes with.

L359. Change “turns to be worse” to: is worse.

L364. This should be: (...) than that of ‘woSSP585’ (...).

Caption of Figure 10. The word “quantile” in the last sentence should be plural.

L412. The word “significantly” should be replaced by substantially or a similar word. That is because, according to the error bars shown in Fig. 10, Q50 and Q83 are also significantly influenced, although the magnitudes are small.

L418. “(...) regardless of the GSAT change and the considered percentile”. I believe this is not true. See for example GSAT 2°, Q17, Narrow Kappa. Please consider revising this sentence.

L443-447. I agree with this analysis. However, it does not explain why woCISM has stronger impacts on performance at GSAT = 2° than at GSAT = 4°. In fact, I find this difference in woCISM influence somewhat surprising, given that the CDF seems more affected at high rather than low slc values (Fig. 11b). I would appreciate if the authors could attempt to explain this, or at least mention this aspect in the manuscript.

L454-456. This sentence is very unclear to me. I read it multiple times, but I cannot understand the message that the authors try to convey. Please rephrase.

L476. Please add the word estimated: the estimated contribution.

L533-536. I appreciate this more extensive discussion on the various types of uncertainties. To make this discussion more complete, I recommend mentioning briefly the influence of irreducible uncertainties on Greenland sea-level contribution projections (e.g., Verjans et al., 2025). It would be valuable to include a short statement on how such irreducible uncertainties can be addressed through the use of emulators.

L775. Please revise the notation of  $q^{\frac{1-\alpha}{2}}(slc|\mathbf{x}^*)$ ;  $q^{\frac{1+\alpha}{2}}(slc|\mathbf{x}^*)$ , by respecting the convention that  $\alpha$  represents the level, not the % of coverage (see comment about L247). Note also that the current notation is in disagreement with the notation of  $Q^{\frac{\alpha}{2}}$ ;  $Q^{1-\frac{\alpha}{2}}$  on L783.

L816. The word “score” in “CRPS score” is redundant, please remove it.

L823-828. Same comment about  $\alpha$  notation as for L247.

From the Supplementary Information.

p2, L7. This sentence does not make sense: “Overall the emulator is of moderate magnitude”.  
Figure S3. I recommend using the same x-axis for all sub-figures.

## References

V. Verjans, A. A. Robel, L. Ultee, H. Seroussi, A. F. Thompson, L. Ackermann, Y. Choi, and U. Krebs-Kanzow. The greenland ice sheet large ensemble (grisLens): simulating the future of greenland under climate variability. *The Cryosphere*, 19(9):3749–3783, Sept. 2025. ISSN 1994-0424. doi: 10.5194/tc-19-3749-2025. URL <http://dx.doi.org/10.5194/tc-19-3749-2025>.