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

This is my second review of this manuscript, following the first round of revisions. First, I note the positive aspects of the revisions.

- Using quantile random forest (qRF) regression is a great implementation. This allows to evaluate not only point estimates of emulator predictions, but also their range, which characterizes emulator uncertainty, and increases/decreases of this uncertainty across experiments.

- The authors have clarified that the Multi-Model Ensemble (MME) is taken from the study of Goelzer et al. (2025). Although the study of Goelzer et al. (2025) is still undergoing peer-review, once it has been validated and published, it will serve as a necessary foundation for the present work.

- The wording now includes more nuance about most of the results being specific to this particular MME, and to the particular emulator used.

- Numerous questionable aspects (e.g., previous D_{h} calculation, errors in figures, etc.) have been corrected.

On the other hand, I believe that the revised manuscript still has major shortcomings, in particular in terms of lack of clarity about the methods used, and interpretation of the results. I hope that my comments will help address these issues. My review is separated into three General comments, followed by Specific comments. Line numbers refer to the revised manuscript without tracked changes.

General comment 1: Arguments supporting the conclusions should be better grounded in the results

One of the main conclusions from this study is "the importance of having diverse ISM and RCM models" (L23). In particular, the performance analysis (Figs. 7, 8) shows the decrease in performance when excluding MAR (woMAR) or excluding CISM (woCISM). However, this does not necessarily argue for the importance of diversity in RCMs and ISMs. In my view, it only shows that extrapolation errors of the emulator to unseen RCMs and ISMs are high. For example, Q^2 relative differences are higher for experiment woCISM than experiment CISM (Figs. 7, 8), while the former includes 3 ISMs and the latter only includes 1. Thus, while experiment woCISM has a more "diverse" set of ISMs than experiment CISM, its performance is worse: this is in direct contradiction with the conclusion about importance of diversity. Instead, in my view, this worse performance is a consequence of woCISM only including 33.5% of the MME simulations versus 66.5% for CISM (Table 2). Thus, training on a smaller set of simulations is the likely cause of decreased performance.

More generally, one of the conclusions is that having different ISMs is beneficial, while training on just a few κ values is inconsequential. And I agree on this point. Let's take a very simple example of a very small MME to predict one unseen case.

- The study shows that for the task of emulating *slc* from the hypothetical configuration (ISM=Elmer, κ=0.5), it is more useful for the emulator to be trained on the 4-member MME set {(CISM, 0.1), (CISM, 0.9), (Elmer, 0.1), (Elmer, 0.9)} than to be trained on the 4-member MME set {(CISM, 0.1), (CISM, 0.5), (CISM, 0.9), (Elmer, 0.1)}.

- However, the study does not show that for predicting *slc* from the hypothetical configuration (GISM, 0.9), training the emulator on the 3-member MME set {(CISM, 0.1), (Elmer, 0.1), (IMAUICE, 0.1)} is more useful than training on the 3-member MME set {(CISM, 0.1), (CISM, 0.5), (CISM, 0.9)}, since GISM is absent from both training sets.

I hope that this example makes the point, despite its simplicity. The key is that the ISM to be emulated is present versus absent in the training set. Note that the exact same argument can be made concerning the "diversity" of RCMs instead of ISMs.

In short, what I try to communicate is that, despite the conclusions from the authors, there is no evidence that higher diversity leads unconditionally to better emulation performance. Instead, performance seems highly sensitive to the training set size, and also depends on the emulation target: it is easier to extrapolate to an unseen κ value than to an unseen ISM. This is not a surprise: the qRF cannot predict results from RCMs or ISMs that it has not seen during training (see General comment 3 below). But no evidence is given that this prediction capability increases with the diversity of RCMs/ISMs used during training, as long as the unseen RCM/ISM is not included. Yet, this is one of the key messages from the manuscript, or, at least, this is how it is communicated.

Therefore, the key takeaways should be grounded in metrics that directly support the specific message that is communicated. They should not be expressed as general conclusions if this generality has not been verified. Alternatively, stronger evidence should be provided to justify the general claims, accompanied by clear explanations grounded in quantitative results.

General comment 2: Lack of clarity regarding the evaluation procedure
I honestly have a lot of difficulties to understand exactly how the emulator performance has been evaluated (Sect. 2.4.1).

(a) Are the n_{test} test samples excluded from the emulator training in each of the 25 iterations of the validation procedure (i.e., are they truly unseen)? This needs to be explicitly specified.

(b) Is the validation procedure performed separately for each of the experiments described in Table 2? Fig. 7 suggests that this is the case, but it needs to be explicitly specified.

(c) Similarly, is the validation procedure performed using the full MME? Fig. 5 suggests that this is the case, but it needs to be explicitly specified.

(d) When validating the emulator for the specific experiments (shown in Fig. 7), are simulations excluded from both the training and test samples, or only from the training samples? For example, for evaluation of woMAR, are all MAR simulations excluded from the training and test samples or only from the training samples?

(e) In my previous review, I raised the question about why not using the traditional 10-fold cross-validation procedure rather than the ad-hoc validation procedure prescribed here. I am unconvinced by the response of the authors: "The reason for proposing an alternative validation procedure is to make sure to reflect the ability of the emulator to perform well over a wide range of GSAT values instead of randomly selected cases".
In 10-fold cross validation, each simulation of the MME is left-out exactly once. As such, the evaluation cases are not "randomly selected cases", they are all the cases. After performing cross-validation, it is easy to aggregate results according to their GSAT range, as is the case in Fig. 5. Note that I may be misunderstanding something here, which is why I ask the

authors to add one or two sentences of clear explanation for why their validation procedure is better "adapted to our objective" (L199).

(f) Figs. 7 and 8 show relative differences in performance. Is this relative to the metric for the exact same test sample as evaluated in the full-MME evaluation? In which case, it means that the 25×200 n_{test} test samples are shared across all the validation experiments? Or is it "relative" to some other quantity? This needs to be specified.

(g) It is nowhere specified if the evaluation process for the experiments (e.g., woMAR, MAR, woCISM, etc.) is with respect to the full distribution of results from the MME, or only to those specific simulations left out for the particular experiment (e.g., only the MAR simulations for woMAR, only the RACMO and HIRHAM simulations for MAR, only the CISM simulations for woCISM, etc.). As I understand it, the evaluation is with respect to the full MME distribution, but this needs to be explicitly specified.

Also, concerning the results from the evaluation procedure, more clarity or emphasis is required for some key points.

(h) Figs. 7 and 8 show very large relative differences in the metrics. In particular, the Q^2 relative difference often exceeds 100%. If I understand correctly, this implies that $Q^2 \leq 0$. This means that the emulator performs worse than simply predicting the mean as a constant output prediction. In other words, the emulator is worthless in such situations.
This is a critical point, which is completely omitted in the manuscript. I recommend that the authors draw a vertical line at the 100% value in Fig. 7b and Fig. 8b,e to emphasize this. They should also mention and discuss this in the main text.

(i) The authors correctly point out that "the RF emulator should be used cautiously over the range of GSAT values around 3°C" (L251). This is a critical point that should be discussed in the Synthesis and Discussion section, and mentioned in the Abstract (e.g., (...) the emulator performance is unsatisfactory at intermediate levels of warming (~ 3°C) (...)).

General comment 3: Lack of details concerning the emulator

In my previous review, I asked for more clarification about the emulator. I thank the authors for the additional information included in the manuscript, but I still believe that some critical details are omitted. I raise this as an important concern, because I believe that some of these aspects might have impactful consequences on the emulator results.

Taking the example of the woMAR experiment, the qRF is trained with RCM cases of HIRHAM and RACMO only. The qRF is constructed using decisions at each split, and some splits may separate based on HIRHAM versus RACMO. Then, at prediction time, if the emulator is tested with a MAR case, how can it make a prediction? In other words, how are unseen categories handled at prediction time?

For this reason, it seems strange to me to emulate *slc* from unseen RCMs or unseen ISMs. I take here two examples from recent emulation efforts of ISMIP6 where the emulator was not designed to predict output from unseen ISMs. First, Edwards et al. (2021) take the full ensemble as an emulation target. In contrast to this study, the emulator of Edwards et al. (2021) does not use the ISM as input to the emulator. Instead, their emulator is trained to predict the ensemble response across ranges of GSAT and κ values, not the response of a specific ISM (their nugget term accounts for inter-ISM differences). Second, Seroussi et al. (2023) emulate specific missing GCM-ISM experiments. However, they only emulated those missing experiments for which some other experiments using the target ISM were available. As such, inputting the ISM as a predictor

variable for a new prediction was actually meaningful, because it could be associated with samples from the training set. In summary, it is important to explain how unseen categories (e.g., ISMs, RCMs) are handled at prediction time, as this would partly illuminate the interpretation of the prediction performances shown in Figs. 7 and 8.

As a side note, I understand that the authors want to make the article as easy as possible to follow for non-experts, and as such move details to Appendices or omit them entirely. However, in my opinion, the level of technical detail in the main text is insufficient. For example, Sect. 2.2 reads more as an introductory paragraph to qRF regression rather than a description of the emulation process. Sects. 2.4.1 and 2.4.2 also lack the necessary detail to really understand the results shown in Figs. 5,6,7,8. I believe that the editor needs to agree that technical details are quasi absent in some important sections of the main text.

Specific comments

Title. Replace "future" by: 2100

L16-18. This sentence should end with a question mark.

L18-19. Specify: (...) to build a random-forest-based emulator of 2100 Greenland sea-level rise contribution (...).

L36-37. Please rephrase: one member cannot span, it is the MME that should span.

L56. In this paragraph, please also refer to the work of Seroussi et al. (2023), which is highly relevant to this study. L71-72. In this sentence, the word "experiments" is used twice to designate two different notions. This can be confusing.

L87-89. This statement needs a citation.

L95. "surface mass balance (SMB) changes" should be: surface mass balance (SMB) anomalies.

L123. Here and in the remainder of the manuscript, why is the wording "credibility interval" used instead of confidence interval? The former suggests that some Bayesian modeling has been performed. Please consider re-wording.

Caption of Fig. 2. Please specify the confidence interval corresponding to the likely range.

L133. Please rephrase because Elmer/Ice is not "more frequent than others".

L134. The minimum resolution of 16 km does not appear in Fig. 4. There is a bar at 20 km, and the most frequent seems to be at 8 km.

L177-179. At the end of this sentence, a brief sentence should be formulated to specify explicitly if the objective is then to evaluate if the emulators constructed from the reduced MME are capable of (i) reproducing the distribution of results from the original MME, or (ii) reproduce the results that have been left-out from the original MME.

L194-197. This last sentence of the 1st paragraph should be moved elsewhere. This Sect. 2.4.1 already includes very little details, so the space dedicated to it should be focused on explaining the performance evaluation procedure.

L201-203. "for each interval, 50 samples are randomly selected. For one iteration of the procedure, a total of n_{test}=200 test samples are randomly selected". I find this phrasing somewhat confusing. I recommend rephrasing: for each interval, 50 samples are randomly selected, resulting in a total of n_{test}=200 test samples.

L204. Specify if the emulator is trained in each iteration on all the MME simulations, except the n_{test} test samples (see General comment 2).

L205. "mean relative error" should be: mean relative absolute error.

L214. How many samples are drawn for the Monte-Carlo random sampling procedure?

Figure 5. The CRPS is a good metric, but not very intuitive (e.g., what does it mean if CRPS is 0.0025?). I believe that it would also be insightful to evaluate if the emulator is under-dispersed, over-dispersed, or well-calibrated. A common and intuitive metric for this is the spread-error ra-

tio (e.g., Stephenson and Doblas-Reyes, 2000). I think that it would add a lot to the analysis to quantify first the calibration of the emulator (in Fig. 5), and second if the emulator tends to become over- or under- dispersed in the experiments (Figs. 7 and 8). Note that since the qRF does not provide standard deviation of the prediction, the spread-error ratio can be approximated as $\frac{\sigma}{\mathrm{RMSE}} \approx \frac{Q_{75}-Q_{25}}{1.35\,\mathrm{RMSE}}$.

Caption of Figure 5. Please specify explicitly that the performance statistics are computed over test samples unseen during emulator training.

L239-240. Please specify here if this is performed using the full MME (in contrast to the reduced MME used for the experiments of Table 2).

Caption of Figure 6. Typo: constructed using the Monte-Carlo based procedure. Also, please specify the confidence interval corresponding to the likely range.

Caption of Figure 7. Please specify explicitly that the performance statistics are computed over test samples unseen during emulator training.

L273. Please remove "Interestingly", as it is preferable to let readers decide what they find interesting.

L278. "twice that of the third most important contributor, i.e., woCISM". The medians of RAE relative difference are very close. Please be more specific in quantification of the performance.

Caption of Figure 8. Please specify explicitly that the performance statistics are computed over the same test samples as in Figure 7.

Figure 8. These results show that (i) excluding SSP126 and SSP245 has negligible impact for predicting in the GSAT range ≥ 3.83°C, and (ii) excluding SSP585 has negligible impact for predicting in the GSAT range ≤ 2.14°C. This should be mentioned and discussed briefly in the text.

L285. Please rephrase: "has the largest impact almost at the same level".

L287. Mention to Table 1 is wrong.

L288. Please remove "it is interesting".

L289. "The analysis of the other GSAT intervals" should be: The analysis of the GSAT interval 3.34 to 3.83°C.

L295. Here, I believe that this analysis applies to the random samples drawn as explained in Sect. 2.4.2, and not to the test samples explained in Sect. 2.4.1. Please specify this explicitly.

L305. "under-estimated by more than 25%": Fig. 9 shows $\sim 22\%$.

L307. Please remove "Interestingly".

L307-311. How can this contrasting result be explained? A perfect performance would mean that all quantiles remain unchanged. As such, why do larger changes in quantiles do not lead to worse performance? The authors should explain this.

Caption of Figure 9. Please specify that these results are computed from the random samples as explained in Sect. 2.4.2 and not from the validation procedure (if I understood correctly).

Table 3 (row SSP-RCP). Please specify that "the strong linearity of the Greenland ice sheet response with global temperature" is valid for the 2100 timescale.

Table 3 (row ISM choice). I believe that the under- versus over-estimation depends on the specific ISM that is excluded. For example, if the CISM model predicts consistently higher *slc* values than other ISMs, then the experiment CISM would over-estimate the left-out *slc* values (as it is the case here). However, if the CISM model was predicting consistently lower *slc* values, then experiment CISM would lead to under-estimation of the left-out *slc* values. And reciprocally for the experiment woCISM. Therefore, I do not believe that such a general conclusion can be made about over- versus under-estimation (this links to General comment 1). Similarly in the Abstract, the word "under-estimations" may be misleading.

L335. "Here, 'woMAR' is not necessarily the highest contributor to the changes". Please explain this (see comment about L307-311).

L342. Please explain the reasons in the MME design that explain why woCISM leads to a larger perturbation of the member distribution than woMAR (e.g., experiments of a specific SSP scenario have only been done with the CISM model, etc.).
L345. "further work should look into this aspect in more detail". This should be done as part of this study.
L360. "This also relates to the question of initialisation (and initial mass loss estimates) where the RCM choice is a key ingredient (e.g., Otosaka et al., 2023)". It is unclear to me what this sentence implies, and which message the authors try to convey.
L363. "First, our study contributes (...) according to the same report". I do not see how this is a contribution of this study. Here, the authors simply provide the Greenland sea-level rise contribution estimates of the ICCP.
L390-392. "Indeed, scenarios based on global warming levels can be potentially better understood by stakeholders than the SSP or RCP scenarios, and also allow users to better make the link with the climate objectives set out in the Paris agreement to stabilize climate change well below 2°C GWL". This reads as a personal opinion of the authors, so please rephrase or remove.
L399. Please specify: future sea level by 2100.
L400. Please specify: high importance for emulator accuracy.
L417-421. This sentence is too long and confusing.
Equation A1. What does s represent in this equation?
L609. "where I(A) is the indicator operator". The parenthesis notation is not used in Equation A2.
L612 and L615. These two sentences repeat the same information.
L624. Please use consistently q_τ or q^τ.
Equation B1. Please define inf as the infimum function.
L628. Typo: "weighed".
L637. Please specify that Step 2.2 approximates the CDF of $slc|\tilde{x}$ (if I understand correctly).
L638. If I understand correctly, $Q_{\tilde{u}}^{\alpha}$ is not specific to a single $\tilde{u}$ sample, but depends on the full set of $\tilde{u}$'s sampled in Step 2.1. This is in contrast to $q^{\tilde{u}}\left(slc|\tilde{x}\right)$ in Step 2.2. If this is correct, then this notation is confusing, and I recommend writing $\tilde{Q}^{\alpha}$ instead.
L642. This "variability" corresponds to the emulator uncertainty about a given quantile level α. But, if I understand correctly, the range $\left[Q^{\alpha/2} : Q^{1-\alpha/2}\right]$ gives the $1-\alpha$ confidence interval of the emulator prediction for $slc|\tilde{x}$. If this is correct, please specify it.
L647. Please specify that the p-value quantifies how unlikely the variable importance in the non-permuted data is with respect to the null distribution of variable importance reached from the permutations.
L657. Typo: "should retained".
Figures S2, S3, and S4. Thes figures are identical. Is this an error?

References

Tamsin L. Edwards, Sophie Nowicki, Ben Marzeion, et al. Projected land ice contributions to twenty-first-century sea level rise. *Nature*, 593(7857):74–82, May 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03302-y. URL `http://dx.doi.org/10.1038/s41586-021-03302-y`.

Hélène Seroussi, Vincent Verjans, Sophie Nowicki, et al. Insights into the vulnerability of antarctic glaciers from the ismip6 ice sheet model ensemble and associated uncertainty. *The Cryosphere*,

17(12):5197–5217, December 2023. ISSN 1994-0424. doi: 10.5194/tc-17-5197-2023. URL http://dx.doi.org/10.5194/tc-17-5197-2023.

David B. Stephenson and Francisco J. Doblas-Reyes. Statistical methods for interpreting monte carlo ensemble forecasts. *Tellus A: Dynamic Meteorology and Oceanography*, 52(3):300, January 2000. ISSN 1600-0870. doi: 10.3402/tellusa.v52i3.12267. URL http://dx.doi.org/10.3402/tellusa.v52i3.12267.