Replies to Referees' 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 the two Referees for taking the time to participate in this third round of reviews. We agree with most of the suggestions and, therefore, we have modified the manuscript to take on board the comments and suggestions. We recall the reviews and we reply to each of the comments in turn (outlined in green). The page and line numbers are those of the document **with** track-changes.

In addition to the corrections suggested by the Referees, we have made some changes:
- corrections of typos and reformulations;
- correction of Fig. 4 where the upper bound of the likely range was corrected (it corresponded to the 95th percentile);
- clarification of the GSAT definition by specifying that we analyse GSAT change relative to 1995-2014.

Referee #1:

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.
We would like to thank Dr Verjans for taking the time to participate in this third round of reviews. We really appreciate the opportunity Dr Verjans has given to continue the revisions based on the valuable comments/suggestions.

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 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.
We thank Dr. Verjans for this suggestion.

We propose the following formulation in the abstract: "These results point to the size of the training set as the key driver of the changes, which supports the need for large ensembles to develop accurate and reliable emulators, hence encouraging large participation to projects such as the Ice Sheet Model Intercomparison Project ISMIP".

In the concluding remarks (on page 24, lines 538-544), we have also emphasised the same message as follows: "They also show that an ensemble designed only with a unique ISM and RCM model, i.e., here with the one that is most frequently selected in the considered MME, has non-negligible implications. These results point to the size of the training set as the key driver of the changes in the emulator performance and percentile estimates, hence underlying the need for building large ensembles to develop accurate and reliable emulators. Broad participation in projects such as ISMIP, with as many simulations as possible contributed by numerous groups, appears to be an effective option to this end."

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.
We have reformulated as follows: "We use these experiments to build a random-forest-based emulator, whose predictive capability to assess Greenland sea level rise contributions in 2100 proves very satisfactory for low and high levels of warming but less effective for intermediate levels."

L20. Typo: "warning".
This has been corrected.

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.
This is now specified as follows: "Quantified criterion changes are then used to rank the different emulator experiments in terms of the magnitude of their impact on emulator performance and emulator-based probabilistic predictions".

L247. Throughout the manuscript, I find the notations CA^{α} and PI^{α} potentially confusing. In statistical terminology, α typically denotes the significance level, so α = 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 α 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}.
We thank Dr. Verjans for this clarification. We have corrected the notation as suggested.

L248. Typo: "fall" should be falls.
This has been corrected.

L257. Typo: "scenario" should be scenarios.
This has been corrected.

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.
This has been replaced by predictive capability.

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.
We confirm that Fig. 6 has been built by using the RF mean, i.e., without including the emulator uncertainty. We have added elements on this aspect in Sect. 3.1 as follows: "The results are computed using the mean of the RF emulator (Appendix A), and do not include uncertainty arising from the emulator itself. The procedure described in Appendix B is further applied to assess the impact of the emulator uncertainty, and shows that the width of the 90% confidence interval for the percentiles considered remains in the order of 0.1 cm, hence indicating minor influence of the emulator uncertainty in this case".

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.
The introduction of Sect. 3.2, on page 14, lines 326-332 has been reformulated as follows: "We analyse in Figure 7 the impact of design decisions on the RF predictive capability and on the reliability of the RF prediction intervals. The decrease of RF predictive capability is measured by the decrease of the relative differences of *RAE* and *CRPS* (Fig. 7a,c) and the increase of the relative differences of Q^2 (Fig. 7b). The reliability of the RF prediction intervals is measured by CA^{90} and CA^{50} (Fig. 7d,e), which are respectively related to the prediction intervals at the 10% and 50% significance level, and by *IQR* (Fig. 7f). This assessment is conducted relative to the performance metrics of the reference solution computed from the validation test applied without excluding the experiments as explained in Sect. 2.4.1".

L330. Change "goes along" to: goes with.
This has been corrected.

L359. Change "turns to be worse" to: is worse.
This has been corrected.

L364. This should be: (...) than that of 'woSSP585' (...).
This has been corrected.

Caption of Figure 10. The word "quantile" in the last sentence should be plural.
This has been corrected.

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.
This has been replaced by "substantially".

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.
We have nuanced the remark by pointing out this exception. Note that most results for Kappa experiments have values on the same order of magnitude than the emulator uncertainty. This is also underlined.

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.
We recognize that Fig. 11 does not explain all aspects of the problem. We now clearly underline this in Sect. 4.1, page 22, lines 475-477 as follows: "Analysis of Fig. 11 reveals certain similarities in the effect of the different emulation experiments, but is not sufficient to explain all aspects of the problem; for example, this type of analysis does not fully explain why '*woCISM*' has a stronger impact on performance at GSAT change of 2°C than 4°C."

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.
We have reformulated (now lines 470-472) as follows: "This suggests that removing members associated with other ISMs / RCMs from the training set has an impact, because these members contain information relevant to the RF emulator capability to make predictions, especially in the situations explained in Sect. 2.3, for levels of categorical variables not seen in the training dataset".

L476. Please add the word estimated: the estimated contribution.
This has been corrected.

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.
We thank Dr. Verjans for this suggestion. We have added this aspect as follows: "To address this question, a wider range of uncertainties should be considered, more specifically model and structural uncertainties (i.e. uncertainty in the formulation of the model and its ability to represent the physics of the system), in addition to uncertainties in model parameters (related to ice dynamics and atmospheric/oceanic forcing), but also irreducible uncertainties such as

internal climate variability as investigated by Verjans et al. (2025) on Greenland sea level contribution projections".

Regarding the question of climate variability, we believe that the emulators could play a useful role to explore the space of climate forcings; more particularly because climate forcing is kept constant in ISMIP6. This should however be done in addition to the exploration of other uncertain factors / parameters. We have added a short statement in this sense as follows: "Here, emulators are expected to play a key role to explore this wide uncertain space thoroughly".

However, we have chosen to keep the statement relatively generic, because addressing this problem may potentially require additional developments, which are out of the scope of our study. Depending on the number of uncertainty sources, combination with adaptive sampling (e.g., Rohmer and Idier, 2012) may be required to improve the uncertain space exploration. The influence of aleatoric uncertainties may be viewed through the lens of stochastic simulation codes (e.g., Marrel et al., 2012), which may also require adapted emulators. These possible lines of further developments are only suggestions, and we believe that further analysis is here required.

References (for the replies)

Rohmer, J., & Idier, D. (2012). A meta-modelling strategy to identify the critical offshore conditions for coastal flooding. *Natural Hazards and Earth System Sciences*, *12*(9), 2943-2955.
Marrel, A., Iooss, B., Da Veiga, S., & Ribatet, M. (2012). Global sensitivity analysis of stochastic computer models with joint metamodels. *Statistics and Computing*, *22*(3), 833-847.

L775. Please revise the notation by respecting the convention that α represents the level, not the % of coverage (see comment about L247). Note also that the current notation is in disagreement with the notation of $Q^{\alpha/2};Q^{1-\alpha/2}$ *on L783*
We have corrected the notation as suggested in the comment for L247.

L816. The word "score" in "CRPS score" is redundant, please remove it.
This has been corrected.

L823-828. Same comment about α notation as for L247.
We have corrected the notation as suggested in the comment for L247.

From the Supplementary Information.

p2, L7. This sentence does not make sense: "Overall the emulator is of moderate magnitude".
This is replaced by "Overall the emulator uncertainty is of moderate magnitude."

Figure S3. I recommend using the same x-axis for all sub-figures.
This has been corrected.

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.
We thank Dr. Verjans for the suggested reference which has been added to the reference list.

Referee #2:

This is the third round of review for this paper. I would like to thank the authors for their responses to my comments and for the changes made to the manuscript.
We would like to thank Referee #2 for taking the time to participate in this new round of reviews. We really appreciate the opportunity Referee #2 has given to continue the revisions based on the valuable comments/suggestions.

Given that the authors have responded positively to my major comments, I recommend that the manuscript be finalized for publication. I still ask the authors to carefully reread their manuscript, as there are still quite a lot of typos in the text (see specific comments below). Provided the authors also respond favorably to the more technical comments addressed by the first reviewer, I would recommend this paper for publication.

Specific Comments

Note: I am using the author's tracked-changes document for the line numbers.

(1) There is inconsistency in the use of hyphens for compound nouns used as adjectives; for example, you use both 'sea level contributions' and 'sea-level rise'. Please stick to one writing style throughout the text.
We thank Referee #2 for noticing this problem. We have corrected by removing the hyphens.

(2) When using mathematical symbols, please use italics. This is particularly relevant for Appendix A ('p' → 'p', 'L' → 'L', ...).
We apologise for this problem. We now use italics for all symbols.

(3) Please avoid using fractions in running text. Either use a '/' symbol or write a full equation.
The notations have been corrected by writing full equations.

(4) You define slc in several places in the text and use it before it is properly defined. You should define it at the beginning of the manuscript once and for all.
The definition is now provided at the beginning of Sect. 2.1. We have also removed redundant definitions from the main text.

(5) [Line 17] 'specific set'→ 'a specific set' or 'specific sets'.
This has been corrected.

(6) [Line 20] 'warning'→ 'warming'.
This has been corrected.

(7) [Line 74] '([...] by Edwards et al. (2010))'→'([...] by Edwards et al., 2010)'.
This has been corrected.

(8) [Line 78] 'as well'→ 'as well as'.
This has been corrected.

(9) [Line 92] '(SSP126, SSP245, SSP585)'→'(SSP1-2.6, SSP2-4.5, SSP5-8.5)'.
This has been corrected.

(10) [Table 1] Follow-up on my previous comment (14): I am not asking to remove the categorical aspect in the column; rather, my suggestion was to use the symbols {a,b} and [a,b] to establish this distinction efficiently. Such notation is used in other multi-ensemble studies.
We thank Referee #2 for this clarification. We have corrected Table 1 using the suggested notations.

(11) [Line 148] '([...] by Aschwanden and Brinkerhoff (2022))'→'([...] by Aschwanden and Brinkerhoff, 2022)'.
This has been corrected.

(12) [Figure 4] Why is there an '(a)' at the beginning of the caption?
This is a mistake. This has been corrected.

Orleans,
November 7th, 2025
J. Rohmer[1] on behalf of the co-authors

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