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 new round of reviews. We really appreciate the opportunity the Referees and the Editor have given to continue the revisions based on the various comments/suggestions. This is underlined in the acknowledgement section. 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.

Referee #1:

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

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

In particular, we have taken care to:

- Provide sufficient details on the methods (Sect. 2.3 and 2.4), without using too technical description so that the article can be followed by non-experts, as suggested by Referee #2;
- Explain better how the RF emulator makes predictions with categorical inputs (Sect. 2.3) and add new performance scores;
- Nuance the conclusions regarding the "diversity" of the ISM and RCM model.

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 Dh 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

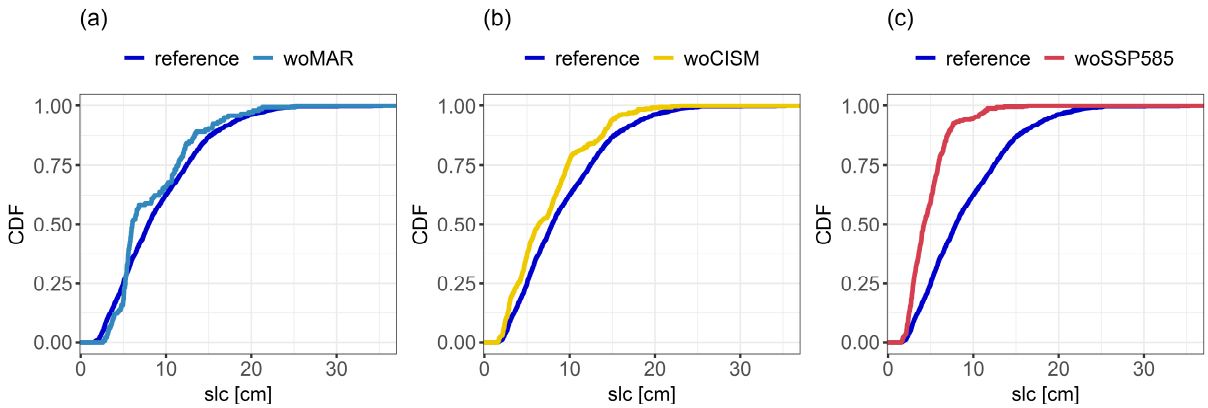
generality has not been verified. Alternatively, stronger evidence should be provided to justify the general claims, accompanied by clear explanations grounded in quantitative results.

We thank Dr Verjans for this useful analysis. This is also helpful for better framing our message.

We have now nuanced the message by specifying that the results are related to the considered MME and to the particular ISM / RCM model used in our study. These corrections have been applied in the abstract, in Sect. 4, and in Table 3.

We agree that the results do not rigorously demonstrate the added value of having diverse RCM or ISM models. Therefore, we propose to remove any reference to "diversity" and to focus the analysis on the results of the experiments and in particular CISM and MAR experiments and their implications.

The description in Sect. 4.1 (on page 21, lines 432-466) has been reworked as follows: "On the other hand, some other conclusions could not necessarily have been anticipated in detail more particularly the implications on the percentile assessment (Sect. 3.3). Our results show that the magnitude of the influence depends on the GSAT scenario considered, the performance criterion and the target percentile level. For the high GSAT scenario, the exclusion of SSP5-8.5 has as much impact as the exclusion of MAR on emulator performance, and is even the biggest contributor to changes in the high percentiles. For the low GSAT scenario, excluding CISM has as much impact as excluding MAR on the emulator performance, and contributes most to changes in the low percentiles. The decrease in MME size induced by 'woCISM' and 'woSSP585' is smaller than that induced by 'woMAR', on the order of 70%, suggesting that it is not only a problem of 'size' but also a problem of the type of information that is removed from the set. Figure 11c shows that, when applying 'woSSP585' experiment, the emulator is learned with *slc* spanning a restricted range lower than that of the original MME. This means that the emulator is built with little information on large *slc* values, and to predict cases associated to high GSAT scenarios, the RF model mainly relies on extrapolation. This is a situation where emulator methods such as RF can fail completely; see e.g., Buriticá & Engelke (2024). Analysis of Figs. 11a and b helps to understand why "woMAR" and "woCISM" induce roughly equivalent changes for the 2°C GSAT scenario, as the *slc* CDF appears to be similarly disrupted by the application of these experiments with a CDF shifted towards low-to-moderate *slc* values, particularly in the *slc* range of ~5 to ~15cm. This means that the emulators are built on members whose *slc* values span approximately the same range.



New Figure 11: Comparison between the Cumulative Distribution Function (CDF) of *slc* in 2100 of the original MME (reference) and of the reduced MME after application of the emulator experiments, 'woMAR' (a), 'woCISM' (b), 'woSSP585' (c).

The oppositive experiments that consist in using MME restricted to members to a specific ISM or a particular RCM, here CISM or MAR respectively, are also informative. Although the corresponding emulator experiments imply a reduction of less than 30% of the MME size, the decline in emulator performance or changes in percentiles cannot be considered negligible. Our interpretation is that this effect is related to the importance of the information removed, i.e., the configurations of all input variables associated with the removed members, which is necessary for the RF emulator to make predictions for levels of categorical variables not seen in the training dataset, i.e., ISMs/RCMs, as explained in Section 2.3.

The interaction between the reduction in the size of the MME and the type of information important for the training of the emulator is however complex to analyse due to the multiple joint effects to be taken into account between the inputs. From a methodological viewpoint, this calls for further developments, in particular by relying on the data valuation domain (Sim et al., 2022). These types of tools aim to study the worth of data in machine learning models based on similar methods as the ones used by Rohmer et al. (2022) in the context of sea level projections. Transposed to the MME context, these tools could be used in future studies to assess the impact of each member in the emulator's predictions, i.e. the worth of each member. From a broader perspective on collaborative research, these results on the influence of RCM and ISM models can be seen as an additional justification for intensifying the model intercomparison efforts initiated in the past, in particular ISMIP6 (Nowicki et al., 2016), which included coupled ISMs as well as stand-alone ISMs in CMIP for the first time. They also support, to some extent, a posteriori, the choices that have been made for the construction of the MME considered here (based on Goelzer et al., 2020)".

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 ntest 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 ntest 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.

We reply here to comments (a) – (g). To clarify the structure, we have added more details about the implementation of the validation procedure for testing the emulator performance (Sect. 2.4.1, on pages 10-11, lines 216-254) and about the implementation of the Monte-Carlo procedure for estimating the quantiles (Sect. 2.4.2, on pages 11-12, lines 255-268).

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.

We thank Dr. Verjans for this useful comment. As suggested, we have added comments on this aspect in the description of the results in Sect. 3.2 as well as in the caption of Fig. 7 and 8.

(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) (...)).

We have specified in the abstract (on page 1, lines 18-21) that "We use these experiments to build a random-forest-based emulator, which shows high predictive capability for assessing 2100 Greenland sea-level rise contributions for low and very high levels of warning".

We have also identified more clearly this aspect as a line for improvement in Sect. 5 (on page 25, lines 537-543) as follows: "Second, our results are based on the use of an emulator, i.e., a statistical approximation of the 'true' chain of numerical models. The RF emulator trained in our study showed satisfactory predictive capabilities for low and high levels of warning (GSAT values of respectively 2 and 4°C). The emulator performance remained however unsatisfactory at intermediate levels of warming (3°C). Despite, the efforts made in our study to nuance the results by including indicators of the emulator uncertainty, the emulator training should be

improved in the future by considering alternative emulator models (e.g., Yoo et al., 2025) but also more robust approaches for hyperparameter tuning (Bischl et al., 2023), and more particularly more advanced categorical variables' encoding (Au, 2018; Smith et al., 2024), which is key to apply the proposed emulator experiments".

Added references

Au, T. C.: Random forests, decision trees, and categorical predictors: the" absent levels" problem. Journal of Machine Learning Research, 19(45), 1-30, 2018.
Smith, H. L., Biggs, P. J., French, N. P., Smith, A. N., and Marshall, J. C.: Lost in the Forest: Encoding categorical variables and the absent levels problem. Data Mining and Knowledge Discovery, 38(4), 1889-1908, 2024.

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.

As rightly pointed out, we are striving here to satisfy two contradictory (but justified!) demands, namely (1) to provide sufficient technical detail so that readers have all the information they need to analyse and understand the results; (2) not to overload the text with overly technical aspects that could hinder readability for non-experts and obscure the practical implications of the work. Therefore we proposed having technical appendices in the previous version.

On the one hand, we partly agree with Dr. Verjans regarding the description of performance scores. We believe that indicating the interpretation of scores rather than equations in the main text should serve our twofold objective.

On the other hand, we totally agree with Dr. Verjans regarding the problem of prediction for unseen categories. This is problem intrinsically connected to the use of random forest models, known as the "absent level" problem (Au, 2018). We propose to clarify this problem in the main text in a relatively non-technical way. To do so, we have re-organized Sect. 2, by first defining the 'emulator experiments' related to the design decisions, then by describing in more details the functioning of the random forest model, and the problem of prediction for unseen categories in relation to the emulator experiments.

Sect. 2.3 (page 9-10, lines 190-202) has been updated as follows: "A key aspect of our study is to be able to handle many categorical variables with large number of levels (unordered values). However, the partitioning algorithm described above tends to favour categorical predictors with many levels (Hastie et al. (2009): chapter 9.2.4). To alleviate this problem, we rely on the computationally efficient algorithm proposed by Wright and König (2019) based on ordering the levels *a priori*, here by their *slc* mean response. A second key aspect is to be able to predict for new levels of the categorical variables, since the emulator experiments defined in Sect. 2.1 involve leaving out specific members from the original MME assigned to a given model, RCM / ISM, or a given SSP-RCP scenario, i.e., some specific levels. This problem is related to the more general 'absent levels' problem for RF models (Au, 2018), which arises when a level of a categorical variable is absent when a tree is grown, but is present in a new observation for prediction. Here, the chosen ordering algorithm of Wright and König (2019) alleviates this problem: by treating the categorical variables as ordinal, levels not present at a given partition during the splitting procedure can still be assigned to a next partition in the next iteration by directing all observations with absent levels down the same branch of the tree (in our implementation, chosen as the "left" branch). In this manner, the observations with absent levels are kept together and can be split down the tree by another input variable".

Dr. Verjans is right to underline that our experiments test the extrapolation capability of the RF model by using it for predicting new categories. However, it is important to note that this type is not so "strange" because this is not a "pure" extrapolation problem: the members that are removed from the training by applying the emulator experiments share information, with those used for training via the other variables. The RF model relies on them to make the prediction as explained above. In our study, this means that the emulator experiments test whether the information left in the MME after removing specific members is sufficient to predict *slc* at a reasonable accuracy (Sect. 2.3, page 10, line 203).

In conclusion, we do not claim to overcome the difficult problem of "absent levels", which is inherent to the RF method. Therefore, we use a standard procedure implemented in several packages for RF modelling. Other options exist (Au, 2018; Smith et al., 2024), and we have clearly pointed out the need for investigating them as an avenue of this work in the concluding remarks (Sect. 5, page 25, lines 543); see also reply to comment (i) described above.

Added references

Au, T. C. (2018). Random forests, decision trees, and categorical predictors: the" absent levels" problem. *Journal of Machine Learning Research*, *19*(45), 1-30.

Smith, H. L., Biggs, P. J., French, N. P., Smith, A. N., & Marshall, J. C. (2024). Lost in the Forest: Encoding categorical variables and the absent levels problem. *Data Mining and Knowledge Discovery*, *38*(4), 1889-1908.

Specific comments

Title. Replace "future" by: 2100
This has been corrected.

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

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

L36-37. Please rephrase: one member cannot span, it is the MME that should span.
This has been corrected as suggested.

L56. In this paragraph, please also refer to the work of Seroussi et al. (2023), which is highly relevant to this study.
We agree with Dr. Verjans. We now refer to this study in the introduction (page 3, lines 66-68).

L71-72. In this sentence, the word "experiments" is used twice to designate two different notions. This can be confusing.
This has been reformulated on page 3 (lines 74-78) as follows: "To address these questions, we take advantage of a large MME of Greenland ice sheet contributions to sea level this century, based on which we define a series of numerical experiments (referred to as emulator's experiments) that are closely related to practical MME design decisions. These experiments consist in leaving out specific results from the original MME assuming that all members have the same weight in the ensemble."

L87-89. This statement needs a citation.
We have added references to (Slater et al., 2019, 2020) and Rahlves et al. (2025).

Added reference

Rahlves, C., Goelzer, H., Born, A., and Langebroek, P. M.: Historically consistent mass loss projections of the Greenland ice sheet, The Cryosphere, 19, 1205-1220, https://doi.org/10.5194/tc-19-1205-2025, 2025.

L95. "surface mass balance (SMB) changes" should be: surface mass balance (SMB) anomalies.
This has been corrected.

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.
We agree with Dr. Verjans. There is no notion of Bayesian modeling and we now refer to "confidence intervals".

Caption of Fig. 2. Please specify the confidence interval corresponding to the likely range.
This has been corrected.

L133. Please rephrase because Elmer/Ice is not "more frequent than others".
This has been corrected.

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.
This is a problem with the endpoint of the y-axis. This has been corrected. We thank Dr. Verjans for noticing this problem.

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.
This part (Sect. 2.2, page 9, lines 170-175) has been reformulated as follows: "To measure the influence of removing specific members from the original MME, we assess if the emulators constructed from the reduced MME are capable of reproducing the results of an emulator trained with the complete original MME, named the 'reference solution' in the following. We analyse the changes in two types of criteria: (1) emulator performance to predict *slc* in 2100 for input configurations unseen during the training; (2) probabilistic predictions for *slc* in 2100 given future GSAT change scenarios, here chosen at 2°C (+/- 0.5°C) or 4 °C (+/- 0.5°C). The details of this assessment are explained in Sect. 2.4".

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.
Section 2.4.1 has been fully re-written by providing more details on the performance evaluation. We chose to keep the comment on GSAT, because it is the motivation for the development of an evaluation procedure adapted to our objective. We hope that the description is now clearer.

L201-203. "for each interval, 50 samples are randomly selected. For one iteration of the procedure, a total of ntest=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 ntest=200 test samples.
This part has been reformulated as suggested.

L204. Specify if the emulator is trained in each iteration on all the MME simulations, except the ntest test samples (see General comment 2).
The description of the procedure at the beginning of Sect. 2.4.1 clarifies this aspect.

L205. "mean relative error" should be: mean relative absolute error.
This has been corrected.

L214. How many samples are drawn for the Monte-Carlo random sampling procedure?
A total of 10,000 random samples are considered. This is now more clearly indicated in Sect. 2.4.2 (page 12, line 262).

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 ratio (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 σ/RMSE ≈ Q75−Q25/1.35 RMSE.

We thank Dr Verjans for his valuable suggestion. Building on this idea of calibration, we propose two additional performance scores:

- The first criterion analyses the coverage of the prediction intervals at the level α. If the coverage reaches the expected value of α, the prediction interval can be considered reliable;
- The spread to error ratio with a formulation using the interquartile distance as proposed by Dr. Verjans.

We have however a concern with the equation proposed. The correction by 1.35 is valid assuming a normal distribution of the predictive distribution provided by the RF emulator. This is not necessarily the case here. In addition, we can find different formulations in the literature. For instance, Bellon-Maurel et al., (2010) proposes a formulation without correction.

We propose to use the formulation without correction and to clearly describe the interpretation used in our study (Sect. 2.4.1, page 11, lines 250-254): "the ratio of performance to the interquartile distance *IQR* (Bellon-Maurel et al., 2010) compares the emulator prediction uncertainty, measured by the difference between the 75th and the 25th quantiles - named interquartile distance, with the prediction error measured by the root mean square error. If $IQR \approx 1$, the interquartile distance provides valuable information about the prediction error. If $IQR < 1$ (>1), this means that the emulator prediction uncertainty under-(over-)estimates the prediction error, i.e., the emulator provides over-(under-)confident predictions".

In addition, we have included a new reference in Appendix D for the formal description of the *CRPS* score, i.e., Bracher et al. (2021).

Added reference

Bracher, J., Ray, E. L., Gneiting, T., & Reich, N. G., 2021. Evaluating epidemic forecasts in an interval format. PLoS computational biology, 17(2), e1008618.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M. and McBratney, A.: Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC Trends in Analytical Chemistry, 29(9), 1073-1081, 2010.

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

This has been specified.

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).

This has been specified. The new description of the procedure in Sect. 2.4.1 and 2.4.2 should help to clarify as well.

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

This has been corrected. The specification has also been added.

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

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

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.
The description of the results has fully been revised.

Caption of Figure 8. Please specify explicitly that the performance statistics are computed over the same test samples as in Figure 7.
This has been specified as well as in the new Fig. 9.

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.
This has been added on page 17, line 367.

L285. Please rephrase: "has the largest impact almost at the same level".
This sentence has been modified.

L287. Mention to Table 1 is wrong.
This has been corrected.

L288. Please remove "it is interesting".
This has been removed.

L289. "The analysis of the other GSAT intervals" should be: The analysis of the GSAT interval 3.34 to 3.83◦C.
This has been reformulated.

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.
This has been clarified in Sect. 3.3, page 18, line 388.

L305. "under-estimated by more than 25%": Fig. 9 shows ∼ 22%.
We thank Dr Verjans for noticing this inconsistency. This has been corrected.

L307. Please remove "Interestingly".
This has been removed.

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.
We recognize that this statement is too simplistic, and we have now improved our analysis. Our results in Sect. 3.2 (page 17, Fig. 9 and 10) show that there are two main drivers of the performance depending on the GSAT scenario considered, i.e., 'woCISM' and 'woMAR' at low GSAT and 'woSSP585' and 'woMAR' at high GSAT. This means that there is no reason to opposite the two analyses, performance (Sect. 3.2) and percentile changes (Sect. 3.3), and there are some consistencies in the results.

We agree with Dr. Verjans that a perfect performance would mean that all quantiles remain unchanged. However, the opposite situation is not necessarily always valid, i.e., having poor performance scores do not necessarily mean that all levels of percentile are affected in the same way. The analysis of the percentile changes highlights this aspect. This is now clarified in Sect. 3.3 and in Sect. 4.1. Please refer also to our reply to comment 1.

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).
This has been specified.

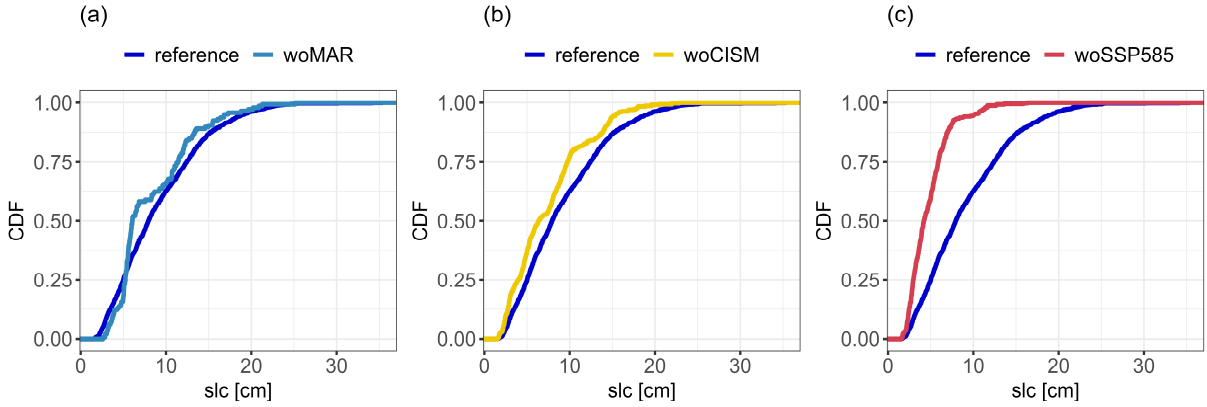
Table 3 (row SSP-RCP). Please specify that "the strong linearity of the Greenland ice sheet response with global temperature" is valid for the 2100 timescale.
This has been specified.

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.
We agree that the relation to the considered MME and the particular RCM or ISM model used should be made clearer. We propose a new formulation in Table 3 as follows: "Excluding the most frequently selected ISM in the considered MME, i.e., CISM, has a significant impact on emulator performance and percentile values with a more pronounced effect for low GSAT values. The opposite situation, i.e., limiting to CISM, leads to changes of lower magnitude."
The term "under-estimation" has been removed from the abstract.

L335. "Here, 'woMAR' is not necessarily the highest contributor to the changes". Please explain this (see comment about L307-311).
We have improved the presentation of the results. Please refer also to the reply to comment 1 above. As underlined in Sect. 4.1, page 21, from line 433: "Our results show that the magnitude of the influence depends on the GSAT scenario considered, the performance criterion and the target percentile level. For the high GSAT scenario, the exclusion of SSP5-8.5 has as much impact as the exclusion of MAR on emulator performance, and is even the biggest contributor to changes in the high percentiles. For the low GSAT scenario, excluding CISM has as much impact as excluding MAR on the emulator performance, and contributes most to changes in the low percentiles".

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.).

We have provided an additional analysis of the *slc* CDFs in Sect. 4.1, on page 21, from line 438 as follows: "The decrease in MME size induced by 'woCISM' and 'woSSP585' is smaller than that induced by 'woMAR', on the order of 70%, suggesting that it is not only a problem of 'size' but also a problem of the type of information that is removed from the MME. Figure 11c shows that, when applying 'woSSP585' experiment, the emulator is learned with *slc* spanning a restricted range lower than that of the original MME. This means that the emulator is built with little information on large *slc* values, and to predict cases associated to high GSAT scenarios, the RF model mainly relies on extrapolation. This is a situation where emulator methods such as RF can fail completely; see e.g., Buriticá & Engelke (2024). Analysis of Figs. 11a and b helps to understand why "woMAR" and "woCISM" induce roughly equivalent changes for the 2°C GSAT scenario, as the *slc* CDF appears to be similarly disrupted by the application of these experiments with a CDF shift towards low-to-moderate *slc* values, particularly in the *slc* range of ~5 to ~15cm. This means that the emulators are built on members whose *slc* values span approximately the same range.



New Figure 11: Comparison between the Cumulative Distribution Function (CDF) of *slc* in 2100 of the original MME (reference) and of the reduced MME after application of the emulator experiments, 'woMAR' (a), 'woCISM' (b), 'woSSP585' (c).

L345. "further work should look into this aspect in more detail". This should be done as part of this study.

Although informative, the CDF analysis described above is not sufficient to fully answer the question. We believe that quantifying the importance of group of members is related to the information removed when applying the experiment, i.e., to which extent the configurations of the input variables associated with the corresponding members are valuable for the RF emulator to make predictions for ISM/RCM unseen in the training dataset as explained in more details in Sect. 2.3. This is underlined in Sect. 4.1, page 22, lines 454-456.

We have underlined in Sect. 4.1 (page 22, lines 457-462), a line for further work as follows: "The interaction between the reduction in the size of the MME and the type of information important for the training of the emulator is however complex due to the multiple joint effects to be taken into account between the inputs. From a methodological viewpoint, this calls for further developments, in particular by relying on the data valuation domain (Sim et al., 2022). These types of tools aim to study the worth of data in machine learning models based on similar methods as the ones used by Rohmer et al. (2022) in the context of sea level projections. Transposed to the MME context, these tools could be used in future studies to assess the impact of each member in the emulator's predictions, i.e. the worth of each member".

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.
We agree that this sentence is outside the scope and have decided to delete it.

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.
We agree that this point was made not sufficiently explicit. We have added in Sect. 4.2 (page 23, lines 480-484) the two following sentences to explain how our study can contribute to a better understanding on the Greenland ice sheet melting to sea-level rise: "Here, we showed that some choices made by modelers, such as the tidewater glacier retreat parameter, have a minor impact on the spread of the Greenland sea-level rise contribution, whereas others, such as using only MAR as a regional climate model, have a large impact. These findings can be useful to inform future modelling experiments, and could help identifying where modelling efforts could focus to better characterize the spread of the projected contribution of the Greenland ice-sheet and to increase our understanding of that spread."

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.
We thank Dr Verjans for this comment. We have clarified that this statement is not a personal opinion but a choice made by adaptation decision makers in at least one country, i.e., France. We have added a clarification in Sect. 4.2 (page 24, lines 506-510) as follows: "For example, the latest adaptation plan in France requires adaptation practitioners to test their adaptation measures against a climate change scenario reaching 2°C in 2050 and 3°C in 2100 globally (Le Cozannet et al., 2025). Motivations for considering these GWLs rather than SSP or RCP scenarios include their perceived clarity for a wide range of adaptation practitioners, as well as the direct links that can be made with the climate objectives set out in the Paris agreement to stabilize climate change well below 2°C GWL."

L399. Please specify: future sea level by 2100.
This has been added.

L400. Please specify: high importance for emulator accuracy.
Results show that this importance is for both situations, emulator performance and percentile estimates. This has now been specified.

L417-421. This sentence is too long and confusing.
The sentence in Sect. 5 (on page 25, lines 545-547) has been simplified as follows: "This could be done iteratively. The procedure could alternate between simulation phases, i.e. either test simulations to assess sensitivity to different inputs, or small exploratory sets that do not use all the available computing time/human/project resources, and training and retraining of the emulator."

Equation A1. What does s represent in this equation?
This is mistake. This has been removed.

L609. "where I(A) is the indicator operator". The parenthesis notation is not used in Equation A2.
This has been corrected.

L612 and L615. These two sentences repeat the same information.
We have removed Line 612.

L624. Please use consistently q_τ or q^τ .
This has been corrected.

Equation B1. Please define inf as the infimum function.
This has been specified.

L628. Typo: "weighed".
This has been corrected.

L637. Please specify that Step 2.2 approximates the CDF of slc|˜x (if I understand correctly).
Dr Verjans is correct about that. We have specified it.

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 $q^{\tilde{u}}(slc|\tilde{\boldsymbol{x}})$ in Step 2.2. If this is correct, then this notation is confusing, and I recommend writing $\tilde{Q}^{\alpha}$ instead.
The original intent was to indicate the dependence to the full set of $\tilde{u}$'s sampled in Step 2.1. From Dr Verjans's comment, we have feeling that its adds more confusion, and we have removed this notation and used the one proposed.

L642. This "variability" corresponds to the emulator uncertainty about a given quantile level α. But, if I understand correctly, the range $[Q^{\alpha/2}; Q^{1-\alpha/2}]$ gives the $1-\alpha$ confidence interval of the emulator prediction for $(slc|\tilde{\boldsymbol{x}})$. If this is correct, please specify it.
This has been specified when describing step 2.3.

L647. Please specify that the p-value quantifies how unlikely the variable importance in the nonpermuted data is with respect to the null distribution of variable importance reached from the permutations.
This has been specified.

L657. Typo: "should retained".
This has been corrected.

Figures S2, S3, and S4. Thes figures are identical. Is this an error?
We confirm that this is not an error. We agree that with this type of representation, little differences can visually be seen. We now propose a new representation with CDFs and a zoom on one case to better highlight the impact of the emulator 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.

Hélene 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.

Referee #2:

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

This article is a revised version of a previously submitted paper. While some changes have been made to address my comments –for which I would like to thank the authors– I do not believe that the current version is ready for publication, as it does not yet meet the standards of clarity and readability required for publication. I continue to believe that the content of this article is noteworthy and valuable for the scientific community, but further work is needed to make it publishable.

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 various comments/suggestions.

In particular, we have taken care to:

- Provide sufficient details on the methods (Sect. 2.3 and 2.4), without using too technical description so that the article can be followed by non-experts;
- Underline the presence of other sources of uncertainty;
- Improve the quality of the figures.

My review is structured in two parts. First, in the general comments section, I address the major changes made by the authors in response to my original remarks. Then, I provide a list of specific comments.

These are mostly minor suggestions and corrections, but they are relatively numerous. In particular, I have not gone through the entire text in detail. To facilitate the next round of revisions, I would therefore ask the authors to carefully proofread their revised manuscript, which would help streamline the revision and correction process. As a side note, the article would be more pleasant to read if the figures were included in vector format or, at least, in higher resolution, so that they do not appear blurry.

We apologize for this problem. This is due to the word-to-PDF conversion of the preprint. We have now converted with a higher quality of the figures. In addition, we have uploaded a zip file with all figures on the github https://github.com/rohmerj/MMEdesign/blob/main/Figures.zip. Note that the Copernicus platform does not allow to upload the figures as separate files.

General comments

General comment 1. I would like to thank the authors for their efforts to reorganize the paper to make it more accessible to a broader audience.

We appreciate this positive feedback. Please note however that Referee #1 asks for detailed technical description, and we are striving here to satisfy two contradictory (but justified!) demands. Most of the technical details are kept in Appendices but additional technical descriptions have been added in Sect. 2.3 to clarify how the RF model makes predictions as well in Sect. 4.1 to provide more details on the validation procedure.

General comment 2. I am pleased to see that a brief discussion on the weight attribution of each ensemble member is now included in the paper. However, unless I am mistaken, this appears primarily in the interpretation of results, and thus appears rather late in the main text (it is first mentioned on page 8). Evaluating the predictive quality of the members is crucial in the context

of future ensemble studies; indeed, it is a key step in linking numerical simulations to observations and constraining the former using the latter. It is also a quite natural step from a Bayesian approach, as it allows for relaxing the assumption of a uniform prior. Therefore, it would make sense to introduce this question earlier in the main text, so that the reader clearly understands how each member is to be compared. I suggest adding such a discussion when the inputs are presented in subsection 2.1, at the end of page 5. The authors could also briefly mention it in the introduction when outlining the scope of the paper.
We thank Referee #2 for this comment. We agree that our assumption of the weight attribution should be better underlined. Therefore, we have specified it in Sect. 2.1 on page 7, lines 146-149 as follows "In this study, we assume that each member has the same weight, in particular, without differentiating members based on their reliability (e.g., low-resolution models compared with high-resolution models) or any observational constraints (as done for instance by Aschwanden and Brinkerhoff (2022)). Under this assumption of uniform weighting, […]".

As suggested, we have also briefly introduced this aspect in the introduction on page 3 in line 76-77: "[…] we define a series of numerical experiments (referred to as emulator's experiments) that are closely related to practical MME design decisions consisting in leaving out specific results from the original MME assuming that all members have the same weight in the ensemble".

We recall also this assumption in the description of the emulator experiments in Sect. 2.2 (page 9, line 168).

Finally, we mention a weighting approach as an avenue of this work in Sect. 5 on page 24 in lines 529-533.

General comment 3. I would like to thank the authors for the additional details concerning the interpretation of the parameter κ. While I am not entirely convinced by the rationale of comparing a single parameter to a full forcing scenario choice, I am fine with the authors retaining this aspect of their analysis.
I still wonder whether the meaning of this parameter could be made clearer by explicitly renaming it as 'ocean forcing' parameter instead of 'retreat' parameter, particularly if it is associated with uncertainties in ice–ocean coupling, rather than being a parameter intrinsic to the ice-sheet model itself.
This would make it clearer that it represents a forcing.
We thank Referee #2 for this comment. We appreciate the point of Referee #2, and understand the reasons for the possible confusion. However, we would like to underline that it is called "retreat" parameter and "retreat" parameterization in Edwards et al. (2021) and Rahlves et al., (2025) and above all it is also how in the community the modelers refer to it since ISMIP6. For sake of consistency, we think it is clearer to keep it as is.

Furthermore, it seems to me that the structural uncertainty of the ice-sheet models has not been addressed in the paper. This omission might bias the results by underestimating the impact of uncertainties related to ice-sheet physics and models.
We agree that this aspect deserves further investigation; We have indicated it in the conclusions on pages 25, lines 533-536, 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)".

Specific Comments

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

(1) [Lines 32–33] '(IPCC: e.g. Lee et al., 2021)'→'(IPCC; e.g., Lee et al., 2021).

This has been corrected.

(2) [Line 36] '(e.g. Knutti et al., 2010)'→'(e.g., Knutti et al., 2010)'.

This has been corrected.

(3) [Line 37] The fact that each member evenly spans a representative set of plausible realizations is somewhat misleading, as this is only true if no additional information is not available (e.g., observations). This would benefit from further clarification; see general comment 2.

Please see our reply to comment 2.

(4) [Line 40] '(e.g. Merrifield et al., 2020)'→ '(e.g., Merrifield et al., 2020)'.

This has been corrected.

(5) [Lines 52–55] This sentence is too long. Consider splitting it in two, maybe after the 'thoroughly'.

This has been split into two sentences as suggested.

(6) [Line 82] 'the main model parameter' is ambiguous: do you mean κ?

This has been reformulated as "[…] the retreat parametrisation described below"

(7) [Line 94] Consider adding what you mean by 'as best as possible', as it is vague on its own. Maybe mention that the misfit between computed and observed surface velocities and/or thicknesses is minimized?

We agree that this may be read like a vague statement, but it is vague on purpose. It goes back to the ISMIP6 protocol, which was intentionally very open to accommodate a range of different modelling approaches. It was really up to the individual modellers to interpret that as they saw fit. We would like to keep the original formulation.

(8) [Line 98] '(Slater et al., 2020, 2019)'→'(Slater et al., 2019, 2020)'.

This has been corrected.

(9) [Line 108] 'parameter values' is ambiguous: do you mean the ice-sheet parameter values?

This has been corrected.

(10) [Lines 109–112] I would remove entirely the discussion about the merge of the two inputs and present your final setup more directly, namely, the use of a GSAT that corresponds to a combination of SSP-RCP and GCM. This would be easier to follow.

We have simplified the presentation by specifying in the main text that: "The inputs below the double line in Table 1 are those used for the building of the RF emulator, in particular with the use of global annual mean surface air temperature change relative to 1995-2014, denoted GSAT, that corresponds to a combination of SSP-RCP and GCM by following a similar approach as Edwards et al. (2021)".

(11) [Line 114] 'The inputs from the double line' is not clear. Do you mean the inputs below the double line?
This has been corrected.

(12) [Table 1] Please be consistent in your system of notations. Some of the names start with capital letters (e.g., 'Sliding'), others do not (e.g., 'thermodin.').
This has been corrected.

(13) [Table 1] 'thermodin'→'thermodyn' or even 'thermo'.
This has been corrected.

(14) [Table 1] One way to simplify the reading of the table would be to use math symbols to clarify whether the variables are categorical or continuous. For example, the ISM models would become {CISM, Elmer/Ice, GISM, IMAUICE}, while the resolution would become [1, 40] km.
We thank Referee #2 for this suggestion. As pointed by Referee #1, the treatment of categorical variables is key in our study, and we believe that Table 1 should specify this aspect explicitly. Therefore, we choose to keep this column.

(15) [Line 118] 'as a particular'→ 'a particular'.
This has been corrected.

(16) [Line 121] 'expressed in meters sea level equivalent SLE'→'expressed in meters sea level equivalent, SLE'.
This has been corrected.

(17) [Figures 2–4] I think it would make more sense to first present the inputs (Figures 3 and 4) before the output (Figure 2).
We thank Referee #2 for this suggestion. This improves readability.

(18) [Figure 2] Vertical label: 'density'→ 'PDF'.
This has been modified.

(19) [Figure 2] Caption: 'Probability density function of the sea level contribution in 2100 (with respect to 2014) from the Greenland ice-sheet (in cm seal level equivalent, SLE) based on the raw MME ensemble data considered in this study' → 'Probability density function of the sea level contribution of the Greenland ice sheet in 2100, with respect to 2014, based on the raw MME ensemble data considered in this study'.
We thank Referee #2 for the suggestion. This has been corrected.

(20) [Line 132] 'highest importance' is ambiguous. Do you mean 'that contributes the most to the uncertainty'?
This has been reformulated as suggested.

(21) [Line 137] Consider adding one sentence that quickly explains why the design of experiments was indeed unbalanced.
We thank Referee #2 for this comment. A sentence has been added (Sect. 2.1, page 6, lines132-134), namely that "this parameter was sampled for only 3 different values by most models (the median, the 25% and the 75% percentile), and the additional 2 values were only sampled by one ISM at a later stage to broaden the parameter range".

(22) [Figures 3, 4, C1 and C2] Consider drawing the plots with the count number in the y axis, as that is more common.
This has been modified for Figures 3, 4 (now Figures 2 and 3) as well for Figure C2. We chose not to change Fig. C1 because of the long list of variables to be specified on the axis.

(23) [Line 154] '(named emulator)' is not necessary here as you have already introduced several times the notion of emulator previously.
This section has been reformulated, and the definition of emulator is in the introduction.

(24) [Line 156] Consider putting the reference to the overview as new separate sentence.
This has been corrected.

(25) [Line 169] I found the reference to a conditional mean to be not very clear. On what is the mean conditioned here?
This is more clearly defined in Appendix A: equation A1. To avoid any confusion with non-expert readers, we have decided to refer to the mean provided by the RF model.

(26) [Line 184] What is meant here by 'tolerance'?
This term is confusing and has been removed here. The "tolerance" is now clarified in the description of Monte-Carlo procedure in Sect. 2.4.2 – step (2).

(27) [Line 184] 'in Sect. 2.4'→ 'in the next section'.
This has been corrected.

(28) [Line 204] You mention AR6 here, but this has not been introduced/defined before.
We have more clearly defined it.

(29) [Line 209] Consider adding an adverb at the beginning of the sentence here (e.g., 'consequently'), so that it is clear that the 200 number is directly linked to the 50 number before, and not a new parameter for the validation exercise.
This has been added.

(30) [Lines 212–218] Given that you introduce three performance criteria, you should use three distinct items, not two.
This has been corrected.

(31) [Line 219] Consider removing 'for a GSAT scenario' from the title of the subsection. Both subsections 2.4.1 and 2.4.2 depend on the GSAT scenario.
This has been corrected.

(32) [Line 266] [4.6; 7.4cm]→ [4.6 cm; 7.4 cm] or [4.6; 7.4] cm.
This has been corrected.

(33) [Line 266] [10.4; 17.0cm]→ [10.4 cm; 17.0 cm] or [10.4; 17.0] cm.
This has been corrected.

(34) [Figure 6] Vertical label: 'density'→ 'PDF'.
This has been corrected.

(35) [Line 269] 'constructed the Monte-Carlo-based procedure': there seems to be a missing word here.
This has been corrected.

(36) [Line 281] 'whatever the performance criteria'→'for every performance criterion'.
This has been corrected.

(37) [Line 286] 'Table 1'→Table 2'.
This has been corrected.

(38) [Line 367] '(based on Goelzer et al. (2020))'→'(based on Goelzer et al., 2020)'.
This has been corrected.

(39) [Lines 410–411] '(Merrifield et al., 2023; Evin et al., 2019)' → '(Evin et al., 2019; Merrifield et al., 2023)'.
This has been corrected.

(40) [Line 411] 'take'→ 'took'.
This has been corrected.

(41) [Line 609] You use slc with a superscript for the index, and then with a subscript later in the text. Please be consistent in your system of notations.
This has been corrected.

(42) [Line 612] 'By nature'→'By construction'.
This has been corrected.

(43) [Line 615] 'squared errors': errors of what?
This has been replaced by the more specific term used for training RF emulators, i.e., the variance.

Orleans,
October 1st, 2025
J. Rohmer[1] on behalf of the co-authors

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