Replies to Referee #2's comments on "Drawing lessons for multi-model ensemble design from emulator experiments: application to future sea level contribution of the Greenland ice sheet" (egusphere-2025-52)

We would like to thank Referee #2 for the constructive comments. We agree with most of the suggestions and, therefore, we will modify the manuscript to take on board the comments and suggestions. We recall the reviews and we reply to each of the comments in turn (outlined in blue). Following the journal's reviewing procedure, the revised manuscript will be provided after the interactive review process, in a second phase.

Additional changes

Since the submission of this manuscript, we had the opportunity to include HIRHAM in the MME as well. In addition to the modifications suggested, we will thus include this third RCM model in the analysis. A major advantage will be to define an experiment where only the members of HIRHAM and RACMO are used ('woMAR' experiment) in addition to the 'MAR only' experiment. We expect that this modification will bring new insights and strengthen our conclusions which are, in the original version of the manuscript, based on two RCMs only. The results discussed in Sect. 4 will thus be modified accordingly.

Referee #2:

This study develops a Random Forest (RF) emulator to emulate Greenland 2100 sea level contribution (slc) output from a Multi-model ensemble (MME). In particular, the RF is trained using a set of 7 inputs, associated with the climate scenario, the ice sheet model (ISM) used, the regional climate model (RCM) used, and different settings of the ISM run. The authors investigate how changing the MME design leads to changes in the emulator performance and in its range of emulated slc. Based on these metrics, they provide guidelines for future MME designs that aim at estimating future ranges of slc from the Greenland ice sheet.

This study addresses an important and difficult question: how can we improve the design of MMEs to provide the best information about the probability density function (PDF) of future slc? The concept underlying this study is that the MME itself does not need to characterize this PDF, but that it should be designed optimally such that an emulator can do this characterization a posteriori. This is a valid and efficient approach to uncertainty quantification. It is also a challenging topic, and work on this topic is important. However, at this stage, I believe that several points need to be improved to make this study a valuable contribution in addressing this question. The authors make recommendations and they "expect these recommendations to be informative for the design of next generations of MME" (L22). But I believe that their recommendations are dependent on many assumptions or choices that they made, without always justifying them and making them clear to the reader. Furthermore, more methodological details about the RF emulator are needed because all the results presented depend on the emulation and, therefore, the RF design influences strongly any interpretation, and thorough RF evaluation is critical as well. I detail my concerns in this review, which consists of Major and Minor comments. I do not provide technical comments at

this stage of the reviewing process because I believe that the more substantial aspects should be addressed first. Line numbers in this review correspond to the preprint manuscript.

We thank Referee #2 for the in-depth analysis of the manuscript. In the following, we provide details on how we will account for them in the revised manuscript.

Major comment 1: Inherent assumptions associated with the MME

Many of the conclusions are strongly dependent on the particular MME used in this study. I have several reservations about this.

First, it is unclear to me how the MME used in this study was acquired and designed. The only details provided about the MME are (L74): "We focus on the sea level contribution from the Greenland ice sheet (GrIS) in 2100 based on a new MME study performed for the European Union's Horizon 2020 project PROTECT (http://protectslr.eu). Some modelling choices are taken from the protocols of the ISMIP6 initiative (Goelzer et al. (2020): in particular, the two main emissions scenarios, and the main model parameter explored."

Has this MME been peer-reviewed? Why are the authors not using the well-established ISMIP6 MME? The latter MME also has the advantage of providing a larger set of experiments, notably including many more ISMs than the MME of this study. At least, why has the MME not been combined with the ISMIP6 MME? Also, given that no publication describing the MME is referenced, I believe that it is important to give many more details about the MME configuration: Did all ISMs run under high- and low-warming forcing? Are the 15 global climate models used in the MME well-balanced across the runs? Etc.

The MME has been peer reviewed within the H2020 Protect project and is currently in preparation for submission to the Cryosphere Special Issue journal. In the revised version of the manuscript, we will refer to the EGUSPHERE preprint for further details. In addition, we will also provide a more detailed description of the Protect MME by paying particular attention to highlight the key differences with ISMIP6 MME.

To specifically reply to Referee #2 comment, we confirm that the experimental design builds on the ISMIP6 protocol with four different ice sheet models and extends it to more fully account for uncertainties in sea-level projections for the GrIS. This is underlined in Sect. 2. The Protect MME is thought as an extension of ISMIP6 MME regarding different aspects:

- we have included a wider range of CMIP6 climate model output, more climate change scenarios (SSP126, SSP245, SSP585);
- we have provided retreat forcing before 2015 that is calculated from reconstructions of past runoff and ocean thermal forcing. This allows for a consistent forcing of the models in past and future and to consider historical retreat of the outlet glaciers, which was an important source of mass loss after 1990;
- we have provided surface mass balance forcing from several RCMs, i.e. MAR and RACMO for the MME used for this study.

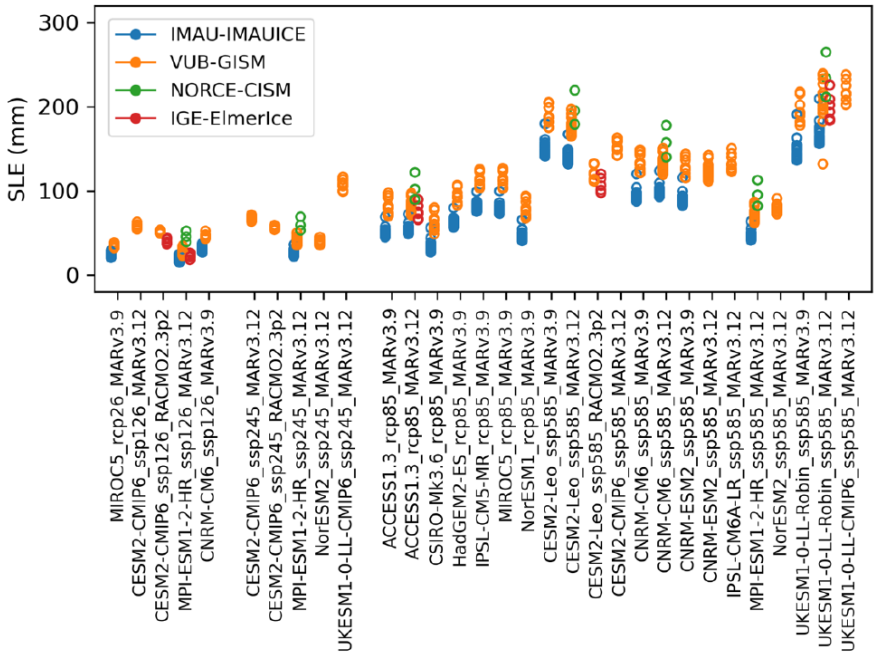
Since the submission of this manuscript, we had the opportunity to include HIRHAM as well. We believe that adding this third RCM model will add new insights and will make our conclusions more robust (based on two RCMs in the original version of the study). The results discussed in Sect. 4 will thus be modified.

To appreciate this extension, the following table gives the repartition of the members with respect to the RCP/SSP scenarios for ISMIP6 MME and for Protect MME (before inclusion of

HIRHAM). In addition, it should be noted that the total number of members has increased by a factor of 4 compared to ISMIP6 MME.

	ISMIP6 MME	Protect MME
RCP26	23	40
RCP85	156	319
SSP126	18	189
SSP245	0	189
SSP585	59	566
Total number	256	1303

As a second illustration, the following figure gives an overview of the results of the ensemble of projections at the year 2100 for all available Earth System models (ESMs), RCMs and ISMs under high, med and low retreat sensitivity. This allows to appreciate, graphically, how well the range of sea level is covered. It should however be underlined that, although the collection of forcing data covers a wide range of variations across different ESMs and scenarios, it ultimately still represents an 'ensemble of opportunity' similarly as for ISMIP6 MME. Our study aims to address the potential implications of this characteristic.

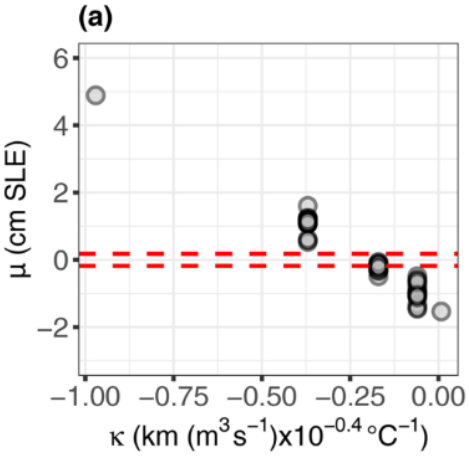


Second, the conclusions of this study are strongly dependent on the initial MME used in the emulation process. For example, the authors argue that there is "a quasi-linear relationship between κ and slc" (L306). But this conclusion is based only on the set of 4 ISMs used in this study: CISM, IMAUICE, GISM, and ElmerIce. Furthermore, given that very little experiments were performed by ElmerIce and GISM, I assume (I need to assume here because no information is provided on the design of the MME) that these two models may well have only been run with a single κ value. In this case, the "quasi-linear relationship" would be derived only from two ISMs. Given that different ISMs can show very different sensitivities to movement of the tidewater glacier front and grounding line positions, this conclusion could well be very different if other ISMs are included. So, if one was to perform a similar study with the ISMIP6 MME, would the "recommendations" for future MME design be different? As another example, I mention above that only 4 ISMs are included in the MME, two of which account for > 90% of

the simulations. By excluding CISM from the training experiments, the authors then make "recommendations" about the ability of the emulator to estimate the slc simulated by ISMs not included in the design. Here also, this evaluation depends critically on how similar simulated slc from CISM is to the simulated slc from IMAUICE. This similarity depends on numerous aspects that are specific to these two particular models. I would expect that the "recommendations" would be very different if other ISMs (ISSM, PISM ...) show more or less similarity with CISM.

We thank Referee #2 for the insightful analysis. Here we feel that some clarifications about κ should be given. As described in Sect. 2, we rely on a standard approach for integrating ocean forcing, i.e. based on an empirically derived retreat parameterization for tidewater glaciers (Slater et al., 2019, 2020). In this approach, κ is not a parameter in the ice-flow model; it rather represents the sensitivity of the ocean forcing as a whole. It may be thought of as defining the sensitivity of the downscaling from global model to local ice sheet scale, similar to the combined parameter choices in RCMs for downscaling climate conditions. In the studied MME, we have different RCMs, which have different sensitivities and produce different melt for the same global forcing.

Regarding the specific comment on the quasi linear behaviour, we expect this relationship not to change much by adding more ISMs because κ is "external" to the ISM as afore-described. To support this result, Referee #2 should refer to the following figure adapted from Fig. 7 of Rohmer et al. (2022) based on ISMIP6 MME: it shows the sensitivity index (denoted μ) that measures the contribution, in terms of sea level equivalent SLE, depending on the value of κ. A quasi-linear trend has here been identified.



We however agree with Referee #2 that a more careful attention should be paid not to 'over-interpret' our results by making the recommendations too general. The conclusions will be nuanced and reformulated in this sense. In addition, we also propose to modify the title to highlight that our results are linked to the specificities of our ensemble as follows: "Lessons for multi-model ensemble design from emulator experiments: application to a large ensemble for future sea level contributions of the Greenland ice sheet".

Reference

Rohmer, J., Thieblemont, R., Le Cozannet, G., Goelzer, H., and Durand, G.: Improving interpretation of sea-level projections through a machine-learning-based local explanation approach, Cryosphere, 16, 4637–4657, https://doi.org/10.5194/tc-16-4637-2022, 2022.

Such assumptions are not made explicit by the authors. This could very well be misleading to the readership targeted by the authors, especially those less familiar with ice sheet modeling (e.g., "stakeholders" (L328) and "coastal adaptation practitioners" (L332)).

We totally agree with Referee #2. The clarification on κ will be added to Sect. 2 to improve our message to stakeholders and coastal adaptation practitioners.

Major comment 2: Characterization of uncertainty

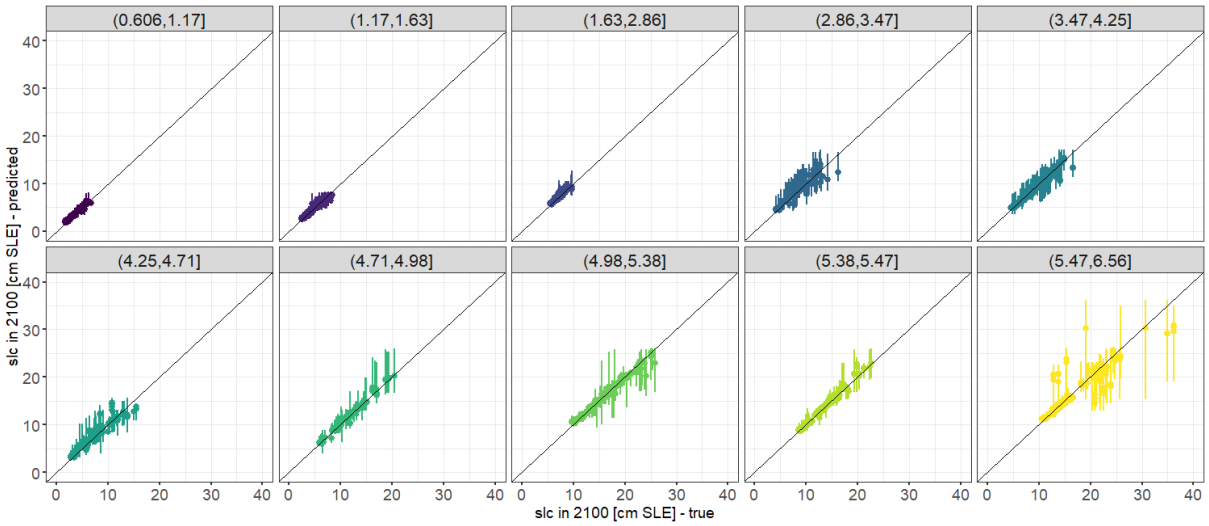
The authors use their random forest (RF) emulator such that "changes in the emulator's predictive performance and the emulator-based probabilistic projections provided information on several aspects" (L18). After reviewing the manuscript, I identify remaining limitations about the RF emulator regarding uncertainty characterization.

The authors use changes in the predictive performance of the RF as a proxy for uncertainty remaining about a hypothetical MME (here, a MME excluding some of the experiments). But this metric is sensitive to the particular machine learning model used for the emulation. Here, the emulation output is thus conditioned on the RF architecture, with a single fixed combination of hyperparameters. Is any decrease in predictive performance of this specific RF therefore a meaningful assessment of uncertainty imputable to the MME design? This question is critical, because the conclusions of this study use this as a fundamental assumption.

This issue is further exacerbated by the fact that the RF does not provide probabilistic output. By this, I mean that the RF only provides a point estimate. There is no uncertainty quantification. Ideally, the design of a MME should target the strongest reduction in posterior covariance (i.e., the uncertainty remaining given the current MME). But this particular RF emulator does not provide such metric. This could be addressed by choosing another architecture (e.g., Gaussian processes, Williams and Rasmussen (2006)), by subsampling techniques for RF models (Mentch and Hooker, 2016), or by adapting the RF to output conditional quantiles (Meinshausen and Ridgeway, 2006).

We thank Referee #2 for this suggestion. We agree that complementing the study with uncertainty quantification of the emulators itself would bring new insights and will allow us to better discuss the results. As suggested, we propose to implement the quantile random forest emulator, denoted qRF, for both the experiments on the emulator's performance (Sect. 3.3) and for the probabilistic projections (Sect. 3.4).

For the former application, the quantile random forest provides estimates of quantiles at any order τ, denoted $q^{\tau}(slc|\mathbf{x}^*)$ for a given instance of the input variables $\mathbf{x}^*$. The quantiles can directly be used to define the prediction intervals at any level α: $[q^{(1-\alpha)/2}(slc|\mathbf{x}^*); q^{(\alpha+1)/2}(slc|\mathbf{x}^*)]$. The following figure is a new version of Fig. 4(a), which allows to verify the satisfactory level of predictability for a large range of GSAT values with Q^2 ranging from 82%, for the largest GSAT values, to >98%.



New Figure 4. Comparison between the true numerically computed *slc* and the emulator's predicted values for the 25 validation tests (described in Sect. 2.4). Each panel corresponds to test samples for a given range of GSAT values (indicated at the top of the panel).

For the latter case, the emulator uncertainty is propagated in addition to the uncertainties to the different input variables based on the following procedure:
(1) Draw a random realization of the input variables $\tilde{\mathbf{x}}$;
(2) Draw a number $\tilde{u}$ between 0 and 1 by assuming a uniform random distribution;
(3) Compute $\widetilde{slc} = q^{\tilde{u}}(slc|\tilde{\mathbf{x}})$ given $\tilde{u}$ using the qRF model.

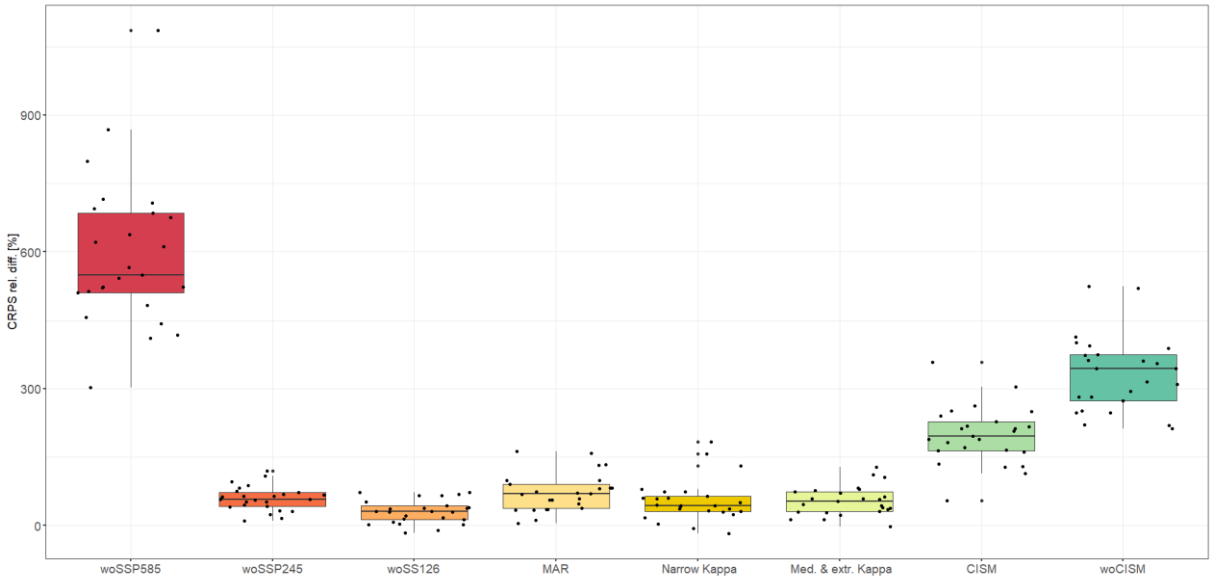
In addition, we propose to add a new performance indicator to analyse the changes in the emulator's performance in terms of reliability of the predictive probabilistic distribution. This is done using the continuous ranked probability score, denoted *crps*, as used for validating probabilistic weather forecast (Gneiting et al., 2005). To evaluate the *crps* score, the formulation based on quantiles (Berrisch and Ziel (2024): Eq. 2) is used:

$$crps = 2\int_0^1 B(q^{\tau}(slc|\mathbf{x}^*), slc^{true})\,\mathrm{d}\,\tau \approx \frac{2}{P}\sum_{\tau\in\Gamma} B(q^{\tau}(slc|\mathbf{x}^*), slc^{true})$$

where the term $B(q^{\tau}(slc|\mathbf{x}^*), slc^{true})$ is defined as $\begin{cases}(1-\tau)(q^{\tau}(slc|\mathbf{x}^*) - slc^{true}) \text{ if } slc^{true} < q^{\tau}(slc|\mathbf{x}^*) \\ \tau(slc^{true} - q^{\tau}(slc|\mathbf{x}^*)) \text{ if } slc^{true} \geq q^{\tau}(slc|\mathbf{x}^*)\end{cases}$, where slc^{true} is the true value of the sea level contribution, and where the quantiles $q^{\tau}(slc|\mathbf{x}^*)$ are evaluated using the trained qRF model at given instance of the input variables $\mathbf{x}^*$ for an equidistant dense grid of quantile levels $(\tau_1, \ldots, \tau_P)$ with $\tau_i < \tau_{i+1}$ and $\tau_{i+1} - \tau_i = 1/P$. In this study, we consider level τ_1=5% and τ_P =95% with 1/P=5%.

This score jointly quantifies the calibration of qRF probability distribution, i.e. the reliability of the estimation, and its sharpness (i.e. the concentration/dispersion of the probability distribution). The lower *crps*, the higher the quality of the qRF probabilistic predictions, with a lower limit of zero.

Our first tests (Figure below) show some similarities with the results for *RAE* and Q^2 performance indicators, but with a clearer effect of ISM experiments (CISM and noCISM). The results will be finalised in the revised version of the manuscript provided in the second phase of the reviewing process. A more thorough discussion will be provided in Sect. 4.



New Figure. Relative difference (in %) for the estimates of RF predictive capability measured by *CROS*, between the RF reference solution and the RF emulators trained by applying the experiment described in Table 2. The dots indicate the results of the 25 repetitions of random validation tests (described in Sect. 2.4).

Finally, we will also elaborate more in the Discussion section on the problem of model uncertainty related to the construction of the emulator; in particular the problem of hyperparameters' tuning and its relatively lesser impact for random forest models (Probst et al., 2019; Bischl et al., 2023).

Added references

Berrisch, J., Ziel, F., 2024. Multivariate probabilistic crps learning with an application to day-ahead electricity prices. International Journal of Forecasting, 40(4), 1568-1586, doi:10.1016/j.ijforecast.2024.01.005

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., et al., 2023. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 13(2), e1484, doi: 10.1002/widm.1484

Gneiting T, Raftery AE, Westveld III AH, Goldman T. 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. Monthly Weather Review 133(5): 1098–1118

Probst, P., Wright, M. N., Boulesteix, A. L., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, 9(3), e1301, doi:10.1002/widm.1301

Minor comment 1: Lack of technical information

All the results and conclusions from the study are dependent on the RF emulator. As such, I find that more information on the RF development and evaluation are needed. I highlight some aspects to prioritize here below.

(a) The evaluation of the RF (L174-183) is assessed through a random sampling evaluation, but I find the details about the evaluation method somewhat unclear. First, the authors mention the "iteration of the procedure" (L180). However, it is not explained what is iterative in this procedure. Later in the manuscript, the authors often refer to "25 validation tests" (e.g., caption of Figure 4). But this number of 25 is not explained in the description of the evaluation

method. Thus, I can only assume that the random validation is iterated 25 times. Second, it is unclear what the validation performance measure shown in Figure 4 represents. In Figure 4a, there are clearly much more than 25 points, but clearly less than 25 × 55 = 1375 points (where 55 is the number of test samples mentioned on L180). Thus, what does each point represent? In addition, why are there much less points shown in Figure 4b than Figure 4a? Finally, the authors explain that there are 55 test samples, but they draw 5 samples for 10% ranges between 0 and 100% (L179-180). As such, there should be 5 × 10 = 50 test samples I believe, not 55.

(b) I wonder why it was decided to use this random evaluation procedure. In particular, the commonly-used 10-fold cross validation procedure would have been a more natural choice. This would also avoid the influence of sampling biases related with the random sampling of relatively few experiments (55 from the 1303 experiments per iteration). Since 10-fold cross validation was used for parameter fitting (L197), I suppose that there is no computational issue for this. Also, it would be straightforward to exclude the members from the 9 training folds as required by the specific experiments (e.g., exclude all SSP5-8.5 when training for woSSP585). Thus, is there any reason to prefer the random evaluation over the 10-fold cross validation?

We reply here to comments (a) and (b). Referee #2 is right to say that 10-fold cross validation would be a more natural choice. 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. This ability is important in our case, because we discuss the performance with respect to the probabilistic projections given fixed GSAT values. The new Figure 4 (Major comment 2) illustrates this type of analysis.

Though our GSAT definition does not strictly correspond to the global warming level defined in AR6, they can help end users to interpret the projections associated to temperature constraints as illustrated by recent projections for France by Le Cozannet et al. (2025).

We also thank Referee #2 for noticing the problem with the number of test cases. The presentation in Sect. 2.4 has been clarified. We also propose to increase the number of test cases, now 10 per bin of GSAT values resulting in 100 test samples covering a larger range of GSAT values (new Figure 4, above).

Added reference

Goneri Le Cozannet, Remi Thieblemont, Jeremy Rohmer and Cecile Capderrey (2025). Sea-level scenarios aligned with the 3rd adaptation plan in France. (in press) https://doi.org/10.5802/crgeos.290

(c) More technical details about the RF emulator construction would be beneficial. In particular, mixing categorical and continuous inputs is not straightforward, and may incur performance sensitivity to the RF design. For example, what is the splitting criterion used: mean absolute error, mean squared error, other? And how did the authors alleviate the potential issue of selection bias towards the inputs that have more possible splits? This could partly influence the different sensitivities to, for example, SSP5-8.5 scenario (global annual mean surface air temperature change, GSAT, is a continuous input with many different values), ISM (categorical input), κ (continuous input with few different values). As such, some information on these technical aspects would help the reader understand how modeling challenges may affect the results or not.

We agree that more technical details should be added.

We use the mean squared error in the loss function of the random forest model. The treatment of categorical variables is based on the recommendation by Hastie et al. (2009): chapter 9.2.4.

We follow the implementation proposed by Wright et al. (2019), who showed that ordering the factor levels a priori, here by their mean response, is at least as good as the standard approach of considering all 2-partitions in all datasets considered, while being computationally faster; It has been shown to be more efficient than dummy coding and simply ignoring the nominal nature of the predictors as well.

However, as Referee #1 highlighted in her/his first comment, "the most readers of this journal are likely geoscientists, who will primarily be interested in the study's results". We propose to move these details in Appendix A on the Random Forest implementation.

Added references

Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning. Springer, New York. 2nd edition.
Wright, M. N., & König, I. R. (2019). Splitting on categorical predictors in random forests. PeerJ, 7, e6339.

Minor comment 2: Use of global mean temperature change

The authors aggregate all the combinations of emission scenario (SSP) and global climate model (GCM) as a value of GSAT. I wonder if this does not risk misrepresenting the climate forcing affecting the Greenland ice sheet (GrIS). In particular, a given GSAT could very well lead to different magnitudes of:
(1) GrIS surface air temperature change
(2) GrIS precipitation
(3) GrIS ocean forcing
I expect that there may well be some substantial differences in these 3 components between different GCMs. It would be interesting to explore whether separating the single GSAT variable into these 3 separate components refines the emulator predictions.

This choice is based on the approach followed by previous emulation studies (Edwards et al. (2021). An alternative would use regional climate variables. Although this would improve the signal-to-noise ratio for the emulator, but would restrict us to using computationally expensive general circulation models from CMIP5/6, for which there only a few tens of models. With the GSAT option, as performed by Edwards et al. (2021), the simple climate model like FaIR can be also used to explore uncertainties in each scenario thoroughly, using the latest assessments of equilibrium climate sensitivity.

Referee #2 is also invited to refer to our reply to major comment 1 on the clarification of κ which is closely related to the sensitivity of the ocean forcing as a whole.

Finally, in order to be more consistent with IPCC practices, we will slightly change the definition of GSAT by computing the difference between the temperature at the considered year and the mean temperature over the period 1995-2014. To ensure consistency, this is also done for *slc*, which is computed relative to the mean *slc* over the same period.

Minor comment 3: Interpretation of some results

I find that the interpretation of results are not always well supported quantitatively. I note that, in some cases, this may simply be due to a lack of clarity in the interpretation. I provide here a few examples.

2.1 The Dh, DS definition

In Figure 6, the authors show the different combinations of decrease in MME size (DS) and deviations from original histograms (Dh) resulting from their model experiments. Firstly, I think that the manuscript would benefit from a clearer definition of Dh. It is defined as "the average difference in the count numbers between the two histograms (normalised by the total number of members)" (L172-173). I believe that the normalization is by the histogram counts, not the total number of members, because otherwise Dh would be proportional to DS. For example, assume that for a given variable, we have a hypothetical 3-category histogram with counts 5, 10, 85 (i.e., n=100). In hypothetical experiment 1, the counts are 0, 10, 85 (i.e., n=95). In this case, DS = 100−5/100 = 0.95 and, following the definition, Dh = 5+0+0 /3 × 1/100 = 1/60 . In hypothetical experiment 2, the counts are 5, 10, 80 (i.e., n=95). In this case, DS = 100−5/100 = 0.95 and Dh = 0+0+5/3 × 1/100 = 1/60 . This shows that taking "the average difference in the count numbers between the two histograms (normalised by the total number of members)" results in an identical pair (DS,Dh) for these two hypothetical experiments. I am probably misunderstanding here, but I think that a more precise definition would help.

We thank Referee #2 for this insightful comment which made us rethink our procedure. As rightly shown by Referee #2, the proposed indicator D_h might fail to reflect the changes in the histograms.

As a remedy, we propose to rely on a well-established criterion for comparing different probability distributions, namely the Kolomogorov-Smirnov (*KS*) statistic, defined as:

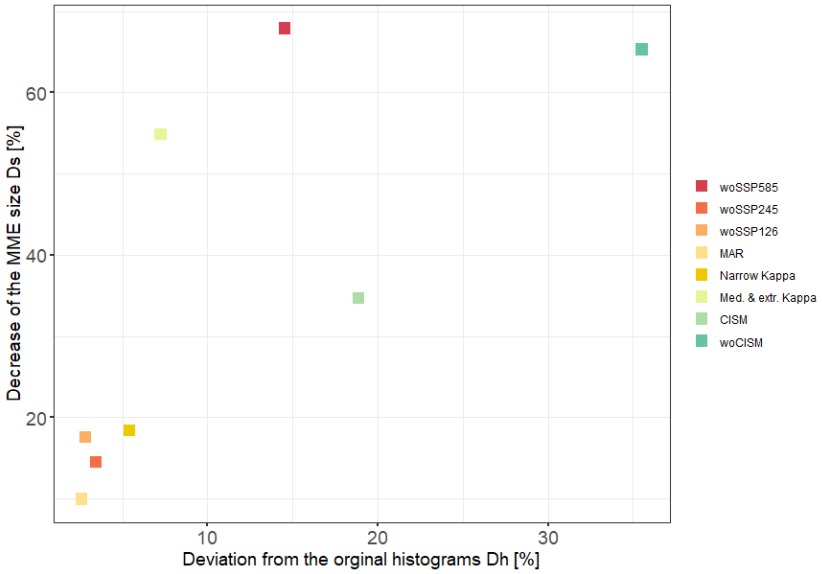
$$KS_X = \max |F_{\text{ref}}(X) - F_{\text{ref}|E_i}(X|E_i)|$$

where F_{ref} is the cumulative probability distribution CDF for the input X considering the reference solution, without perturbation; $F_{\text{ref}|E_i}(X|E_i)$ is the cumulative probability distribution CDF for X when applying the emulator experiment, i.e. when removing members. For categorical inputs, the CDF is defined by assigning each level an integer index.

The new D_h criterion is then defined as the average value of KS_X over all Xs. By doing so, we make sure that D_h reflects the average value of the deviations in the members' distributions.

The new Figure shows how each experiment affects the MME. The use of the new D_h criterion is more 'discriminative' by highlighting different groups of points:

- The group of points at the bottom, left corner corresponding to MAR, Narrow Kappa, woSSP245 and woSSP126. They all have minor-to-moderate impact on both criteria, D_S and D_h;
- The group of points with decrease in size D_S on the same order of magnitude, i.e. between 55 and 65% with different behaviours in terms of histograms' perturbation as shown by the wide range of D_h values. This group corresponds, in increasing order of D_h to 'Med. and Extr. Kappa', woSSP585 and woCISM;
- The point at an intermediate position with moderate D_S and D_h changes which corresponds to CISM.



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

2.2 The Dh, DS results

I do not understand the interpretation of the impact from Dh, DS on the emulator performance (Sect. 3.3). First, the authors write "Excluding the extreme SSP scenario SSP5-8.5 (experiment 'woSSP585') has the largest impact in terms of RAE relative difference with respect to the original RF performance (Sect. 3.1), where RAE is increased of ~ 10% compared to the original RAE value (Fig. 4)" (L245). However, Figure 7 shows a ~275% relative difference in RAE, so it is not clear to me where the value "~ 10%" comes from. Second, I do not follow the logic of the arguments. The authors write that (i) the high DS of woSSP585 causes large errors. But then, (ii) they argue that "this 'size effect' is not the only contributor to the performance impact, as shown by the 'woCISM' experiment, which removes an equivalent number of members to the 'woSSP585' experiment (Fig. 6), and the resulting RAE increase reaches half that of 'woSSP585' experiment" (L253). And (iii) that the woCISM experiment has the largest Dh value. However, when I interpret Figures 6 and 7, I find that (a) woSSP585 and woCISM have similar DS values (i.e., (ii)), (b) woCISM has higher Dh than woSSP585 (i.e., (iii)), but (c) that the errors from woSSP585 ar much higher than those of woCISM (Figure 7). So, it seems that the lower Dh of woSSP585 is accompanied by larger errors. This is the opposite message to that conveyed in the text: "This shows that the second important factor here is the diversity among the members within the MME after applying the experiment. The Dh indicator remains, however, a first-order approximation of this diversity (...)" (L256). The statement of greater diversity leading to lower errors, is not supported by the larger errors of woSSP585 compared to woCISM. To summarize: DS(woSSP585) ≈ DS(woCISM), Dh(woSSP585) < Dh(woCISM) where low Dh implies greater "diversity", but RAE(woSSP585) >> RAE(woCISM).

In the revised version of the manuscript, we will discuss in more details the links between the groups highlighted above and the implications with respect to the emulator's performance as well the influence on the probabilistic projections. Note also that the results will be updated taking also into account new members with HIRHAM RCMs.

2.3 Figure S3 (in Section S2)

The authors write "The analysis of an alternative indicator of emulator's predictive capability in Supplementary materials S2 confirms these results" (L261). However, in my view, Figure 6

(RAE results) and Figure S3 (Q2, coefficient of determination results) show contrasting conclusions. For example, RAE of woCISM, CISM, and MAR are comparable (Figure 6). However, Q2 is clearly lower for woCISM than for MAR and CISM (Figure S3). This indicates differences when evaluating relative errors versus explained variance. Thus, these differences are potentially interesting to analyze, instead of being discarded as is done in the main text. In particular, they could relate to the emulator performance sensitivity to high versus low slc (the latter being more influential on relative metrics), or its sensitivity in the ability to predict values away from the mean value, or other aspects that would require investigation. Note that this links back to my general comment about the importance of understanding the RF emulator, because the interpretation of the results depends strongly on this understanding.

We thank Referee #2 for this valuable suggestion. We totally agree with Referee #2 and will include in the analysis not only *RAE* but also Q^2. In addition we also propose to analyse the performance indicator *crps* that measures the quality of the emulator's predictive probability as well. Referee #2 is invited also to refer to our detailed reply to Major comment 2.

2.4 Figure 8

There are many aspects that I find puzzling or questionable in Figure 8. Firstly, the results do not correspond to what is shown in Figure S4, where the Q5% and Q95% are shown with the black error bars. For example, in the column ΔGSAT=+3◦, Q95% of woCISM, woSSP245, and woSSP585 are clearly strongly different from the Q95% labeled "original" (Figure S4). But Figure 8 shows that these differences are ≤ 1%. I believe that there is an inconsistency here, or something that I misunderstand about Figure 8.

Secondly, I do no understand how it is possible that the changes in median and quantiles at ΔGSAT=+4◦ are so small for woSSP585. In this design experiment, the RF model has presumably not even seen such levels of warming during training because the SSP 5-8.5 scenario has been excluded. But, by definition, tree models (including RF) predict slc based on decision rules seen during training. Thus, it is not clear how the RF can predict relatively similar slc values under ΔGSAT=+4◦ when excluding SSP 5-8.5 as when it is not excluded. I am probably misunderstanding something here, but I believe that the authors should explain this counter-intuitive aspect of their results.

We thank Referee #2 for his careful analysis. We confirm that, for some experiments, there are some discrepancies that have revealed a bug in our scripts for the plotting. We will of course update the results in the new manuscript.

Minor comment 4: Some conclusions need to be put into perspective

For different aspects, I find that better communication and/or more context about the conclusions is needed. I highlight some key examples here.

(a) Concerning κ, the authors argue for "the lesser importance of the choice in the range of the Greenland tidewater glacier retreat parameter" (L21). However, they compare it with the influence of the SSP scenario and of the ISM choice. It is expected that a single parameter should have much less influence than a global warming scenario and than a full ice sheet model.

We thank Referee #2 for this comment. We feel that some clarification about κ should here be given. Here, κ is not thought as a parameter of the ice-flow model, it rather represents the sensitivity of the ocean forcing as a whole. It may be thought of as defining the sensitivity of the downscaling from global model to local ice sheet scale, similar to the combined parameter choices in RCMs for downscaling climate conditions. In the studied MME, we have different RCMs, which have different sensitivities and produce different melt for the same global forcing.

Since we have only one approach to 'downscale' the ocean forcing, κ is sampling that uncertainty in a similar way.

In addition, our previous study (Rohmer et al., 2022) highlighted the high importance of κ compared to other uncertainties. Referee #2 is also invited to refer to our detailed reply to major comment 1 as well to Referee #1's comment 3.

Reference

Rohmer, J., Thieblemont, R., Le Cozannet, G., Goelzer, H., and Durand, G.: Improving interpretation of sea-level projections through a machine-learning-based local explanation approach, Cryosphere, 16, 4637–4657, https://doi.org/10.5194/tc-16-4637-2022, 2022.

(b) It should be better emphasized that the probabilistic ranges shown by the authors are not probabilistic projections of Greenland slc. Instead, they show a range of emulator predictions (thus conditioned on the emulator architecture) assuming a uniform distribution over the different inputs (L186). Thus, it does not represent calibrated uncertainty accounting for model-observation misfits (e.g., Aschwanden and Brinkerhoff, 2022). And neither does it represent the slc PDF from the MME, because the uniform distribution over the input space is not representative of the MME itself (e.g., the minimum spatial resolution is clearly not uniform between 1 and 40 km, see Fig. 3). As such, I believe that the true meaning of the PDFs shown in Figure 5 should be explained explicitly in order to avoid any reader misinterpreting those PDFs.

We agree with this comment. To do so, we propose to clarify the caption of Fig. 5 as well as the description in Sect. 2.4.

(c) The authors make a conclusion on "the utmost importance of including the SSP5-8.5 scenario, due to the large number of simulations available and the range of global warming they cover" (L19- 20). However, I do not think that the authors have proven the co-existence of these two points. For example, could it be that including only a few training simulations with high global warming forcing would be sufficient to drastically decrease the errors of woSSP585 shown in Figure7? In other words, maybe the emulator needs only a few high-warming training examples to correctly interpolate in the existing range of warming scenarios. Or maybe, as the authors write (L19-20), it is also the high number of experiments that is important. However, as far as I understand, the results presented in this study do not allow to evaluate the relative importance of these two aspects.

We agree with Referee #2 and it also goes in the same line than the comment on the dependence of the results to the MME (Major comment 1). We will nuance our conclusions in this sense. We believe that the influence of the MME size is shown by our results, but disentangling this effect from the range of global warning remains too complicated at least with the procedure proposed here.

References

Andy Aschwanden and DJ Brinkerhoff. Calibrated mass loss predictions for the greenland ice sheet. Geophysical Research Letters, 49(19):e2022GL099058, 2022.

Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. Journal of machine learning research, 7(6), 2006.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. Journal of Machine Learning Research, 17(26):1–41, 2016.

Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for machine learning, volume 2. MIT press Cambridge, MA, 2006.

We thank Referee #2 for the suggested references.

Orleans,
May 5th, 2025
J. Rohmer[1] on behalf of the co-authors

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