Review of "Drawing lessons for multi-model ensemble design from emulator experiments: application to future sea level contribution of the Greenland ice sheet" by Rohmer et al.
Reviewer: Vincent Verjans

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.

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://protect-slr.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.

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 paticular models. I would expect that the “recommendations” would be very different if other ISMs (ISSM, PISM ...) show more or less similarity with CISM.

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)).

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).

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 $25 \times 55 = 1375$ 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 $5 \times 10 = 50$ 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?

(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.

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.

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 D_h, D_S definition
In Figure 6, the authors show the different combinations of decrease in MME size (D_S) and deviations from original histograms (D_h) resulting from their model experiments. Firstly, I think that the manuscript would benefit from a clearer definition of D_h. 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 D_h would be proportional to D_S. 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, $D_S = \frac{100-5}{100} = 0.95$ and, following the definition, $D_h = \frac{5+0+0}{3} \times \frac{1}{100} = \frac{1}{60}$. In hypothetical experiment 2, the counts are 5, 10, 80 (i.e., n=95). In this case, $D_S = \frac{100-5}{100} = 0.95$ and $D_h = \frac{0+0+5}{3} \times \frac{1}{100} = \frac{1}{60}$. 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 (D_S, D_h) for these two hypothetical experiments. I am probably misunderstanding here, but I think that a more precise definition would help.

2.2 The D_h, D_S results
I do not understand the interpretation of the impact from D_h, D_S 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 $\sim$ 10% compared to the original RAE value (Fig. 4)" (L245). However, Figure 7 shows a $\sim$ 275% relative difference in RAE, so it is not clear to me where the value "$\sim$ 10%" comes from. Second, I do not follow the logic of the arguments. The authors write that (i) the high D_S 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 D_h value. However, when I interpret Figures 6 and 7, I find that (a) woSSP585 and woCISM have similar D_S values (i.e., (ii)), (b) woCISM has higher D_h 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 D_h 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 D_h 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: D_S(woSSP585) $\approx$ D_S(woCISM), D_h(woSSP585) $<$ D_h(woCISM) where low D_h implies greater "diversity", but RAE(woSSP585) $>>$ RAE(woCISM).

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 (Q^2, coefficient of determination results) show contrasting conclusions. For example, RAE of woCISM, CISM, and MAR are comparable (Figure 6). However, Q^2 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.

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 $\leq 1\%$. 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.

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.

(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.

(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.

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.