Replies to Referee #2's comments on "Improving interpretation of sea-level projections through a machine learning-based local explanation approach". (egusphere-2022-435)

We would like to thank Referee #2 for the constructive comments. We agree with most of the suggestions and, therefore, we have modified the manuscript to take on board their comments. We recall the reviews (black, italic) and we reply to each of the comments in turn (blue).

Referee #2:

Following is a review of, "Improving interpretation of sea-level projections through a machine-learning-based local explanation approach" by J. Rohmer et al. In this manuscript, the authors describe a strategy for the interpretation of an ensemble of Greenland Ice Sheet model projections, in particular, those created for the ISMPI6 experiment. The goal is to use a novel machine-learning approach (SHAP) to analyze the existing ensemble, and bring insight to the ice sheet modeling community, in terms of what modeling choices affect model results and when. This manuscript focuses on the ISMIP6 high end (RCP8.5) projections forced with MIROC5 output, which are provided through the year 2100. The authors find that different modeling assumptions influence results during varying epochs of the projection. In particular, model results are sensitive to the retreat parameter, especially after the first 30 years of simulation, as well as the choice of ice flow equation. A significant dependence of results on minimum grid cell spatial resolution is also found. The authors conclude that the SHAP approach is a promising method for analysis of earth system multi-model ensembles (MME), especially in terms of extracting information about how modeling assumptions may drive simulation results. They note that, with caution, the analysis can offer valuable insight, and they offer suggestions on how to improve upon the approach for future studies. The manuscript is well written and organized, and the figures are of good quality. The authors especially take care in describing the methods, including a schematic to describe the procedure adopted for this study.

Overall, the manuscript is successful in illustrating that the SHAP approach can be used to help researchers interpret results of MME experiments, like ISMIP6. The methods are novel, especially in the adaption of relatively new machine learning techniques to ice sheet model projections of sea-level change. The introduction offers a thorough explanation of the background of the adopted approach and the data section adequately describes the ISMIP6 experiments and model assumptions chosen for this study. However, I find that some additional explanation could be added to the application section, to help lead the reader through the analysis results. I also find that the discussion could be expanded to add context to the analysis results, particularly with respect to how the results might be compelling for the ice sheet modeling community or how they might impact future ice sheet model intercomparison projects. Overall, I recommend publication of this manuscript with minor revisions.

We are grateful to Referee 2 for the positive analysis.

As recommended, to improve the readability of our analysis, we added both in the core text and in the captions of Figures 6-10 some guidance on how to read the results. To further demonstrate that the approach can be useful to the ice sheet modelling community, we have expanded the

practical recommendations in Sect. 5. We have clarified how the different analyses can be of interest for different users, namely the ice-sheet modelling community (individual modellers, or modelling group in charge of the design of experiments), but also adaptation practitioners. Below we provide details on how we took into account both comments.

Below I have some specific comments and suggestions for the authors:

Line 35: Could you please explicitly define what is meant by global vs. local for this context? This is not necessarily terminology that some readers would be familiar with.
We agree that this terminology deserves further explanations. We have completed the introduction as follows: "Commonly-used approaches to improve interpretability usually focus on measuring how important modelling assumptions are for prediction (e.g., Lundberg et al., 2020). Two main approaches exist, either global or local. In the global approach, the objective is to explore the sensitivity over the whole range of variation of the considered modelling assumption, i.e. to assess the variable importance across the whole MME dataset. This can be done by quantifying the MME spread and by identifying its origin (see among others, Murphy et al., 2004; Hawkins and Sutton, 2009; Northrop and Chandler, 2014). For this objective, popular statistical approaches generally rely on variance decomposition (ANOVA); see e.g., Yip et al., 2011 for an introduction. To complement these global methods, we adopt in this study the second approach, i.e. the local approach, which aims at measuring the importance of the variables at the level of individual observations. This means that the local approach focuses on how particular modelling assumptions (i.e. value of a given model parameter, a given ISM formulation, etc.) influences the considered prediction".

Section 2 title, figure 7 caption, and line 468 (and maybe others): Since these are produced from simulations, they are not really "Data", but model output.
We have changed the title of section 2 to better reflect the type of data we are using, namely "Multi-model ensemble case study".

Line 98: more accurately, this can be referred to as "global mean sea-level equivalent"
It is our understanding that "global mean sea-level" is attached to the idea of spatial average. In this case, we rather refer to the mean of an ensemble at a given time. Therefore, we chose to keep the original formulation.

Line 126: maybe "predicted" or "modeled" or "simulated" sea-level change?
This has been specified as "numerically simulated sea level change"
r
Line 357: "sea-level" contribution
This has been specified as follows: "We explore how the magnitude of the modelling assumption's contribution to *sl*".

Line 360: Could you add a statement about what this might mean from the modeling perspective (similar to what is done to conclude the next paragraph about Fig 7b)? For example, does having a negative sea-level contribution result from a low κ value suggest anything to modelers, or is it possibly too dependent on the specific warming scenario being tested (i.e. RCP8.5 from MIROC5)?
We have added a clarification as follows: "We also note that for setting κ above -0.17 km.$(m^3.s^{-1})^{-0.4}$ °C even impact negatively the *sl* prediction, which means that this modelling assumption pushes the prediction lower than the mean value for 2100".

Referee #1 also suggests the possible dependence of this result to the specific warming scenario being tested. Though we find this suggestion very relevant, it is worth underlying that our primary objective is to test the feasibility of our approach, and not to produce finalised results for GrIS, i.e. our work is methodological. In addition, given the number of MME results for RCP2.6, MIROC5 (of 23, i.e. almost 60% less MME results than for RCP8.5, MIRCO5), the validity of this hypothesis can hardly be explored. Therefore, we have clearly highlighted in the conclusion the need for multiplying the application cases (in particular by varying the ESM and the RCP choice).

Line 372: First, please take care to lead the reader through your logic, in particular, it would be helpful to explicitly remind the reader that conclusions are based on the red envelope derived from your analysis. In terms of the stated conclusions in this paragraph, it looks to me that 3cm is a value suggested by the fitting curve envelope. It also appears that 3cm is significantly above the plotted interval (maybe ~2.5 cm SLE is a more accurate number here?). Also, if we follow the logic of using the fitted curve to drive conclusions, then it looks like to me that there are values for >5 km resolution that should still be considered negligible. The curve actually suggests that perhaps >7.5km might be a more appropriate cutoff for the >5km statement? In addition, it seems that since the few results from the 3-4km range are driving the 2 km conclusion (as noted in the text). Because of this, I suggest softening the statement to say that results support a minimum grid size of 5km for sure, but they also reveal that a minimum resolution of as fine as 2km may be required, with more investigation needed. Overall, please consider revising this paragraph's wording, in general, with more accurate statements to reflect the plotted output. (This is important as the results are highly pertinent to ice sheet modelers and may be referenced to support modeling decisions in the future.)
We agree that the smoothing regression in Fig. 7b may add some confusion and we have removed it. We now analyse the influence of the minimum and maximum grid size by identifying a region where their influence becomes significant.
Sect. 4.3.2 has been re-written as follows: "Fig. 7c,d give insights into the influence of the spatial resolution by showing a zone of low-to-moderate influence defined for a minimum and a maximum grid size <5km and 16km respectively. In this zone, the average value of $|\mu|$ across the cases is 0.55cm and 0.27cm for the minimum and maximum value respectively (with a maximum value up to ≈1.1cm both grid sizes). The influence can even be considered non-significant for 40% of the cases falling within the +/- *MAE* range for the maximum grid size. From a modelling perspective, this analysis suggests that there is clear interest in running high resolution simulations. This means that if spatial grid resolution is too coarse, this choice may highly influence the results of sea-level projections; $|\mu|$ can be as high as 1.60 and 2.50cm for the minimum and maximum grid size respectively. A comparison with the contributions of the other modelling assumptions in Figure 8 further suggests that the influence of spatial resolution may dominate all other modelling choices, since their contributions do not exceed +1cm, i.e. they are smaller than those of the identified area of minimum and maximum grid sizes".

In addition, we have added some words of caution in the Discussion section (Sect. 5) regarding the values around 2km as follows: "Additional computer experiments are worth conducted in future modelling exercises to confirm the identified trends; in particular for minimum grid size ranging from 3 to 4km […]."

Line 376: This final statement is a bit awkward. Please consider revising. Maybe something like: "if spatial grid resolution is too coarse, this choice may highly influence the results of sea-level projections."
Thank you for the suggestion. This has been corrected in this sense.

Lines 383-384: This sentence is awkward, and it is not explicit what is meant by "mask" in this context. Please reword. For example, "Fig. 8 further suggests that the contribution of minimum grid size might dominate over (?) all the other modeling choices, since they do not exceed those contributions associated with minimum grid sizes of 8km (?) or greater." Or something similar.
Thank you for the suggestion. This has been rephrased as suggested.

With respect to this statement, don't the ice flow extreme responses technically exceed that of the extremes of the available minimum grid size? That is, aren't the results in Fig 7b within 10km and 15km artificially high, as an artifact of the fitted curve? This might just be an issue of reworking this paragraph to lead the reader through the stated conclusions in a clear way, but as written, the logic is not obvious.
We agree that the smoothing regression in Fig. 7b may add some confusion and we have removed it in the new Fig. 7.

Line 396: be noticeable in the -> "to impact the"?
This has been corrected.

Figure 9f title: For consistency, please reference κ in addition to (or instead of) Retreat parameter
This has been corrected.

Line 424: *At this point in the discussion, it would be helpful to add some sentences to put these results into context for ice sheet modelers. That is, what are the implications for some of these findings and suggestions? Fleshing out some of these ideas and expanding upon them during the discussion would broaden the audience who can benefit from this type of study.*
We agree that some additional clarifications are needed to better show what kind of results can be meaningful to ice sheet modellers. Our primary motivation was to develop an approach to better understand why a given instance of the modelling assumptions leads to a certain prediction (as underlined in the introduction). In this view, we believe that the diagnostic plots (Fig. 6) can be helpful. In addition, by aggregating all these diagnostic, we further provide information on the model structure at a given prediction time (Level 2 of the procedure), and globally over time (Level 3). Both analyses give insights on the sensitivity of *sl* to a considered modelling assumption across the MME results, which help understanding better the numerical model for deriving the MME and these lessons can serve as guidance for future computing exercises.

To reflect both aspects (and in alignment of Referee #1's recommendation), the following modifications were undertaken:

1. We have elaborated on the analysis of the diagnostics provided in Sect. 5 (and Figure 6) to better show how a critical analysis of the *sl* projections can be conducted;
2. The discussion section (Sect. 5) has further been expanded to better highlight how our results can be useful to different users.

"Improving the interpretability of sea level projections is a matter of high interest given their importance to support decision making for coastal risk management and adaptation. To this end, we adopt the local attribution approach developed in the machine learning community to provide results about the role of various modelling choices in generating inter-model differences in MME. These results are intended for different potential users.

First, the diagnostics illustrated in Fig. 6 (and all provided by Rohmer (2022) for MIROC5,RCP8.5-forced GrIS MME in 2100) help the individual modellers involved in the modelling exercise to understand and quantify the impact of their particular assumptions. Figure 6b-d illustrate situations where the SHAP approach allows such critical analysis including checking that the same modelling assumptions have a similar impact on nearby *sl* values.
Second, aggregating all diagnostic (Level 2 and 3 of the proposed approach) provides guidance to the modelling group involved in the definition of experimental protocols for MME (such as ISMIP6, Nowicki et al., 2020). Some key aspects are identified and deserve to be taken into account in future model developments and modelling exercises:

- our results confirm the need for simulations that are sufficiently spatially resolved: *sl* results are largely affected by too coarse grids (here with minimum and maximum grid size not larger than 5km and 16km respectively) regardless of the prediction time;
- the influence of the modelling assumptions depends on the considered prediction time: in the short/medium term (before 2050), initialisation and ice flow's type primarily contribute to *sl*, whereas in the long term , initial year and κ are the key contributors;
- some modelling choices have little impacts on the *sl* values (in average across the considered MME results), in particular choosing a finite element or finite difference numerical scheme or the dataset for bed topography;
- additional computer experiments are worth conducted to better explore given part of the parameter space in the view to confirm the identified trends; in particular for minimum grid size ranging from 3 to 4km and for κ ranging from -0.97 to -0.37 km.$(m^3.s^{-1})^{-0.4}$ °C.

Finally, framing the diagnostic results with narratives is expected to facilitate the communication between modellers and end–users. What is 'easily explained' through narratives is expected to increase the end-user's level of trust in the model, and eventually their engagement in the decision-making process (e.g. Jack et al., 2020). The narratives can follow the example of the GrIS study (Fig. 6(a)): "the largest *sl* predicted value is 19.1cm by 2100 and is mainly attributable (by a positive factor of almost 50% of the ensemble mean) to setting κ to its largest absolute value, i.e. a large contribution of outlet glacier retreat, while the other modelling assumptions have only moderate influence". More broadly, this provides a clear message for risk-adverse stakeholders interested in the upper tails of the distribution (named "high-end" sea level scenarios, Stammer et al., 2019), namely the importance of the dynamics of ice sheet processes on projected high *sl* values, especially in the second half of the century. This message then calls for intensified future research work to reduce uncertainty related to these processes".

Added references

Nowicki, S., Goelzer, H., Seroussi, H., Payne, A. J., Lipscomb, W. H., Abe-Ouchi, A., et al.: Experimental protocol for sea level projections from ISMIP6 stand-alone ice sheet models. The Cryosphere, 14(7), 2331-2368, 2020.
Stammer, D., Van de Wal, R. S. W., Nicholls, R. J., Church, J. A., Le Cozannet, G., Lowe, J. A., et al.: Framework for high-end estimates of sea level rise for stakeholder applications. Earth's Future, 7(8), 923-938, 2019.

Lines 425-435: It would also be beneficial to the manuscript to lead the reader and explicitly explain about how these points pertain to the particular study case here (as opposed to only referring back to earlier sections).
We agree that the reference to the introduction may add some confusion. We have removed it. We have re-written this part by focusing on the applicability of our approach as follows: "These results were obtained by overcoming two major difficulties. The first one is related to the incomplete and unbalanced design of the experiments (Sect. 4.1). Applying more common

statistical methods, namely the linear regression model or the ANOVA-based approach, would hardly be feasible. […] The second difficulty is related to the presence of statistical dependencies (as outlined in Sect. 4.1) […]".

For example, is the choice of method mostly appropriate because of the specific model assumptions that were chosen? Though these assumptions do have inter-dependencies, there are many other model choices which could be studied but are much more inter-dependent (for example different physically based parameters involved with various processes related to ice dynamics).
We agree with Referee #2 that different physically based parameters could have been selected in the analysis. This comment also aligns with Referee #1's comment. To address this problem, we have re-conducted our analysis by adding three new variables (maximum grid size, initial year and type of bed topography dataset). The impact of the selected input variables is also discussed in Sect. 5. See a more detailed analysis in the reply to Referee 1's comments.

Can you say anything about what type of ice sheet model parameters this method would and would not be appropriate for diagnosing, based on the experience gained in this study? While the language currently included is cautionary, I would like to see more discussion geared towards ice sheet models in particularly, like whether the results may be highly specific to the chosen ensemble (e.g., RCP8.5 MIROC5 climate forcing) or what type of "right" or "wrong" conclusions an ice sheet modeler designing a new intercomparison project might take from the method presented here.
Regarding the results relevant to the ice-sheet modelling community, we believe that the reply for "Line 424" (see above) should clarify the added value of our approach.

Regarding the specificity to the chosen ensemble, we agree with Referee #2 that additional tests should be carried out. We have underlined this aspect in the concluding remarks as follows: "This study should however be seen as a first assessment of the potential of the SHAP-based approach, and in order to bring the SHAP-based approach to a fully operational level, we recognise that several aspects deserve further improvements. First, a common pitfall of any new tool is its misuse and over-trust on the results (as highlighted by Kaur et al. (2020)). Future steps should thus concentrate on multiplying the application cases (in particular by varying the ESM and the RCP choice) with an increased cooperation between the different communities, namely ice sheet modellers, MLs, human-computer interaction researchers and socio-economic scientists".

Regarding the "right" or "wrong" conclusions, we have clarified in Sect. 5, the two key prerequisites for our approach to work well, namely the high predictive capability, and the choice of which input variables to include in the analysis.

Minor notes:

Line 23: GCM is more typically used to stand for General Circulation Model. If a more general acronym is desired here, I recommend using something like ESM (Earth System Model) for this context.
We agree with this comment. But for sake of homogeneity of the terminology with the ISMIP references (Goelzer et al., 2020; Nowicki et al., 2020), we preferably use the term "Atmosphere–Ocean General Circulation Model (AOGCM) output".

Line 91: relevance "of"

This has been corrected.

Line 145: units
This has been corrected.

Line 226: gets
This has been corrected.

Line 323: allows "us" (?)
This has been corrected.

Line 353: values
This has been corrected.

Line 359: values
This has been corrected.

Line 364: "indication of" where
This has been corrected.

Line 374: "a" few
This has been corrected.

Line 419: setting "of" the minimum grid size
This has been corrected.

Line 425: helps "alleviate"
This has been corrected (actually in Line 435 and not Line 425).

Orleans,
September 12th, 2022
J. Rohmer[1] on behalf of the co-authors
[1] BRGM, 3 av. C. Guillemin - 45060 Orléans Cedex 2 – France