Replies to Referee #1'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 #1 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 #1:

The manuscript "Improving interpretation of sea-level projections through a machine learning-based local explanation approach" provides a novel approach to analyzing multimodel sea level projections, using machine learning. The method put forward in this study has two parts: (1) using ML models as surrogates for the projection ensemble, (2) explaining the role of various modeling assumptions in generating the differences between sea level projections. The method is novel for our community and provides a potentially powerful was to analyze MIP output in a way that can guide potential future development (though this part needs more elucidation). I have some questions about particular choices in the methods and in general the writing would benefit from clarification. However, overall I find the contribution to be interesting and appropriate for The Cryosphere.
We thank Referee #1 for their positive analysis.

Major questions:
1. Within the context of an uncalibrated multi-model ensemble, the explanation approach adopted here explains why individual models deviate from the median projection. However, it does not say how an individual model deviates from the "truth" (since the models are not compared against reality or some benchmark). Thus, the authors need to be clear about this fact in their claim that this method can help in model development. How can it help it model development? If modelers now that factor X is causing their model to deviate from the ensemble median in some way, how will this facilitate development or adjustment of factor X? The authors needs to be more specific about this claim, which is essentially the motivation for using this method.
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, 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 the lessons drawn from our analysis as follows:

"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.

2. As explained in the manuscript, the SHAP method is always guaranteed to produce a set of quantified contribution (mu) that add up to exactly the total projection. This has some nice mathematical properties, but it also means that the quantified contributions are themselves dependent on the choice of which factors (x) to include in the analysis. For complex ice sheet models, this includes many, many more differences than the six included in this analysis. This leads me to three questions:

We agree with this analysis and are grateful for the suggestions. Below we describe how we accounted for this problem.

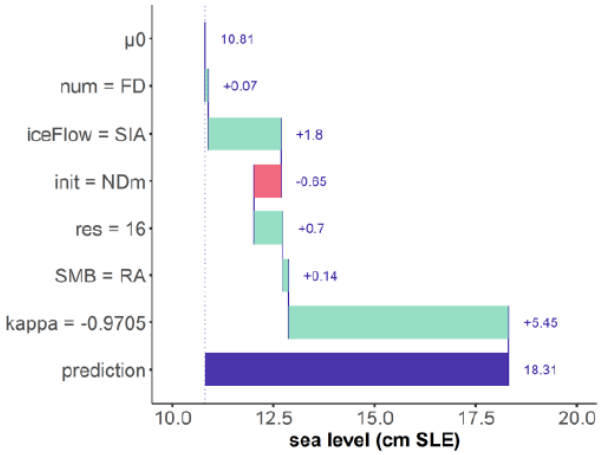
(a) The criteria you used to select which factors to include was those "without empty entry in Table A1 and with a sufficient number of variation across the models." The first makes sense - you need labels (though perhaps some discussion would be in order about whether it is possible to add labels by talking to the modelers response for these simulations). The second is unclear to me - it seems that you left out some potentially very important factors like initialization year and bed topography used with complete labels. Ultimately this selection, which would seem to be very important in the overall, seemed quite ad-hoc.

As rightly underlined by Referee #1, the SHAP approach aims at exactly decomposing the *sl* prediction. This means that for a given set of input variables, the SHAP approach will attribute a given contribution to each selected input variable. The approach is able to identify input variables of low influence, but it cannot handle "hidden factors". This is now better underlined in Sect. 3.3 as follows: "This has several implications in the MME context: (1) any input will be assigned a Shapley value (defined by Eq. 2); (2) if $\mu_i^* = 0$, it indicates the absence of influence of the i^{th} input (related to the 'dummy player' property of the method); (3) the sum of the inputs' contributions is guaranteed to be exactly $f(\mathbf{x}^*) - \mu_0$ (related to the 'efficiency' property of the method). This also means that the selection of the input variables in the local importance analysis is an important step because the quantified contributions are dependent on the choice of which input variables are included in the analysis (see discussion in Sect. 5)."

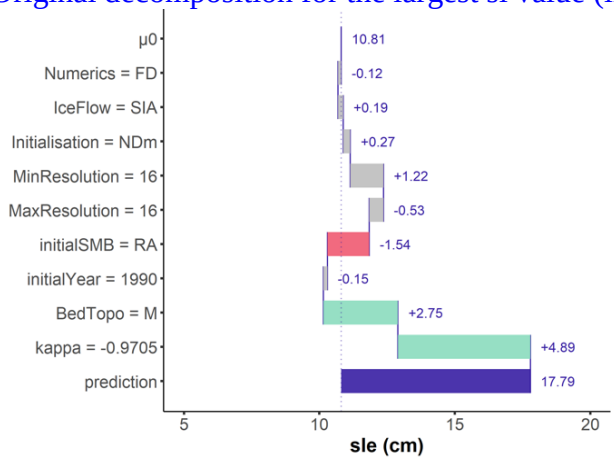
Regarding the selection of the factors in our analysis, we agree that our preliminary selection seems quite *ad-hoc*: this was primarily motivated by "having a sufficient number of variation across the models" to be able to validate the assumption of using a ML model (Sect. 3.2). This is now more clearly specified in Sect. 2 as follows: "Note that some preliminary groupings of categories were carried out to ensure a minimum of variation across the experiments with at least two experiments associated to a given category (specified in the last column of Table 1) which is needed to properly conduct the performance analysis of the ML model (see further details in Sect. 3.2)".

In more details, to validate our approach, we evaluate the ML model's predictive capability, i.e. its capability to predict "unseen" inputs' configurations, which have not been used for the training. Note that this is a more difficult task than checking that the ML model adequately reproduce the results used for its training. The analysis is here done by cross-validation procedures, which consist in randomly splitting the MME results into a training and a test set. When this procedure is carried out for categorical variables, care should be taken when the frequency of categories is unbalanced. For instance bed topography presents a higher number of cases in the "M" category (~90% of the MME results). In this case, during the cross-validation procedure, it is likely that the randomly-defined test set could content members with categories that are not present in the training set, hence making the ML model unable to extrapolate.

To overcome the afore-described problem, we propose to rely on a leave-one-out cross validation procedure instead of a 10-fold one, because it is less data-demanding (one MME result is extracted at each iteration instead of 5 ones). By doing so, we are now able to implement Referee #1's recommendation and to include all the input variables that are commonly shared across the Model in Table A1. Thus, the analysis has been re-conducted with these 3 new variables (maximum grid size, initial year and type of bed topography dataset) and the results have been re-analysed.
As expected, the values of the contributions changed by including these new variables, but overall our conclusions about the most influential variables remain unchanged. See an example below with the decomposition of the largest *sl* value in 2100.



(a) Original decomposition for the largest sl value (in 2100)



(b) New decomposition for the largest sl value (in 2100)

Despite the inclusion of additional variables, we still show that

- κ is a major contributor to *sl* especially in the second half of the century;
- Changing the type of numerical scheme (finite element or difference) remain the least impactful modelling assumptions;
- The influence of grid size should not be overlooked.

Given the importance of this selection, we now specify that:

- Sect 3.1: "We assume that the different models (part of the MME) share the same characteristics corresponding to p different modelling assumptions (e.g. choice in initial SMB / ice flow formulation, value of the grid size, etc.). In our case p=9 (see Sect. 2)";
- Figure 3: we add the pre-selection of the variables as a preliminary step (see the reply below);

- Sect. 3.3: "The interest is that the sum of the Shapley values for the different inputs is equal to the difference between the prediction and the global average prediction $\sum_{i=1}^{p} \mu_i^* = f(\mathbf{x}^*) - \mu_0$, which means that the part of the prediction value, which is not explained by the global mean prediction, is totally explained by the inputs (Aas et al., 2021: Appendix A). This has several implications in the MME context: (1) any input will be assigned a Shapley value (defined by Eq. 2); (2) if $\mu_i^* = 0$, it indicates the absence of influence of the i^{th} input (related to the 'dummy player' property of the method); (3) the sum of the inputs' contributions is guaranteed to be exactly $f(\mathbf{x}^*) - \mu_0$ (related to the 'efficiency' property of the method). This also means that the selection of the input variables in the local importance analysis is an important step because the quantified contributions are dependent on the choice of which input variables are included in the analysis (see discussion in Sect. 5)".
- Sect. 5: "the set of quantified contribution is always guaranteed, by construction (see Sect. 3.3), to add up to exactly the total *sl* projection. This has the practical advantage to ease the interpretation and communication of the results. However, this also means that the quantified contributions are themselves dependent on the choice of which input variables to include in the analysis. This raises the question of the impact of some missing input variables that are important with respect to *sl* prediction, i.e. the influence of some 'hidden factors'.".

In addition, we complement the analysis with a robustness study (see below the reply to comment (c)) to discuss the impact of the variables' selection.

(b) Ultimately, I think much of the issue with (a) could potentially be solved by adding one more factor, which is the model itself (e.g., AWI-ISSM, NCAR-CISM, etc.). There are many modeling choices that are different between those models (e.g., how certain processes are parameterized, etc) which is not described by the six existing categories. Including this as a category would determine how large these inter-model differences are. Hopefully, they would be small contributors to differences, but if they aren't small, then this indicates that more effort is required to label these differences in a meaningful way. Right now, these hidden factors are likely being lumped in together with other factors where there are intermodel differences.

We are grateful for this suggestion. We agree that adding a new variable corresponding to "*ModelID*" (see 1st entry of Table A1: Appendix A) could reflect the inter-model differences. However, the problem with this approach is related to the low-to-moderate size of the MME (here 55 members) to perform the cross validation procedure.

As afore-described, during the random splitting of the cross validation procedure, it is likely that the randomly-defined test set could content members with *ModelID* level that are not present in the training set. Using here a leave-one-out cross validation procedure does not solve the problem because some *ModelID* categories are only present once in the MME. Therefore, in order to be able to properly evaluate the predictive capability of the ML model, we chose not to implement this option.

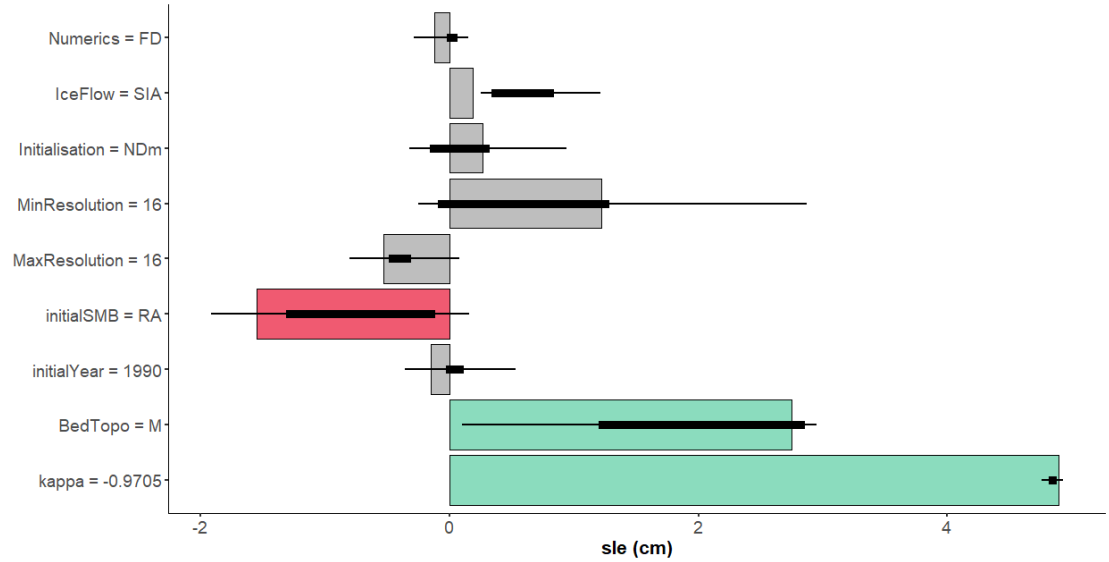
As second reason for not implementing this solution is that the effect of hidden factors is partly taken into account in the cross-validation procedure. The cross validation error already includes the error related to the selection of the input variables for building accurate ML model: if important variables (for the prediction error) are not included, the corresponding ML model is expected to perform poorly and to be assigned a high cross validation error. In the analysis of the results, we paid attention to nuance the SHAP results with respect to the cross validation error. This allows us to identify situations where the contributions cannot be distinguished from

the prediction error. This is now more clearly outlined in the core text and more particularly with a grey color in the diagnosis plot (see new Figure 6 provided below).

Finally, it is important to note that we performed the analysis with given MME results (GrIS MME was defined prior our study). A more ideal situation would have been to plan the analysis from the start, i.e. when designing the computer experiments. Yet, this solution can face some challenges underlined in our conclusion by referring to the discussion by Aschwanden et al. (2021).

(c) Ultimately, there needs to be much more discussion of the sensitivity of the results to the choice of factors to include in the analysis. This can be done by a leave-one-out validation exercise to see how the results change if one of the "known" factors is converted to a "hidden" factors. This would help other researchers hoping to do a similar analysis to decide how to make this choice of factors (or how much effort to put into labelling unlabelled factors) at the beginning of the exercise.
We thank Referee #1 for this valuable suggestion. We have implemented this option and we now use the results to support the discussion (Sect. 5) regarding the pre-requisite of our approach, i.e. the exhaustive selection of the input variables as follows:

"Second, the set of quantified contribution is always guaranteed, by construction (see Sect. 3.3), to add up to exactly the total *sl* projection. This has the practical advantage to ease the interpretation and communication of the results. However, this also means that the quantified contributions are themselves dependent on the choice of which input variables to include in the analysis. By construction, variables whose influence is negligible will be assigned a low contribution, but this does not address the issue of the impact of some missing input variables that are important for the *sl* prediction, i.e. the influence of some 'hidden factors'. The proposed cross validation error partly addresses this problem since high cross validation error reflects any difficulties in approximating the mathematical relationship between the inputs, which includes the afore-mentioned problem. To provide additional discussion, we conducted a robustness analysis by re-running the local attribution approach (and ML model fitting and selection) for the largest simulated *sl* value in 2100 (Case (a) in Figure 6); at each iteration, one of the nine input variables being removed in turn. Figure 10 provides the changes in the quantified contributions represented by a horizontal black error-bar. The comparison with the width of the horizontal coloured bar (representing the value of the original analysis including all nine input variables), confirms the high robustness of the large κ contribution (regardless of the selection of the input variables), and shows the lack of robustness of most of the input variables that were identified as non-significant with respect to the prediction error (coloured in grey). In addition, though the variability is higher, the contribution of the second and third larger contributor (initial SMB and bed topography dataset) show consistent results with the original study".



New Figure 10: Robustness analysis of the local importance analysis for the largest simulated *sl* value in 2100 (Case (a) in Figure 6). The horizontal coloured bars correspond to the quantified contributions by including all input variables (results of Fig. 6a). The endpoints of the thick and thin horizontal black error-bar are respectively the minimum / maximum, and the percentiles at 25 and 75% computed when iteratively excluding one of the nine input variables.

However, one disadvantage of this type of robustness analysis is the much higher computational cost, which prevents from a direct application to the whole MME results and requires further research work related to the active research area of 'sensitivity of the sensitivity analysis' (e.g., Razavi et al. 2021). This difficulty is also highlighted in the text at the end of Sect. 5.

Added reference

Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., et al.: The future of sensitivity analysis: an essential discipline for systems modeling and policy support, Environmental Modelling & Software, 137, 104954, 2021.

3. One aspect of the procedure that I found confusing is the procedure for training and then selecting the ML models. First, it is unclear why an ML surrogate is need. The manuscript says "our knowledge on the mathematical relationship f(.) is only partial and based on the n MME results." I'm not exactly clear on why this is a problem, so more explanation would be appreciated.

We agree that the use of ML models deserves further justification.

First, it should be noted that the exact computation of the Shapley value (Eq. 2) implies covering all subsets *S*, which grows exponentially with the number of factors denoted *k*, i.e. 2^k. In the MME results, all these subsets are not present, and implies using a surrogate model (i.e. the ML model) in place of the true function f to overcome this problem.

Second, computing the exact Shapley values for a single prediction requires solving integrals of the type in Eq. (3), which are of dimension 1 to $k-1$.

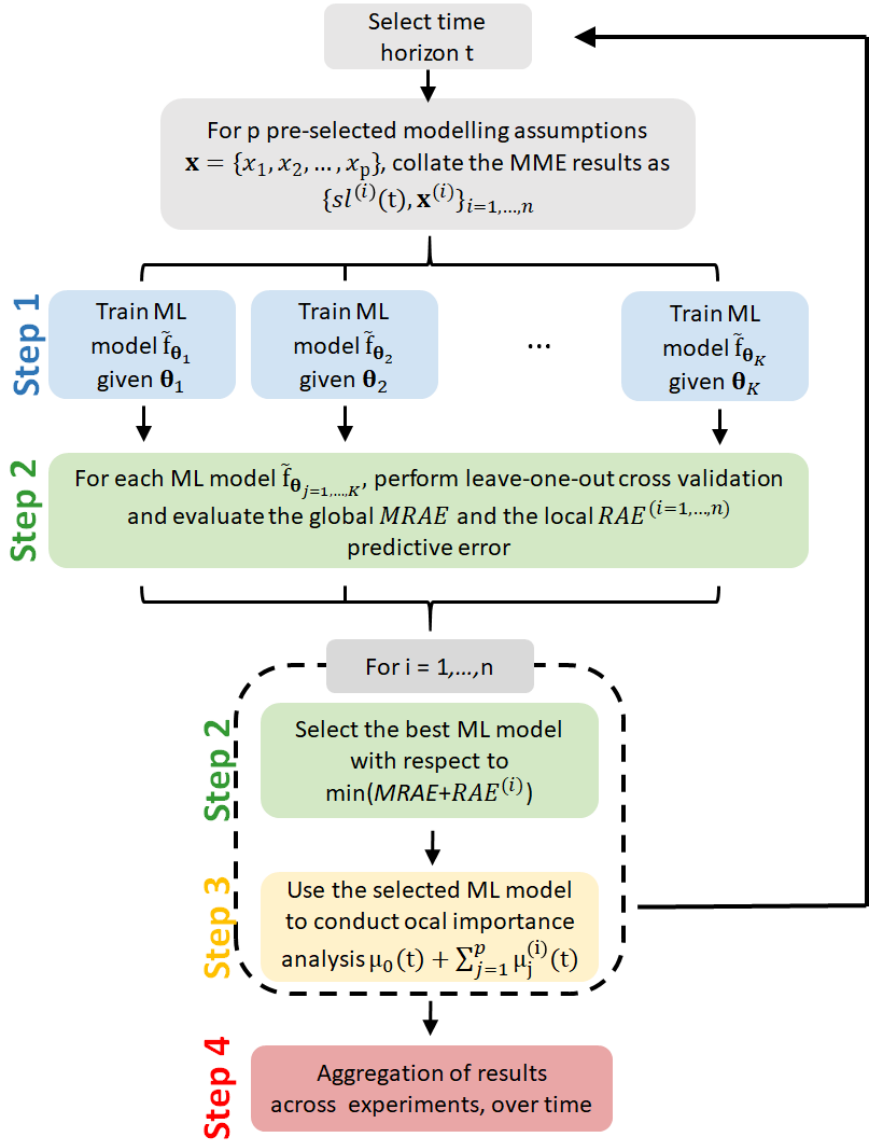
$$\mathrm{val}(S) = \mathrm{E}\left(\tilde{f}(\mathbf{x})\middle|\mathbf{x}_S = \mathbf{x}_S^*\right) = \mathrm{E}\left(\tilde{f}(\mathbf{x}_{\bar{S}}, \mathbf{x}_S^*)\middle|\mathbf{x}_S = \mathbf{x}_S^*\right) = \int \tilde{f}(\mathbf{x}_{\bar{S}}, \mathbf{x}_S^*) p(\mathbf{x}_{\bar{S}}|\mathbf{x}_S = \mathbf{x}_S^*) d\mathbf{x}_{\bar{S}}\,, \quad (3)$$

To account for dependence among the inputs, a sampling-based approach is here used (as described by Aas et al., 2021). Again, this means that the function f should be evaluated for inputs' configurations that are not necessarily present in the MME results, and the use of a ML model is here needed.

These explanations have been added at the end of Sect. 3.3.

Furthermore, you then fit a range of ML models and for each MML result, you pick a different ML model fit depending on which does a better job at reproducing this particular result. It is unclear why you do it in this way as opposed to just picking the ML model which best fits all the results. In particular, this seems to conflict with your prior statement that you need knowledge on "the mathematical relationship f(.)" in order to perform the explanation part of the procedure. Overall, this whole aspect of the procedure is quite confusing and could do with clearer explanations as to its purpose in the overall study.

We thank Referee #1 for this comment which led us to re-think our procedure. The best performing ML model is now selected with respect to two performance indicators: (1) the mean relative absolute error (denoted *MRAE*), which is a global indicator computed with results of leave-one-out cross validation procedure (Sect. 3.2); (2) the relative absolute error (denoted *RAE*) for the considered MME result (as outlined in the diagnostic plot in Fig. 6). Sect. 3.2 has been re-written in this sense and the workflow description has been updated (Sect. 3.1 and Figure 3).



New Figure 3: Schematic overview of the different steps of the procedure.

Note that a practical implication is that the level of prediction error between the original and this new analysis may differ. In particular, we achieved lower prediction error for the case (a) in Fig. 6 because in the original study, we did not account for the global performance indicator for selected the best performing ML model.

4. Similar to point #3, the introduction of the ML model quality metrics (Q, RAE, MAE) is confusing. It is unclear exactly how each of these metrics is used in order to perform model selection. Is MAE even used? If not, then its not clear why it is introduced.
We agree that some confusion may arise and we have clarified the description by focusing only on the performance indicators / quality metrics related to the absolute errors, either local metrics, *AE* (and its normalized variant relative absolute error *RAE*), and global metrics, *MAE* (and its normalized variant *MRAE*). Q^2 was thus removed from the analysis. Sect. 3.2 has been re-written in this sense.

To summarise, depending on the situation and on the level of sensitivity analysis (Level 1-3, described in Sect. 3.1), a different error metric is used.

- To measure the global predictive capability, we use *MRAE*;
- To measure the local predictive capability, we use *RAE*;
- To conduct the level 1 of the analysis (see Sect. 4.3.1 and Fig. 6), we use *AE* because it is directly expressed in physical unit (i.e. in centimeters);
- To conduct the level 2 of the analysis (see Sect. 4.3.2 and Figs. 7,8), we compare the importance given different MME members. For sake of comparability, we thus use *MAE*;
- To conduct the level 3 of the analysis (see Sect. 4.3.3 and Fig. 9), we preferably analyse the absolute value of a normalised version of μ^*, i.e. $\mu_n(\mathrm{t}) = \mu^*(\mathrm{t})/(sl^*(\mathrm{t}) - \mu_0(\mathrm{t}))$ to be able to compare the influence between the different predictions across time. The error indicator is thus adapted to this situation, namely we define a variant of $RAE(\mathrm{t})$, namely $RAE_n(\mathrm{t}) = \left|\frac{e(\mathrm{t})}{sl(\mathrm{t})-\mu_0(\mathrm{t})}\right|$.

These explanations were added at the end of Sect. 3.2.

Other points:
Line 10-12: confusing sentence
We have simplified the phrasing as follows: “To this end, we adopt the local attribution approach developed in the machine learning community known as ‘SHAP’ (SHapley Additive exPlanation).”

Line 24: can be->are
This sentence has been rephrased as follows: “To cover the large spectrum of uncertainties that impact the outcomes of these numerical models, a popular approach is to perform common sets of numerical experiments by considering various initial conditions and/or model design”

Line 28: ice-sheet->ice sheets
This has been corrected.

Line 33: the key to what?
And
Line 33: its not clear that this method can tell us why the numerical delivered some particular results (in the absolute sense), but rather why it differs from other model results

We have clarified this sentence as follows: “Therefore, the key to improving interpretation of sea-level projections is not only to deliver modelling results, but also to explain why the numerical model delivered some particular results given the set of chosen modelling assumptions”

Line 35: "this" view? which?
This term has been removed.

Line 39: how particular modelling
This has been corrected.

Line 43: positive with respect to model sensitivity?
We have rephrased as follows: ‘if the measure of local importance is positive, then the considered modelling […]”

Line 60: local importance of what?
This has been rephrased as follows: “our objective is to compute measures of local importance for each considered modelling assumptions”

Line 80: Is this a test case or just the point of the study?
This has been rephrased as follows: ‘To test our approach, we use a test case defined based on the MME study”.

Line 92: more discussion could be useful about why GCM choice is not included. I can see that this would be done to reduce complexity, but ultimately GCM choice is important to the full ensemble of results from ISMIP. At a minimum, better explanation of why this difference is omitted would be helpful.
We agree with Referee #1 that this choice should be clarified. In the introductive part of Sect. 2, we explain that *“*so that a sufficient number of MME results are available to validate our approach. For this case, a total of 55 numerical experiments were extracted to analyse the time evolution of sea level changes with respect to 2015 (Fig. 1); each of these results is associated with different modelling choices represented by different ISMs that are described in Appendix A: Table A1. In addition, for the selected AOGCM, we are able to analyse the sensitivity to the parameter based on the availability the experiments denoted exp05, exp09 and exp10 in Goelzer et al. 2020: Table 1”.

Though we find the suggestion to test the impact of the GCM choice very relevant, it is worth underlying that our primary objective is the test the feasibility of our approach for a given MME study, 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), additional tests can hardly be done here. Therefore, we have highlighted in the conclusion the need for multiplying the application cases (in particular by varying the AOGCM and the RCP choice).

Line 105: explain further about "preliminary groupings"
We have added some clarifications as follows: “Note that some preliminary groupings of categories were carried out to ensure a minimum of variation across the experiments with at least two experiments associated to a given category (specified in the last column of Table 1)

which is needed to properly conduct the performance analysis of the ML model (see further details in Sect. 3.2)".

Line 127: the f(.) notation is generally very confusing to me, not sure what "." means in this context
This confusing notation has been removed in the whole manuscript as well as in Figure 3.

Line 151: why are these particular ML models used? Also, I think its a stretch to call a linear regression an ML model (I get that its a simple benchmark, but most people would just call it a statistical model)
We agree that this choice deserves further explanations. We chose tree-based ML models because they have shown to perform the best on the type of data we are dealing with i.e. tabular data. See e.g. the recent extensive benchmark by Grinsztajn et al. (2022). This is now clarified in the description of Step 1 of the procedure.

Finally, note that the categorization of linear models as ML models is used in the core textonly for the sake of terminological homogenization. We agree that linear models are statistical models in themselves.

Added reference
Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data?. arXiv preprint arXiv:2207.08815.

Line 157-158: not sure what this sentence means
This has been rephrased as follows: ", i.e. different values of the hyperparameters θ can be defined for each of the considered ML models."

Line 163: explain what you mean by "strong theoretical basis"
We have added a reference to Aas et al., 2021:Appendix A where a description from a modeller's perspective can be found.

Line 181: the objective of this section
This has been corrected.

Line 181: ...by h_theta through the use of...
We are not sure what the Reviewer #1 means and have specified that f_theta is here a ML model.

Section 3.3: I generally really liked this explanation of Shapley approach
We thank Referee #1 for the positive statement; we hope that this description will also be warmly welcomed in the community of ice-sheet modelling.

Line 237: from the global mean prediction to what?
This part of the sentence has been replaced by a more explicit term, namely "$E(f(\mathbf{x}))$".

Line 235: The definition of the Shapley values guarantees that the sum...
We have added some additional description as follows: "This has several implications in the MME context: (1) any input will be assigned a Shapley value (defined by Eq. 2); (2) if $\mu_i^* = 0$, it indicates the absence of influence of the i^{th} input (related to the 'dummy player' property of the method); (3) the sum of the inputs' contributions is guaranteed to be exactly $f(\mathbf{x}^*) - \mu_0$

(related to the ‘efficiency’ property of the method). This also means that the selection of the input variables in the local importance analysis is an important step because the quantified contributions are dependent on the choice of which input variables are included in the analysis (see discussion in Sect. 5)”.

Line 242: where is equation (5)?
This is a mistake. This has been corrected.

Section 3.4: it would help here to give an example of the sort of dependence between inputs you are thinking of. I know you talk about this later in the results, but its hard to understand what you mean by dependence here without a concrete example.
We have provided an example as follows: “A commonly encountered example is when the values for the minimum and maximum grid sizes are correlated.”

Line 266: the procedure describe in...
This has been corrected.

Line 269: sensitivity of projected SLR to modeling assumptions at difference levels
This has been rephrased as follows: “we summarise the results to provide different levels (detailed in Sect. 3.1) of information on sensitivity (Steps 3-4, Sect. 4.3)”.

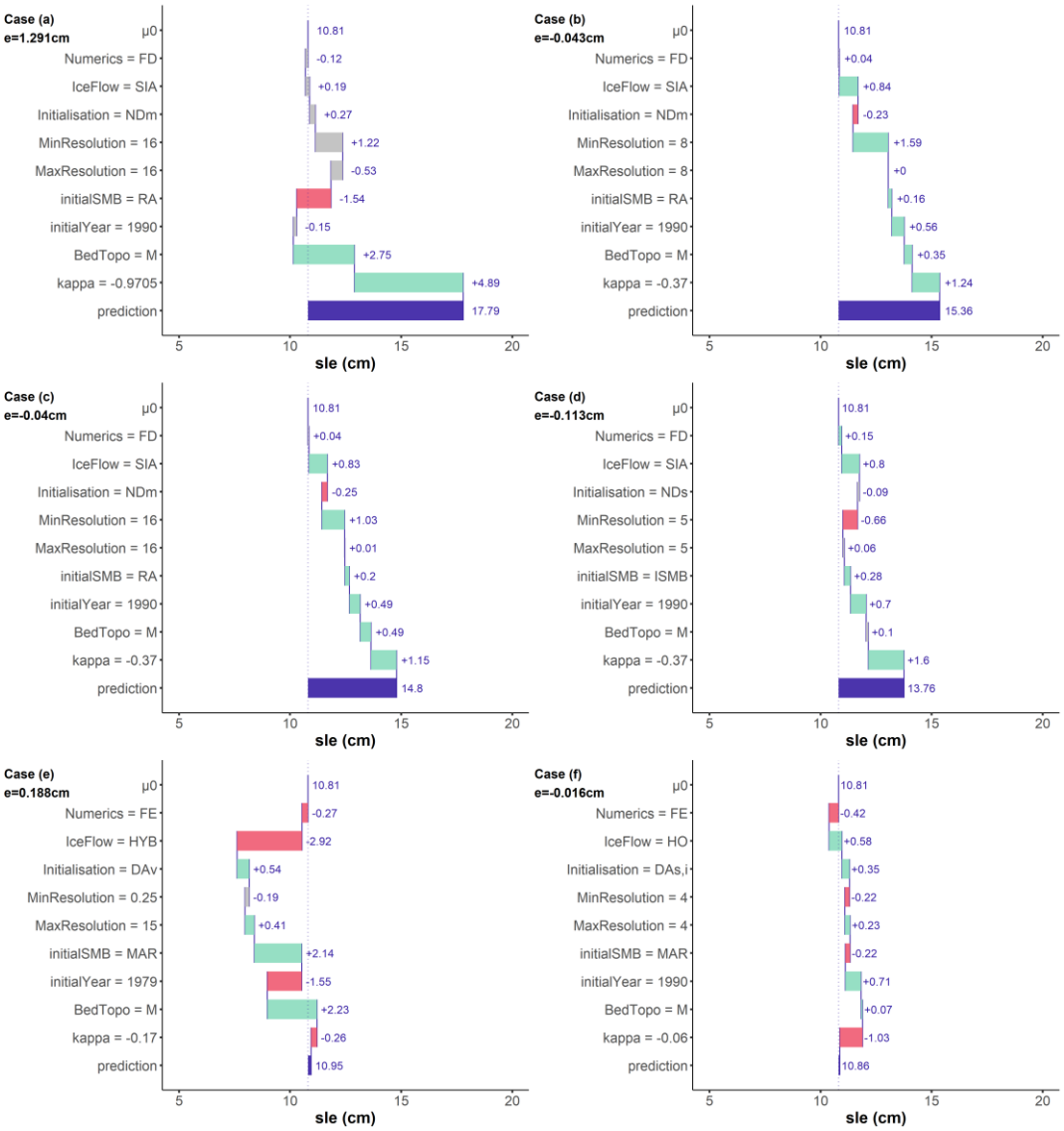
Line 295-298: The results described here are confusing, because from Figure 5a, it seems like the best performing models are all RF.
This is related to the use of Q^2 as an error metric. We now analyse the mean relative absolute error *MRAE* which is more sensitive to small changes between each ML model performance.

Line 304: I could not find anything outlined in orange (maybe I'm not looking hard enough...)
This is a mistake and the right color is red.

Line 319: the use of "perform" here is confusing
This has been rephrased as follows “In this section, we first compute the measures of local importance for each experiment in the MIROC5,RCP8.5-forced GrIS MME […].”

Line 324: sensitivity of what?
We have here specified sea level change.

Figure 6: Can you add a label to each panel indicating which model experiment it is. Also, it seems like it would make sense to color bars which are "negligible" (in terms of being less than the error) as grey or something similar to indicate that their importance is below the threshold that you have defined as being interpretable. This would help reduce the amount of information the reader has to absorb when looking at this very interesting but dense figure.
We agree with this suggestion and Figure 6 (and the caption) has been modified in alignment with this suggestion.



New Figure 6: Diagnostic of particular ML-based *sl* predictions using SHAP for year 2100 considering six different settings of the modelling choices (indicated in the vertical axis). The horizontal blue bar corresponds to the ML-based *sl* prediction (the difference with the true value is indicated by the error term *e* expressed in cm SLE). Each row shows how the positive (green bar) or negative (red bar) contribution of each input moves the prediction from μ_0, i.e. the unconditional expectation of *sl*. The grey colour indicates that the contribution cannot be distinguished from the predictive error, because its absolute value is below the absolute error.

Line 359: large absolute value of kappa
This has been corrected.

Line 365: the results are non-existent
We apologize but we do not understand this comment.

Fig 7 and Line 371: I'm not sure the smooth fitting is defensible. There is clearly overfitting happening here, and you even talk about (for example) mu=+3 cm in Figure 7b, when such values are clearly only a product of the smoothing in an area of parameter space with no simulation.
Figure 7b: the dip in mu between 2 and 4 km resolution make me wonder whether the modeling of input dependence has completely removed this effect. My guess is that the three results with 3 km resolution are all a single type of model which exhibits idiosyncratic behavior for other

reasons beside resolution. This gets to point 2b above, but also why I think it is inadvisable to fit a smooth curve to these plots.
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 the region where their influence becomes significant, i.e. for minimum and maximum grid size >5km and 16m respectively. We agree that the interval 2-4km presents a complex behavior that is related to the low number of MME results here. Adding more computer experiments in this region is now formulated as a recommendation in Sect. 5.

Line 383-384: Confusing sentence
This sentence has been re-written as follows: "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 region of minimum and maximum grid sizes."

Figure 9: this figure needs better explanation as I am not sure what the box plots represent compared to the red dots, etc.
We have clarified in the text Sect. 4.3.3 as well as in Fig. 9 caption how to read the boxplot with respect to the red crosses as follows: "if the boxplot depicted in Fig. 9 does not overlap with the region defined by the interval between the lower and the upper red cross, this means that the influence measured by $|\mu_n|$ can be considered significant with respect to the ML prediction error".

Line 410: which provides narratives about the role of various modeling choices in generating inter-model differences in the GrIS MME
Thank you for the suggestion. This has been added.

Line 419: affected by too coarse grids (here coarser than 5 km)
Thank you for the suggestion. This has been added.

Line 424: could not have
This has been added.

Line 431: comparison of ANOVA with
This has been added.

Line 435: it would be useful to quantify the extent to which this problem is alleviated or not - several prior comments relate to the possibility that input dependence is still present in results
For this problem, we build on the very satisfactory results of the SHAP-CTREE approach in the previous studies by Redelmeier et al. (2020) and Aas et al. (2021). Both studies have shown how different the results can be when dependency is not taken into account. However, this raises the practical question of the most appropriate mathematical approach to this problem. We believe that this open question deserves a more extensive analysis that is beyond our scope here.

Therefore, we have underlined in Sect. 5: "Here the SHAP-CTREE combined approach developed by Redelmeier et al. (2020) helps alleviate this problem by explicitly incorporating the dependence in the computation of the Shapley values (Sect. 3.4; see also Aas et al. (2021) for an extensive study of this problem). In light of the different algorithms available in the literature (Aas et al., 2021; Frye et al., 2020), an interesting line of future research could focus

on a more systematic analysis of the effect of input dependence in the context of MME and on defining clear recommendations on how to treat it".

Added reference:

Frye, C., de Mijolla, D., Cowton, L., Stanley, M., and Feige, I.: Shapley-based explainability on the data manifold, 2020. [arXiv:2006.01272].

Line 436: for the high performance of our approach is
This has been corrected.

Line 443: In this study, we described the use
This has been corrected.

Line 444: assumptions in sea-level projections
This has been corrected.

Line 455: global effects of what?
This has been clarified as "the global effects of the modelling assumptions".

Orleans,
September 12th, 2022

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