Improving interpretation of sea-level projections through a machine-learning-based local explanation approach

Jeremy Rohmer[1], Remi Thieblemont[1], Goneri Le Cozannet[1], Heiko Goelzer[2], Gael Durand[3]

[1]BRGM, 3 av. C. Guillemin, 45060, Orléans, France
[2]NORCE Norwegian Research Centre, Bjerknes Centre for Climate Research, Bergen, Norway
[3]Univ. Grenoble Alpes, CNRS, IRD, Grenoble INP, IGE, 38000 Grenoble, France

Correspondence to: Jeremy Rohmer (j.rohmer@brgm.fr)

Abstract. Process-based projections of the sea-level contribution from land ice components are often obtained from simulations using a complex chain of numerical models. Because of their importance in supporting the decision-making process for coastal risk assessment and adaptation, improving the interpretability of these projections is of great interest. To this end, we adopt the local attribution approach developed in the machine learning community known as 'SHAP' (SHapley Additive exPlanation). We apply our methodology to a subset of the multi-model ensemble study of the future contribution of the Greenland ice sheet to sea-level, taking into account different modelling choices related to (1) numerical implementation, (2) initial conditions, (3) modelling of ice-sheet processes, and (4) environmental forcing. This allows us to quantify the influence of particular modelling decisions, which is directly expressed in terms of sea level change contribution. This type of diagnosis can be performed on any member of the ensemble, and we show in the Greenland case how the aggregation of the local attribution analyses can help guide future model development as well as scientific interpretation, particularly with regard to spatial model resolution and to retreat parametrisation.

1 Introduction

Process-based projections of ice sheets' contributions to sea-level changes generally rely on numerical models that simulate the gravity-driven flow of ice under a given environmental (atmospheric and oceanic) forcing derived from Atmosphere–Ocean General Circulation Model (AOGCM) output. 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 a range of forcing conditions (e.g. Barthel et al., 2020), various initial conditions and/or model design (i.e. different choices in the modelling assumptions including different ice sheet model (ISM) formulations, different input parameters' values, etc.) within a multi-model ensemble (MME) approach. This results in an ensemble of realizations, named ensemble members. Recent MME studies have analysed, within the Ice Sheet Model Intercomparison Project for CMIP6 (ISMIP6), the future evolution of the ice sheets of Greenland (Goelzer et al., 2018; 2020), and Antarctica (Seroussi et al., 2020).

Providing such projections using numerical models is challenging because the considered physical processes are highly complex, and may involve nonlinear feedbacks operating on a wide variety of time scales. Due to the importance of these projections to support coastal adaptation (Kopp et al., 2019), improving their interpretability is of high interest.

When dealing with interpretability, the key is generally 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 (Molnar, 2022). Commonly-used approaches to improve interpretability usually focus on measuring the importance of modelling assumptions 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 a second approach named "local" because it aims at measuring the importance of the input variables locally at the level of individual observations (and not globally across all observations unlike the first approach). 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. This is the local attribution approach adopted by the machine learning community (e.g., Murdoch et al., 2019), and named "situational" in the statistical literature (Achen, 1982). ~~It aims to better understand why a given instance of the modelling assumptions leads to a certain prediction.~~ As described by Štrumbelj and Kononenko (2014), if the measure of local importance is positive, then the considered modelling assumption has a positive contribution (increases the prediction for this particular instance), if it is negative, it has a negative contribution (decreases the prediction), and if it is 0, it has no contribution.

A possible local attribution approach can follow a 'one-factor-at-a-time' procedure, which consists of analysing the effect of varying one model input factor at a time while keeping all other fixed (see an example performed by Edwards et al., 2021). Though simple and efficient, this approach presents several shortcomings (dependence to the chosen base case, to the magnitude of variations, failure when the model is non-linear, etc. see an in-depth analysis by Štrumbelj and Kononenko (2014)). A more generic approach has emerged in the domain of explainable machine learning (Murdoch et al., 2019), named SHapley Additive exPlanation SHAP (Lundberg and Lee, 2017). SHAP has successfully been used in many domains of application, such as finance (Bussmann et al., 2021), medicine (Jothi and Husain, 2021), land-use change modelling (Wieland et al., 2021), mapping of tropospheric ozone (Betancourt et al., 2022), digital soil mapping (Padarian et al., 2021), etc.

SHAP builds on the Shapley values that were originally developed in the cooperative game theory for "fairly" distributing the total gains to the players, assuming that they all collaborate (Shapley, 1953). Making the analogy between a particular prediction and the total gains, SHAP allows breaking down any prediction as an exact sum of the modelling assumptions'

contribution with easily interpretable properties (see a formal definition in Sect. 3); each contribution then reflects the influence of the considered modelling assumptions for the particular prediction.

In this study, our objective is to compute measures of local importance for each considered modelling assumption using SHAP applied to MME of sea-level projections. Applying SHAP in this context faces however several difficulties. First, it is not the prediction provided by the modelling chain (used to generate the MME) that is decomposed by SHAP, but it is a machine-learning-based proxy (named ML model) that relates the modelling assumptions (termed as 'inputs' in the following) to the equivalent sea-level changes (denoted *sl*). Validating the use of this proxy is one key prerequisite of the approach. Second, building the ML model relies on the analysis of the available MME results, which are limited (typically up to 50-100 ensemble members), due to the large computational time cost of the modelling chain. This results in MMEs that are incomplete and unbalanced: i.e. several combinations of modelling assumptions are missing in the MME while some are more frequent than others. Statistically, this incompleteness and unbalanced design might result in statistical dependence among the input variables (related to the modelling assumptions). Overlooking this dependence structure might mislead the interpretation of the inputs' individual influence; see an extensive discussion by Do and Razavi (2020). To overcome the afore-described difficulties, we propose a SHAP-based procedure combined with cross-validation procedure (Hastie et al., 2009) and appropriate techniques for modelling the dependence (Aas et al., 2021; Redelmeier et al., 2020). Through aggregation of the SHAP-based local explanations, we further show how they can be helpful for both improving the scientific interpretation and guiding future model developments. The proposed procedure is applied to sea-level projections for the Greenland ice sheet (Goelzer et al., 2020) by considering the time evolution of sea-level contributions.

The paper is organized as follows. We first describe the sea-level projections used as application case and the corresponding design of numerical experiments (Sect. 2). In Sect. 3, we provide further details in the statistical methods that are used to estimate the local explanations. In Sect. 4, we apply the methods and provide some approaches to combine the local explanations to get global understanding of the MME results across time.

2 Multi-model ensemble case study

To test our approach, we define a case study based on the MME study carried out by Goelzer et al. (2020) in the framework of the ISMIP6 initiative. In the following, we only provide a brief summary of the GrIS MME dataset and the interested reader is invited to refer to Goelzer et al. (2020) and references therein for further details.

To compute the annual time evolution of sea-level contributions from the Greenland ice sheet (GrIS) up to 2100, the modelling chain combines different models: (1) a number of AOGCMs that produce climate projections according to given greenhouse gas forcing scenarios; (2) a Regional Climate Model (RCM) that locally downscales the AOGCM forcing to the GrIS surface; (3) a range of ISM models (initialised to reproduce the present-day state of the GrIS as best as possible from a given initial year to end of 2014) that produce projections of ice mass changes and sea-level contributions. Given bed topography across the ice-ocean margin around Greenland, the ISMs are forced by surface mass balance (denoted *SMB*)

anomalies from the atmospheric RCM-derived forcing and by an empirically derived parameterisation that relates changes in meltwater runoff from the RCM and ocean temperature changes from the AOGCMs to the retreat of tide-water glaciers (Slater et al., 2020). The parameter that controls retreat is denoted κ and is used to sample uncertainty in the parameterisation (Slater et al., 2019).

As the primary objective of this work is to evaluate the relevance of the 'SHAP' approach, we focus on a subset of the original GrIS MME study based on one AOGCM, namely MIROC5 forced under the most impactful climate scenario RCP8.5, because 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 κ based on the availability of the numerical experiments denoted *exp05*, *exp09* and *exp10* in Table 1 of Goelzer et al. 2020.

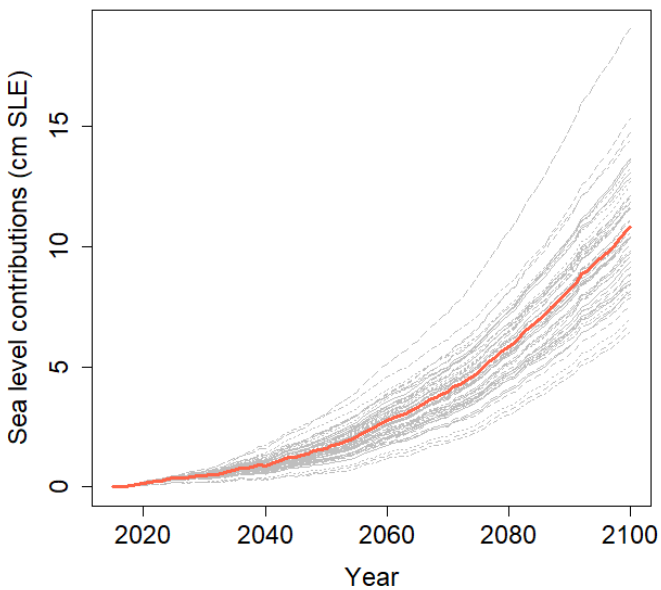

Figure 1: (a) Time evolution of the sea level contribution (with respect to 2015) from the Greenland ice-sheet (in cm sea-level equivalent, SLE). The results are the MIROC5,RCP8.5-forced MME of Goelzer et al . (2020). The red straight line is the temporal ensemble mean.

The analysis is focused on nine main modelling assumptions related to different aspects of the modelling chain (Table 1), namely numerical implementation, initial conditions, modelling of ice-sheet processes, and environmental forcing. Only the modelling assumptions that are commonly shared by all models described by Goelzer et al. (2020): Appendix A were considered, i.e. without empty entry in Table A1~~ and with a sufficient number of variation across the models~~. 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).

Table 1. Modelling assumptions considered in the MIROC5,RCP8.5-forced GrIS MME

Type	Modelling assumption	Symbol	Value range / Categories	Grouping of categories
Initial conditions	type of initialisation method	*init*	Data assimilation of velocity (*DAv*); Nudging to ice mask (*NDm*); Nudging to surface elevation (*NDs*), and a category denoted *DAs,i* that group data assimilation of surface elevation, data assimilation of ice thickness, spin-up, and transient glacial cycles	
Initial conditions	initial surface mass balance (SMB)	*SMB*	Different RCMs among RACMO, either RACMO2.1 or 2.3 (*RA*); *MAR*; HIRHAM5 (*HIR*); and implied SMB (*ISMB*).	Experiments that use climatology and historical spin-up from BOX but historical experiment from either MAR (or RACMO) anomalies were assigned to *MAR* (respectively *RA*) category
Initial conditions	Initial year that is used to compute the present-day until the end of 2014	*Year0*	From 1979 to 2008	
Numerical implementation	Numerical method	*Num*	Finite difference (*FD*) or Finite element (*FE*).	Only one modelling team has used a numerical scheme of finite volume type: this choice was grouped with *FE*
Numerical implementation	Minimum value of the grid size	*res_min*	From 0.25 to 16 km	
Numerical implementation	Maximum value of the grid size	*res_max*	From 0.90 to 30 km	
Ice sheet processes	Type of ice flow	*iceFlow*	Shallow-ice approximation (*SIA*), shallow-shelf approximation (*SSA*), higher order (*HO*), *SIA* and *SSA* combined (*HYB*)	
Ice sheet processes	Bed topography	*Bed*	Two datasets are considered: BedMachine v3 by Morlighem et al. (2017) (*M*); and the one by Bamber et al. (2013) (*B*)	

Environmental forcing	value of the retreat parameter	κ	From -0.9705 to +0.0079 km.(m^3.s^{-1})$^{-0.4}$ °C	

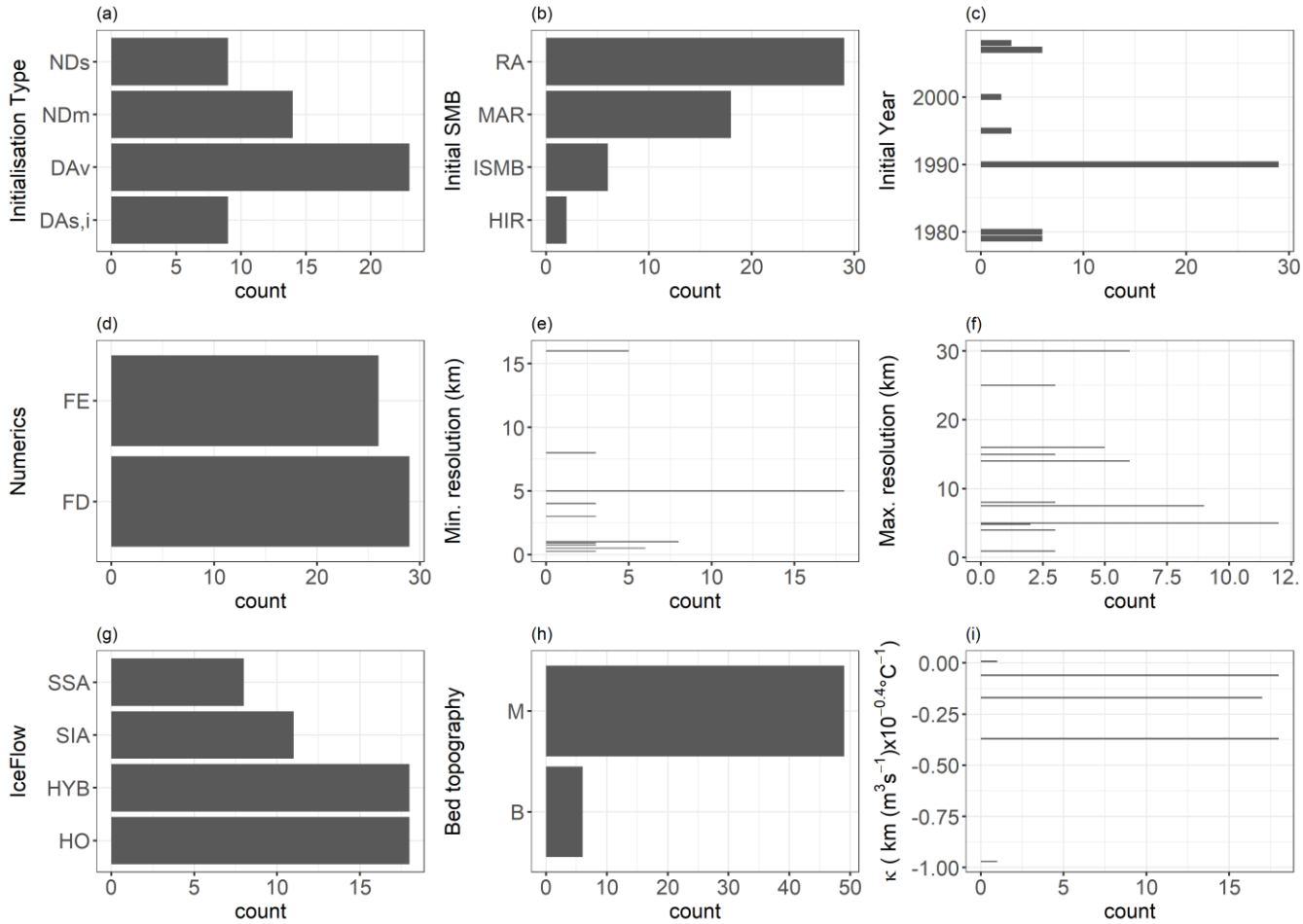

Figure 2: Count number of the MIROC5,RCP8.5-forced GrIS MME members with respect to the different modelling assumptions described in Table 1.

In the following, we name "inputs" the choices made for each of these modelling assumptions. One input setting defines an experiment of the MME. Formally, the inputs are either treated as continuous variables (for κ, minimum and maximum resolution and initial year), or as categorical variables (for the five other ones). Figure 2 shows that the design of experiments is unbalanced: some categories (like *RA* for instance, Fig. 2b) or some values (like minimum resolution at 5km, Fig. 2e) are more frequent than others. The design is also incomplete with large gaps in the histograms. This is for instance the case for κ between -0.9705 and -0.3700 km.(m^3.s^{-1})$^{-0.4}$ °C (Fig. 2i), because this parameter was sampled for only 3 different values by most models (the median, the 25% and the 75% percentile), and the additional 2 values were only sampled by one ISM.

3 Methods

3.1 Overall procedure

Let us consider $sl(\mathrm{t})$ the sea level change (with respect to a reference date) at a given time t that is numerically simulated from the chain of models, denoted f, described in Sect. 2. 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). To each of these modelling assumptions is assigned a random variable x. The vector of p input variables (p modelling assumptions) is denoted by $\mathbf{x} = \{x_1, x_2, \dots, x_\mathrm{p}\}$. We consider n different experiments; each of them associated to a particular $\mathbf{x}^{(\mathrm{i})}$. The MME results at a given time t are $\{sl^{(\mathrm{i})}(\mathrm{t}), \mathbf{x}^{(\mathrm{i})}\}_{\mathrm{i}=1,\dots,\mathrm{n}}$ with $sl^{(\mathrm{i})}(\mathrm{t}) = \mathrm{f}(\mathbf{x}^{(\mathrm{i})})$. This means that our knowledge on the mathematical relationship f is only partial and based on the n MME results. To overcome this difficulty, we replace f by a machine-learning-based proxy (named ML model) built using the MME results; the advantage being to make some predictions for input configurations that are not present in the original MME dataset at a low computation time cost. The ML model is denoted $\tilde{\mathrm{f}}_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta}$ correspond to the ML model's parameters (named hyperparameters, see Appendix B).

Given a specific setting $\mathbf{x}^*$ (i.e. an instance of modelling choices made by the modellers for each of the considered assumptions), we follow the additive feature attribution approach that has been developed for ML models (e.g., Štrumbelj and Kononenko, 2014; Lundberg and Lee, 2017). This approach proposes to improve the interpretability of a particular prediction $\mathrm{f}(\mathbf{x}^*)$ for a given time horizon t by decomposing it as a sum of the inputs' contributions $\mu_i^*(\mathrm{t})$ (specific to $\mathbf{x}^*$) as follows:

$$sl^*(\mathrm{t}) = \mathrm{f}(\mathbf{x}^*) \approx \tilde{\mathrm{f}}_{\boldsymbol{\theta}}(\mathbf{x}^*) = \mu_0(\mathrm{t}) + \sum_{\mathrm{j}=1}^{\mathrm{p}} \mu_\mathrm{j}^*(\mathrm{t}), \qquad (1)$$

where $\mu_0(\mathrm{t})$ (named base value) is a constant value (see definition in Sect. 3.3).

It is important to note that Eq. (1) does not aim to linearize f, but to compute the contribution of each input to the particular prediction value $\mathrm{f}(\mathbf{x}^*)$. This means that the decomposition provides insights into the influence of the particular instance of the inputs $\mathbf{x}^*$ relative to $\mathrm{f}(\mathbf{x}^*)$: (1) the absolute value of $\mu^*(\mathrm{t})$ informs on the magnitude of the influence at time t directly expressed in physical units (for instance in centimetres for sea level), which eases the interpretation; (2) the sign of $\mu^*(\mathrm{t})$ indicates the direction of the contribution, i.e. whether the considered modelling assumption pushes the prediction higher or lower than the base value $\mu_0(\mathrm{t})$.

In order to quantify $\mu^*(\mathrm{t})$ in Eq. 1, the different steps of the proposed approach (schematically represented in Fig. 3) are as follows.

Step 1 Build and train ML models. At a given time horizon t, a ML model $\tilde{\mathrm{f}}_{\boldsymbol{\theta}}$ is built using some supervised ML techniques (see Hastie et al., 2009 for an overview). We rely here on three types of ML models, namely linear regression model, denoted LIN (because of the simplicity of its implementation), and two tree-based approaches, random forest regression method, denoted RF (Breiman, 2001), and Extreme Gradient Boosting for regression denoted XGB (Chen and Guestrin

2016)), which have shown high performance in diverse benchmark exercises (e.g. Grinsztajn et al. (2022) and references therein). See Appendix B for further details on these techniques and their respective hyperparameters $\boldsymbol{\theta}$;

Step 2 Evaluate the predictive capability and select the best performing ML model. The decomposition described in Eq. 1 is only meaningful provided that the assumption of replacing f by $\tilde{f}_{\boldsymbol{\theta}}$ is valid. In this view, we propose to assess this assumption's validity by measuring the predictive capability of $\tilde{f}_{\boldsymbol{\theta}}$ using a leave-one-out cross validation procedure (Hastie et al., 2009). This validation is performed by considering the different parametrisations of the ML methods, i.e. the validation is performed by considering different values of the hyperparameters $\boldsymbol{\theta}$ for each of the considered ML models. Two indicators are computed, namely a local one related to the considered i^{th} MME result, which measures the relative absolute error (denoted $RAE^{(i)}$), and a global one (denoted $MRAE$) defined as the average value of the $RAE^{(i)}$ values computed across all n MME results. Then, for the ith MME result, the ML model that performs the best with respect to the minimum value of $MRAE + RAE^{(i)}$ (i.e. both globally and locally for the considered i^{th} MME result) is retained for the next step. The results of *Step 2* is also useful to characterize the ML prediction error. Further details are provided in Sect. 3.2;

Step 3 Local importance analysis. This step aims to perform the additive decomposition (Eq. 1) using the selected ML model. Among the different available methods (Molnar et al., 2019), we rely on the SHAP approach proposed by Lundberg and Lee (2017) because of its strong theoretical basis (see further details in Sect. 3.3 as well as Aas et al., 2021:Appendix A for a description from a modeller's perspective) as well as its multiple use in various application areas (see introduction). ~~Further details are provided in Sect. 3.3.~~ A special care is paid to the impact of the inputs' dependence by application of methods described in Sect. 3.4;

Step 4 Summarise local explanations. The local explanations are combined and aggregated to provide insights into the model structure and to inform on the sensitivity of *sl*(t) to the modelling assumptions at each time horizon t. Inspired by Lundberg et al. (2020), the sensitivity analysis is conducted at different levels:

- *Level 1 Locally at a given prediction time* by analysing the value and sign of μ_i^* for a particular experiment. An application is provided in Sect. 4.3.1;
- *Level 2 Model structure at a given prediction time* by analysing how the influence measured by μ_i^* (magnitude and sign) evolves as a function of the ith input value. An application is provided in Sect. 4.3.2;
- *Level 3 Globally over time* by analysing how the magnitude of the influence measured by $|\mu_i^*|$ evolves across time by considering all experiments. To be able to compare the influence between the different predictions across time, we preferably analyse the absolute value of a normalised version of μ^*, i.e. $\mu_n(t) = \mu^*(t)/(sl^*(t) - \mu_0(t))$. An application is provided in Sect. 4.3.3.

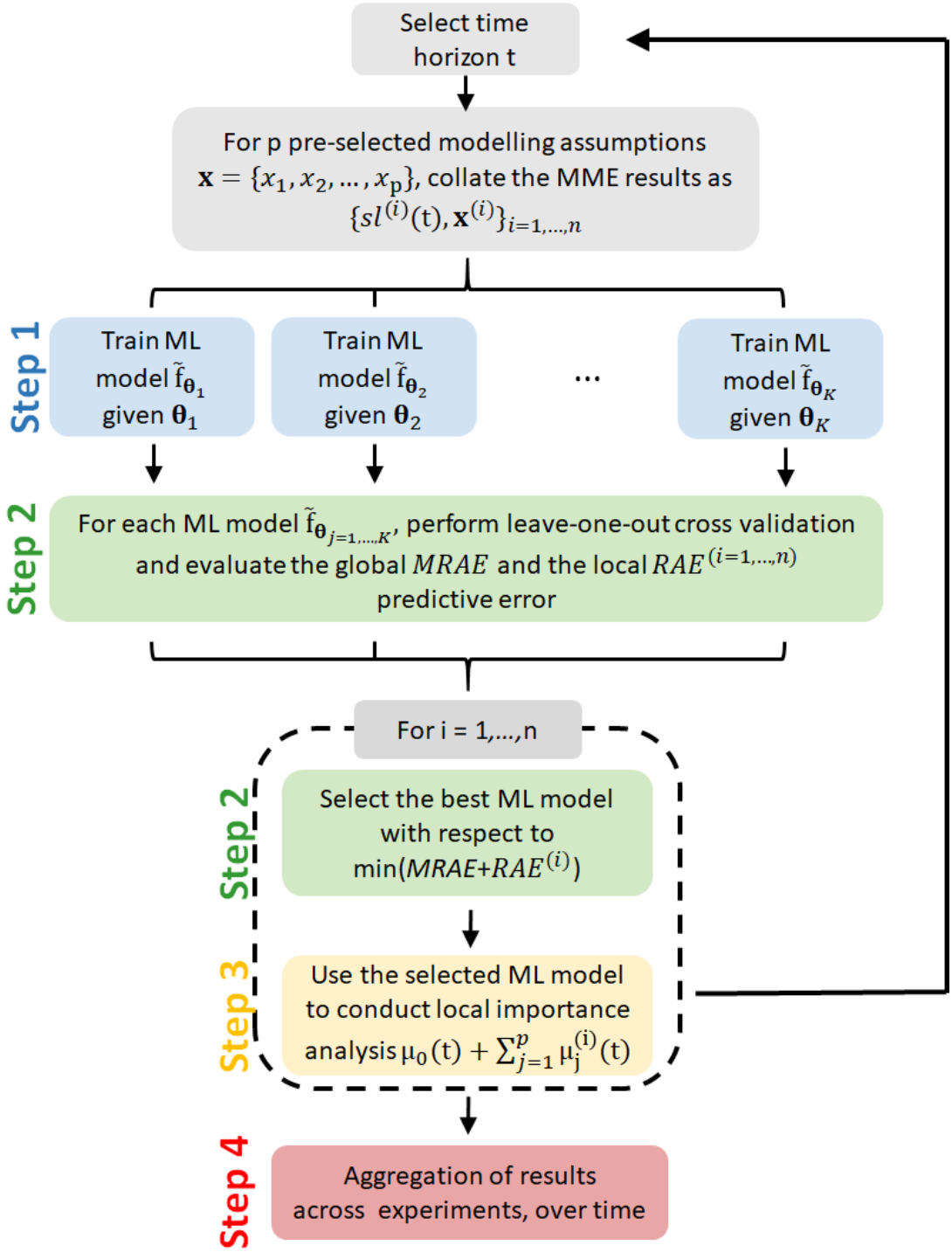

Figure 3: **Schematic overview of the different steps of the procedure.**

3.2 Predictive capability of the ML models

The objective of this section is to assess the validity of replacing f by a ML model $\tilde{f}_{\boldsymbol{\theta}}$ (with $\boldsymbol{\theta}$ being the ML hyperparameters). To do so, we aim to quantify the predictive capability of $\tilde{f}_{\boldsymbol{\theta}}$, i.e. whether $\tilde{f}_{\boldsymbol{\theta}}$ is capable of predicting *sl* with high accuracy given yet-unseen instances of the modelling assumptions (inputs). If this predictive capability is high, replacing f by $\tilde{f}_{\boldsymbol{\theta}}$ can be considered a valid assumption. The predictive capability of the ML model is commonly assessed using some global performance indicators calculated for a given test set $\boldsymbol{T}$. Ideally, the analysis can be done by defining an independent test set $\boldsymbol{T}$ in addition to the MME results. In the absence of such independent dataset, we preferably rely on a leave-one-out cross validation procedure (Hastie et al., 2009) that uses part of the available MME results to train the ML model $\tilde{f}_{\boldsymbol{\theta}}$, and a different part to test it. At a given time t, the procedure holds as follows.

- Step 1. Extract the ith MME result;
- Step 2. Train $\tilde{f}_{\boldsymbol{\theta}}$. using the other n−1 parts of the data, and the prediction error measured by $e^{(i)}(\mathrm{t}) = sl^{(i)}(\mathrm{t}) - \widehat{sl}^{(i)}(\mathrm{t})$ is calculated when predicting the ith part of the data;
- Step 3. The procedure is re-conducted for i= 1,2,...,n and performance indicators are calculated by combining the n estimates of the prediction error.

We use two performance indicators, namely a local one, that measures the local predictive capability related to the considered ith MME result, and a global one, that measures the predictive capability computed across all n MME results. The interest is twofold: the local indicator gives confidence in the local importance analysis for the considered ith case, and the global one gives confidence in the computation of the Shapley values, which require making predictions for inputs' configurations that are not necessarily present in the original MME dataset (see Sect. 3.3 and 3.4).

On the one hand, the local performance indicator is chosen to be the absolute error $AE^{(i)}(\mathrm{t}) = |e^{(i)}(\mathrm{t})|$. To be able to compare the results across time and across the experiments, its normalised version will also be used, i.e. the relative absolute error $RAE^{(i)}(\mathrm{t}) = \left|\frac{e^{(i)}(\mathrm{t})}{sl^{(i)}(\mathrm{t})}\right|$. On the other hand, the global performance indicator is chosen to be the mean absolute error $MAE(t) = \frac{1}{n}\sum_{i=1,\dots,n}|e^{(i)}(\mathrm{t})|$ (and by its normalised version, the mean relative absolute error $MRAE(\mathrm{t}) = \frac{1}{n}\sum_{i=1,\dots,n} RAE^{(i)}(\mathrm{t})$). For a given case i and at a particular time t, the ML model that minimises $MRAE + RAE^{(\mathrm{i})}$ is then retained for the local explanation analysis described in Sect. 3.3. This means that only the ML model that both performs the best globally (across the n MME results) and locally (for the considered ith MME result) is selected for the local explanation analysis.

Finally, it should be noted that no matter how much effort is put in increasing the ML predictive capability, a perfect match to the true model is rarely achievable in particular due to difficulties in approximating the mathematical relationship between the inputs and *sl* or due to the absence of input variables that are important with respect to the *sl* prediction error. Thus, a residual degree of prediction error may still remain. This has implications for the interpretation of low $|\mu_j^*(\mathrm{t})|$ values. In

theory, $|\mu_j^*(t)| = 0$ means that the jth input has no impact on the prediction at time t, i.e. it has negligible influence. In practice, the absence of influence can be concluded only up to a given threshold that is related to the residual prediction error. This means that low contribution values cannot be distinguished from the predictive error. In the following, we propose to use different performance indicators given the level of the sensitivity analysis (Step 4 described in Sect. 3.1) to assess the significance of the inputs with respect to the prediction error: for Level 1, we use $AE^{(i)}(t)$; for Level 2, we use $MAE(t)$; for Level 3, we analyse a variant of $RAE(t)$, namely $RAE_n(t) = \left|\frac{e(t)}{sl(t)-\mu_0(t)}\right|$.

3.3 Shapley additive explanation

We follow the approach developed by Lundberg and Lee (2017) who proposed to define $\mu_i^*(t)$ in Eq. 1 using the Shapley values (Shapley, 1953). The Shapley value is used in game theory to evaluate the "fair share" of a player in a cooperative game, i.e. it is used to fairly distribute the total gains to multiple players working cooperatively. It is a "fair" distribution in the sense that it is the only distribution satisfying some desirable properties (Efficiency, Symmetry, Linearity, 'Dummy player', see proofs by Shapley, 1953, see Aas et al., 2021: Appendix A for a comprehensive interpretation of these properties from a ML model perspective).

Formally, consider a cooperative game with k players and let $S \subseteq K = \{1, \dots, k\}$ be a subset of $|S|$ players. Let us define a real-valued function that maps a subset S to its corresponding value $\mathrm{val}: 2^S \to \mathbb{R}$ and measures the total expected sum of payoffs that the members of S can obtain by cooperation. The gain that the i-player gets is defined by the Shapley value with respect to val:

$$\mu_i(t) = \frac{1}{k}\sum_{S\subseteq K\setminus\{i\}} \binom{k-1}{|S|}^{-1} (\mathrm{val}(S \cup \{i\}) - \mathrm{val}(S)), \tag{2}$$

Equation 2 can be interpreted as a weighted mean over contribution function differences for all subsets S of players not containing player i. This approach can be translated for the ML-based *sl* prediction by viewing each model input (each type of modelling assumptions) as a player, and by defining the value function val as the expected output of the ML model conditional on $\mathbf{x}_S^*$ i.e. when we only know the values of the subset S of inputs (Lundberg and Lee, 2017), namely:

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

where $\bar{S}$ is the complement of S such that $\mathbf{x}_{\bar{S}}$ is the part of $\mathbf{x}$ not in $\mathbf{x}_S$, and $\mathrm{p}(\mathbf{x}_{\bar{S}}|\mathbf{x}_S = \mathbf{x}_S^*)$ is the conditional probability distribution of $\mathbf{x}_{\bar{S}}$ given $\mathbf{x}_S = \mathbf{x}_S^*$.

In this setting, the Shapley values can then be interpreted as the contribution of the considered input to the difference between the prediction $\tilde{f}(\mathbf{x}^*)$ and the base value μ_0. The latter can be defined as the value that would be predicted if we did not know any inputs (Lundberg and Lee, 2017), and is chosen as the expected prediction for *sl* without conditioning on any inputs, i.e. the unconditional expectation $\mu_0 = \mathrm{E}(f(\mathbf{x}))$. In this way, μ_i^* in Eq. 1 corresponds to the change in the expected model prediction when conditioning on that input and explains how to depart from $\mathrm{E}(f(\mathbf{x}))$. 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 for the ith 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 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).

In practice, the computation of the Shapley value may be demanding because Eq. (2) implies covering all subsets S (which grows exponentially with the number of factors denoted k, i.e. 2^k), and Eq. (3) requires solving integrals, which are of dimension 1 to $k-1$. For both reasons, the calculation is performed using a surrogate model (i.e. the ML model) in place of the true function f, because the design of computers is rarely complete (i.e. it rarely contains the results for the different configurations of the inputs that are needed for the calculation). To further alleviate the computational burden in this study, we rely on the kernel SHAP method of Lundberg and Lee (2017), which allows a computationally tractable approximation, and a simple method for estimating the value function in Eqs. 2-3. For this purpose, we use the R package '*shapr*' (Sellereite and Jullum, 2020) with accounts for inputs' dependence (see Sect. 3.4).

3.4 Accounting for inputs' dependencies

In the case considered in this study, there exists some dependence among the inputs. A commonly-encountered example is when the values for the minimum and maximum grid sizes are correlated. Additional examples are provided in Sect. 4.1. In this case, the interpretation of the SHAP decomposition provided by the kernel SHAP method might give wrong answers (Aas et al. 2021) because it relies on the independence assumption for calculating the conditional probability $p(\mathbf{x}_S|\mathbf{x}_S = \mathbf{x}_S^*)$ in Eq. (3). In our case, the dependence cannot be neglected (see Sect. 4.1 for the application to GrIS MME) and we rely on the improved kernel SHAP method proposed by Redelmeier et al. (2020) using conditional inference trees, denoted CTREE (Hothorn et al., 2006) to account for the dependence structure of input variables that are of mixed types (i.e. continuous, discrete, ordinal, and categorical) in the calculation of Eq. 3.

Conditional inference trees belong to the class of decision trees that use a two-stage recursive partitioning algorithm, namely (1) partition of the observations by univariate splits in a recursive way; (2) fit a constant model in each cell of the resulting partition (for regression problem). Different splitting procedures exists and we use here the one proposed by Hothorn et al. (2006), that uses a significance test to select input variables rather than selecting the variable that maximizes the information measure (such as the Gini coefficient, Breiman, 1984). In this approach, the stopping criterion is based on p-values of the significance test; for instance the p-value must be smaller than a given value (typically of 5%) in order to split the considered node. The advantage of CTREE is to avoid a selection bias towards covariates with many possible splits or missing values (see Hothorn et al., 2006 for further details).

To identify the dependence structure, we proceed as follows. We first consider the 1st input variable to be the response, and fit a CTREE model by viewing the remaining input variables as the predictor variables. If the resulting tree model includes one of the predictor variable, this means that there is some dependence with the considered response (i.e. the 1st variable in this example). Otherwise, the resulting tree model is empty. This approach is re-conducted by considering each of the input variables as the response in turn. As a result, the procedure identifies the non-empty tree model(s) that represent the dependence structure between some input variables.

4 Application

In this section, we apply the procedure described in Sect. 2 (schematically depicted in Fig. 3) to the MIROC5,RCP8.5-forced GrIS MME. We first analyse the dependence between the different modelling assumptions (Sect. 4.1). Then, we train and build ML models and select the best performing ones by following Steps 1-2 of the procedure (Sect. 4.2). On this basis, we apply the local attribution approach to measure the local importance and summarise the results to provide different levels (detailed in Sect. 3.1) of information on sensitivity (Steps 3-4, Sect. 4.3).

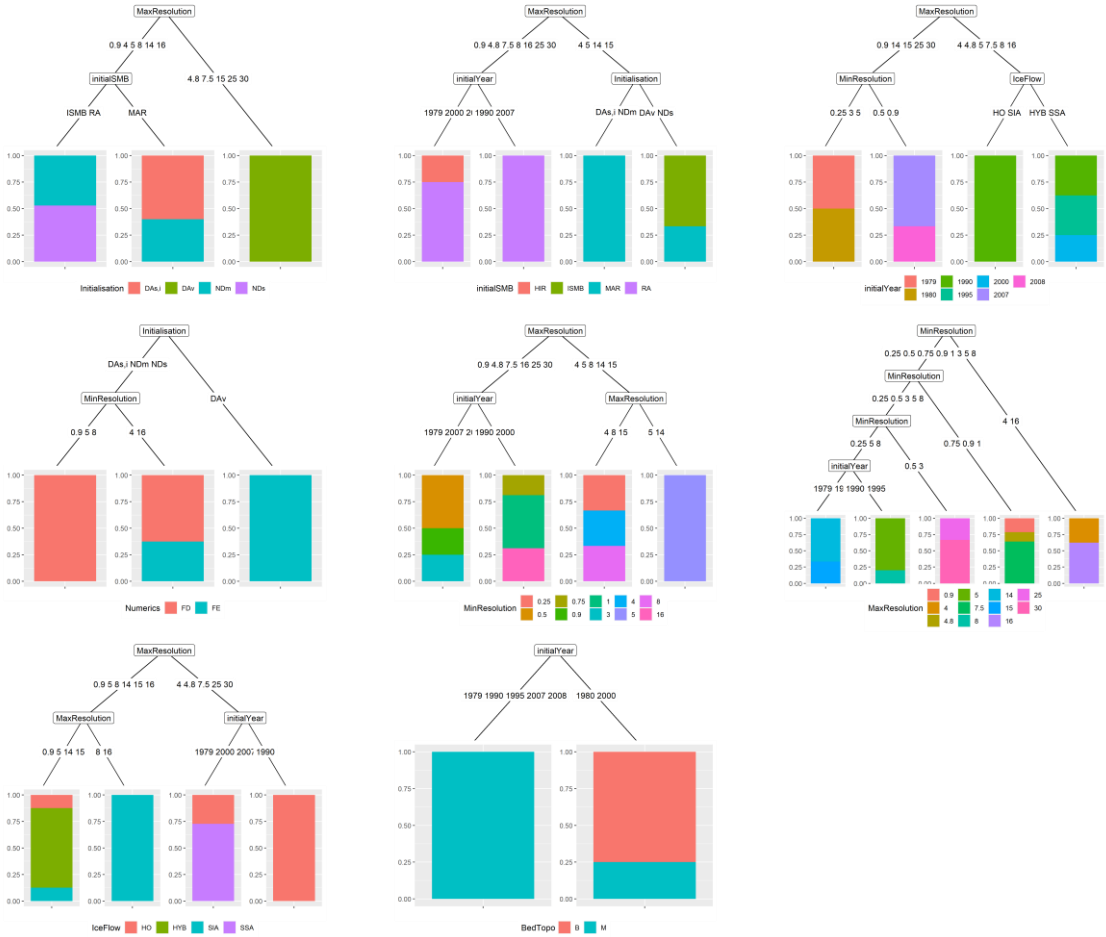

Figure 4: Tree models representing the dependence between the different modelling assumptions (indicated at the bottom of each tree). The bottom nodes (leaf nodes) provide the proportion of experiments given the modelling choices defined along the branches of the tree model. Each colour corresponds to a different category of the considered modelling assumption. For instance, the centre, left tree provides the relation between the choice in the numerical method with the type of initialisation and the minimum grid size. The blue (respectively red) colour is related to the finite element *FE* (respectively finite difference *FD*) category.

4.1 Inputs' dependencies

We first analyse the statistical dependence among the modelling assumptions (inputs) by applying the CTREE approach described in Sect 3.4 (using a split criterion threshold of 95% and Bonferroni-adjusted p-values). Figure 4 shows the resulting tree models for the different modelling assumptions. We show here that all inputs are statistically dependent at the exception of κ for which the tree model is empty, which indicates the absence of (significant) dependence between this parameter and the other modelling assumptions. The different tree models should be read by following the example of the centre, leftmost tree in Fig. 4. This tree provides the relation between the choice in the numerical method with the type of initialisation and the minimum grid size. The bottom nodes (leaf nodes) provide the proportion of experiments given the

combination of modelling choices defined along the branches of the tree model. The blue (respectively red) colour is related to the finite element *FE* (respectively finite difference *FD*) category. This tree model indicates for example that all models with initialisation of type *DAv* have a numerical method of type *FE* (rightmost branch) and all models with initialisation different of *DAv* and a minimum resolution of 0.9, 5 or 8km have a numerical method of type *FD* (leftmost branch).

4.2 Predictive capability of the ML models

Using the results of the MIROC5,RCP8.5-forced GrIS MME, we train a series of ML models to predict *sl* across time. The following ML model with corresponding hyperparameters (see Appendix B for details) are considered:

- Nine RF regression models with hyperparameters *ns*=5 or 10, m_{try}= 1, 3, 6 or 9 and n_{tree}=2,000;
- Thirty XGB models with hyperparameters maximum depth = 2, 3, 6, or 9, learning rate = 0.025, 0.1 or 0.25 and maximum number of boosting iterations = 250 or 450;
- One LIN model.

To assess the predictive capability of the considered ML models at each time instant, we apply a leave-one-out cross validation approach by following the procedure of Sect. 3.2. Figure 5a depicts the time evolution of the performance indicator *MRAE* for all considered ML models. Depending on the type of ML model (and corresponding parametrisation), the global performance can reach satisfactory levels below 10% in particular for some XGB models.

As explained in Sect. 3.2, satisfying the global performance criterion does not necessarily ensure that the ML model gives an accurate approximation of all *sl* predictions. For some cases, the discrepancies can be too large to properly analyse the local explanations. This is illustrated with Fig. 5b that shows the comparison between the true *sl* value and the corresponding ML-based prediction for 2100. For instance, we note that the predictions for the largest *sl* value largely departs from the 1:1 line except for the LIN model (outlined in black in Fig. 5b). This is also the case for the lowest *sl* values for which a given parametrisation of the XGB model performs the best (outlined in red in Fig. 5b). Thus, to further increase our confidence in replacing the 'true' numerical model by the ML model, we apply the filtering approach (described in Sect. 3.2) based on the joint minimisation of the global and of the local performance indicators. The retained predictions are outlined in blue in Fig. 5b.

In total, LIN, XGB and RF models are retained respectively 3.4%, 24.6% and 72% of the total number of experiments (in average over time). After applying this procedure, the *MRAE* criterion (shown in blue in Fig. 5a) reaches values below 10% in average over time (with a maximum value not larger than 15% for year 2040). Note that the *MRAE* curve after this selection is not necessarily the lowest one, because the selection procedure not only implies minimising $MRAE$ but also the local performance $RAE^{(i)}$ (see Sect. 3.2).

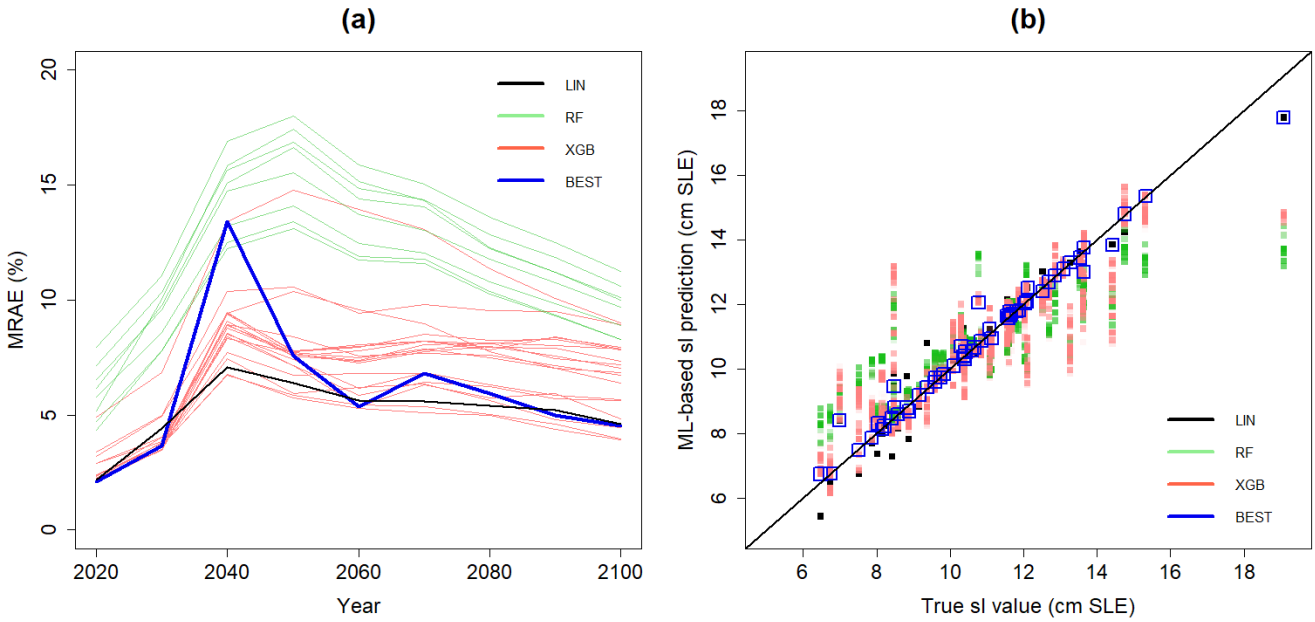

Figure 5: (a) Time evolution of the performance criterion *MRAE* (expressed in %) computed using a leave-one-out cross validation procedure that assesses the predictive capability of all considered ML models with different parametrisations (RF models in green, XGB in red and LIN in black). The blue-coloured lines are related to the performance criterion after selecting the best performing ML model with respect to the joint minimisation of the global and of the local performance indicator described in Sect. 3.2; (b) Comparison between the true and the ML-based predicted *sl* value for 2100 by considering all ML models. The blue colour outlines the retained results after selecting the best performing ML model.

4.3 From local to global explanations

In this section, we first compute the measures of local importance for each experiment in the MIROC5,RCP8.5-forced GrIS MME for a given prediction time (here 2100); such type of diagnostic (*Level 1* of the procedure) helps to understand and quantify the impact of particular assumptions made by the modellers (Sect. 4.3.1). Then, we analyse in Sect. 4.3.2 how the influence of each modelling assumption evolves as a function of the considered input value (*Level 2* of the procedure). This analysis allows us to deepen our understanding of the model structure for a given prediction time. Finally, Sect 4.3.3 summarises all results over time (*Level 3* of the procedure) to provide a global insight (i.e. across all MME members) in the sensitivity of *sl* to the modelling assumptions.

4.3.1 Level 1. Local explanations at a given prediction time

We first illustrate the application of SHAP to a selected set of ML-based *sl* predictions for 2100. Figure 6 provides the SHAP-based decomposition of the ML-based prediction (blue horizontal bar) into the positive (green bar) or negative (red bar) contribution (μ value defined in Eqs. 2-3) of each input using the 2100 ensemble mean of μ_0=10.8cm as base value. The inputs' setting are indicated in the vertical axis for each of the considered Cases (a) - (f). The grey colour indicates that the contribution cannot be distinguished from the predictive error, because its absolute value is below the absolute error.

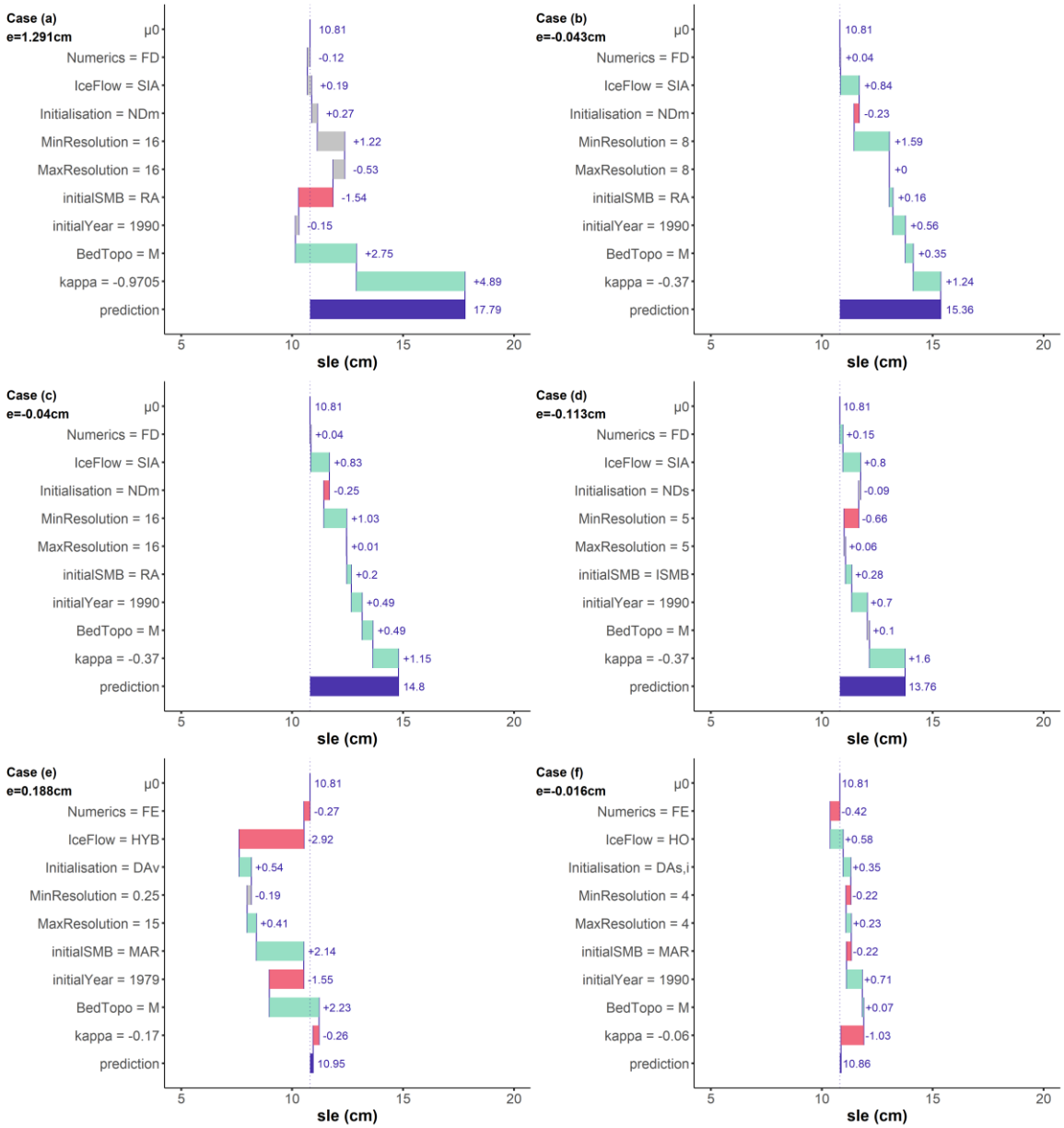

Figure 6: Diagnostic of particular ML-based *sl* predictions using SHAP for year 2100 considering six different settings of the modelling choices (indicated in the vertical axis). The horizontal blue bar corresponds to the ML-based *sl* prediction (the difference with the true value is indicated by the 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.

The analysis of Figure 6 illustrates how the SHAP-based approach can be used to diagnose the MME results:

- Case (a) corresponds to the largest *sl* value (of 19.08cm) that is predicted by the ML model at 17.79cm (with a prediction error $e \approx 1.30$cm). Fig 6(a) confirms the physically expected result regarding κ influence: the largest *sl* is mainly attributable to κ whose absolute value is the largest, i.e. 0.9705 km.(m^3.s^{-1})$^{-0.4}$ °C. This choice pushes the *sl* value higher than the base value by μ=+4.89cm, i.e. by ≈45% of μ_0. In this case, the two other largest contributors to *sl* (with an influence of respectively +2.75cm, and -1.54cm) are related to using the *M* dataset for bed topography and initial *SMB* of type *RA*. The other modelling choices all have absolute contributions below |e|, which indicates that their contributions are not significant in comparison to the prediction error level (outlined in grey in Fig. 6a);
- Case (b) (Fig. 6(b)) corresponds to the second largest *sl* value (of 15.32cm) that is predicted by the ML model at 15.36cm (with a prediction error e≈0.04cm). All modelling choices are similar to Case (a) except κ here set up to a lower absolute value of 0.37 km.(m^3.s^{-1})$^{-0.4}$ °C and the minimum grid size set up to a lower value of 8km. Contrary to Case (a), the influence of κ drops here to low-to-moderate value (+1.24cm), and it is the choice in the minimum grid size that contributes the largest to *sl* (μ=+1.59cm). We note that all contributions can be considered with confidence because their absolute values are all above the absolute prediction error;
- Case (c) presents the same setting that case (b) except for a larger minimum grid size (here of 16km). This results in a lower influence of the minimum grid size (μ drops to +1.03cm), but the contributions of all modelling assumptions remain, to some extent, similar to Case (b);
- Case (d) corresponds to a *sl* value close to the one in Case (c) and illustrates that, despite the differences with Case (c) (i.e. initial SMB, initialisation type and minimum resolution), the contribution of largest contributors to *sl*, i.e. ice flow's type, initial year and κ, remains of the same order of magnitude between both cases;
- The comparison between Cases (b) to (d) also points out that, for relatively close predicted values, the modelling choices contribute equivalently to the prediction despite some minor differences in the setting of the modelling assumptions;
- Cases (e) and (f) illustrate however that, when the dissimilarity in the settings is larger, the modelling choices contribute differently to the prediction although the predicted values are very close (here close to the ensemble mean of 10.8cm). In Case (f), all modelling assumptions contribute equivalently to *sl*, whereas it is mainly ice flow's type and the type of dataset for bed topography in Case (f).

Such type of diagnostic can be performed for any MME results (they are all provided by Rohmer (2022) for year 2100) to inform the modellers on the most and least impactful modelling choices for any *sl* prediction; such information being helpful to explain why a given instance of modelling choice lead to a given *sl* value.

4.3.2 Level 2. Model structure at a given prediction time

We explore in Fig. 7 and in Fig. 8 how the magnitude of the modelling assumption's contribution to *sl*, as well as the direction, change depending on the value of the considered input by applying the SHAP dependence plot proposed by

Lundberg et al. (2020). To judge the significance of the contribution, we compare the results to the range defined by +/- *MAE*=0.18cm (calculated from the leave-one-out cross-validation procedure, see Sect. 4.2): contributions falling within this range (outlined by the red dashed horizontal lines in Fig. 7) indicates that they cannot be distinguished from the predictive error.

We first analyse the continuous variables. Fig. 7a confirms the large influence of κ (of several cm) for large absolute values of κ. We also note that setting this parameter to -0.17 km.(m^3.s^{-1})$^{-0.4}$ °C leads to quasi-negligible influence, because μ falls within the range of *MAE*. A clear trend can be noticed: κ influence decreases with increasing value in a quasi-linear manner (with slope of ~-8 cm per unit of retreat parameter). We also note that for setting κ above -0.17 km.(m^3.s^{-1})$^{-0.4}$ °C even impacts negatively the *sl* prediction, which means that this modelling assumption pushes the prediction lower than the mean value for 2100. Finally, Fig. 7a provides indication of where to perform additional numerical experiments to confirm influence, namely over the range -0.97 to -0.37 km.(m^3.s^{-1})$^{-0.4}$ °C (where the results are scarce).

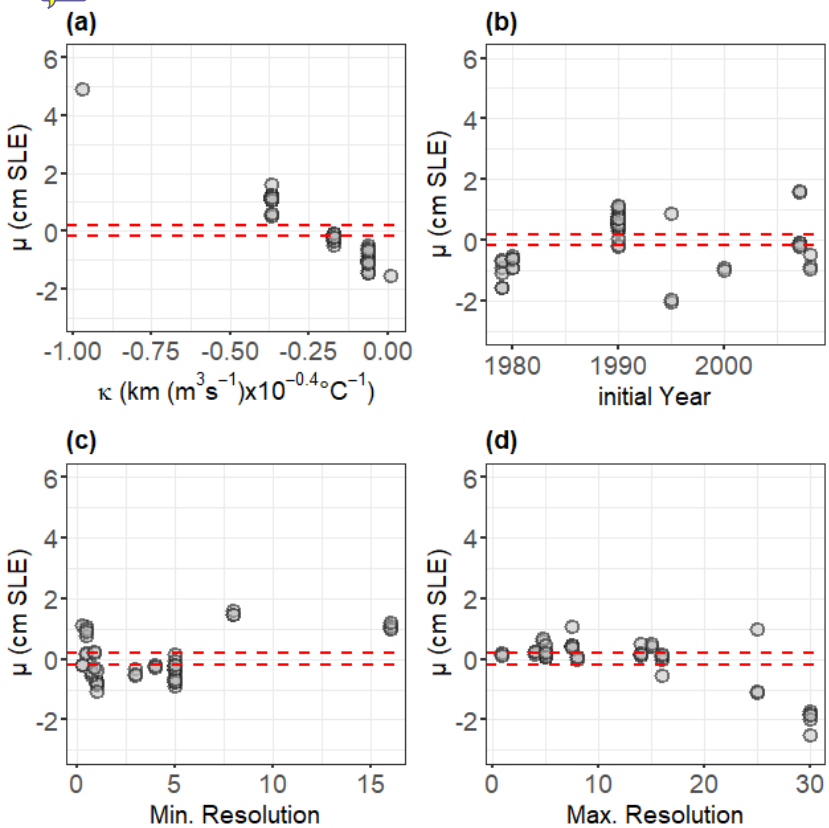

Figure 7: Application of SHAP additive explanation to all members of the MIROC5,RCP8.5-forced GrIS MME for year 2100. Each panel provides μ (y-axis) as a function of the value of the minimum and maximum grid resolution (a,b), of the initial year (c), and of the retreat parameter κ (d). The horizontal dashed red lines indicate the limits defined by +/- *MAE* calculated from the leave-one-out cross-validation procedure: contributions falling within this range indicates that they cannot be distinguished from the predictive error. ~~The blue line indicates the smooth regression fitted to the data (using locally estimated scatterplot smoothing method), and the grey envelope is the 95% confidence interval.~~

Though a trend in the (initial year - μ) mathematical relationship is not straightforward to detect, Fig. 7b shows that the influence can be considered significant with respect to the predictive error *MAE* for some particular cases; $|\mu|$ reach low-to moderate values not larger than 2cm.

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 resolution respectively (with a maximum value up to ≈1.1cm for both grid sizes). The influence can even be considered non-significant with 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 (i.e. if the minimum and maximum grid resolution is outside the identified zone), 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 zone.

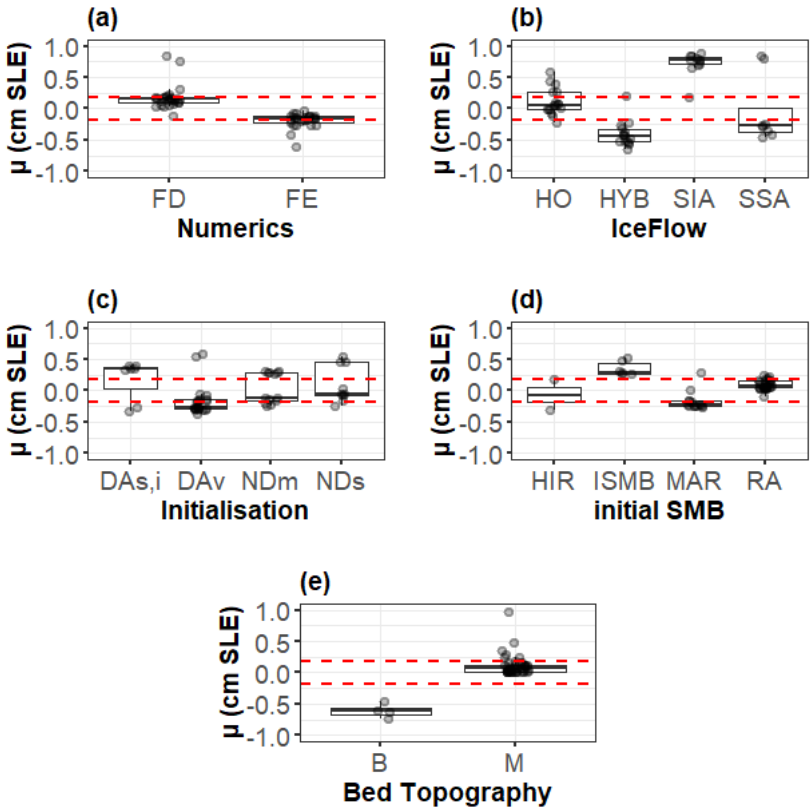

Figure 8: Application of SHAP additive explanation to all members of MIROC5,RCP8.5-forced GrIS MME for year 2100. Each panel provides the boxplots of μ values given the modelling choice for the numerical method (a), the ice flow (b), the initialisation

(c), the initial SMB (d) and type of bed topography dataset (e). Each dot corresponds to a given MME member. The horizontal dashed red lines indicate the limits defined by *MAE* calculated from the cross-validation procedure.

Focusing on the categorical input variables, Figure 8 further indicates that the most impactful modelling assumption for *sl* is the ice flow's choice, either of *SIA* or *HYB* type with positive or negative contribution, and the *B* dataset for bed topography: the corresponding boxplots in Fig. 8b and in Fig. 8e are well outside the +/- *MAE* range. Finally, Fig. 8 also points out some modelling choices with contributions that are hardly distinguishable from the prediction error; namely any type of numerical method, *FD* or *FE* (Fig. 8a), *NDm* and *NDs* for initialisation (Fig. 8c), *HIR* or *RA* for initial *SMB* (Fig. 8d), and *M* dataset for bed topography though some specific cases present low-to-moderate values (see grey dots outside the box in Fig. 8e).

4.3.3 Level 3. Global explanations over time

The analysis of Sect. 4.3.2 is performed for all members of the MIROC5,RCP8.5-forced GrIS MME for any prediction time. As indicated in Sect. 3.1, to be able to compare the influence between the different predictions across time, we analyse in Fig. 9 the statistics of the absolute value of $\mu_n(\mathrm{t}) = \mu(\mathrm{t})/(sl(\mathrm{t}) - \mu_0(\mathrm{t}))$. To judge on the negligible level of the influence with respect to the ML prediction error, we analyse the quartiles of $RAE_n(\mathrm{t}) = \left|\frac{e(\mathrm{t})}{sl(\mathrm{t})-\mu_0(\mathrm{t})}\right|$ calculated at each time instant for all members of MIROC5,RCP8.5-forced GrIS MME. 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.

Considering initial conditions, Fig. 9a,c shows that it is the initialisation type that has the largest impact in the medium term (before 2050/2060), and after this date, it is the choice in the initial year that impacts the most. Conversely, in the long term (after 2050/2060), the influence of the initialisation type reduces up to negligible level (compared to the prediction error). Fig. 9b shows that the influence of the initial *SMB* is low (even negligible) regardless of the considered prediction time at the exception of some particular cases outlined by black dots lying outside the boundaries of the whiskers (illustrated in Fig. 6a,e).

Considering numerical implementation, the choice in the numerical method has here small (even negligible) contributions to *sl* values (Fig. 9d) especially in the medium / long term (after 2050). We note also that the moderate influence of the minimum and maximum grid size remains quasi-constant over time (Fig. 9e,f), hence suggesting that the grid size's influence is time-invariant, i.e. all modelled processes are affected by the spatial resolution in a similar way, independently of the prediction time.

Finally, considering ice-sheet processes and environmental forcing, an important influence of κ is shown only after 2030/2040 (Fig. 6h) with a quasi-constant value after this date. An increasing influence over time is also identified for the ice flow's type; though the temporal trend is only clear up to year 2070. We also show that the type of bed topography dataset

has only low (even negligible) influence compared to the prediction error, at the exception of some particular cases (illustrated in Fig. 6a,e) and outlined by black dots lying outside the boundaries of the whiskers.

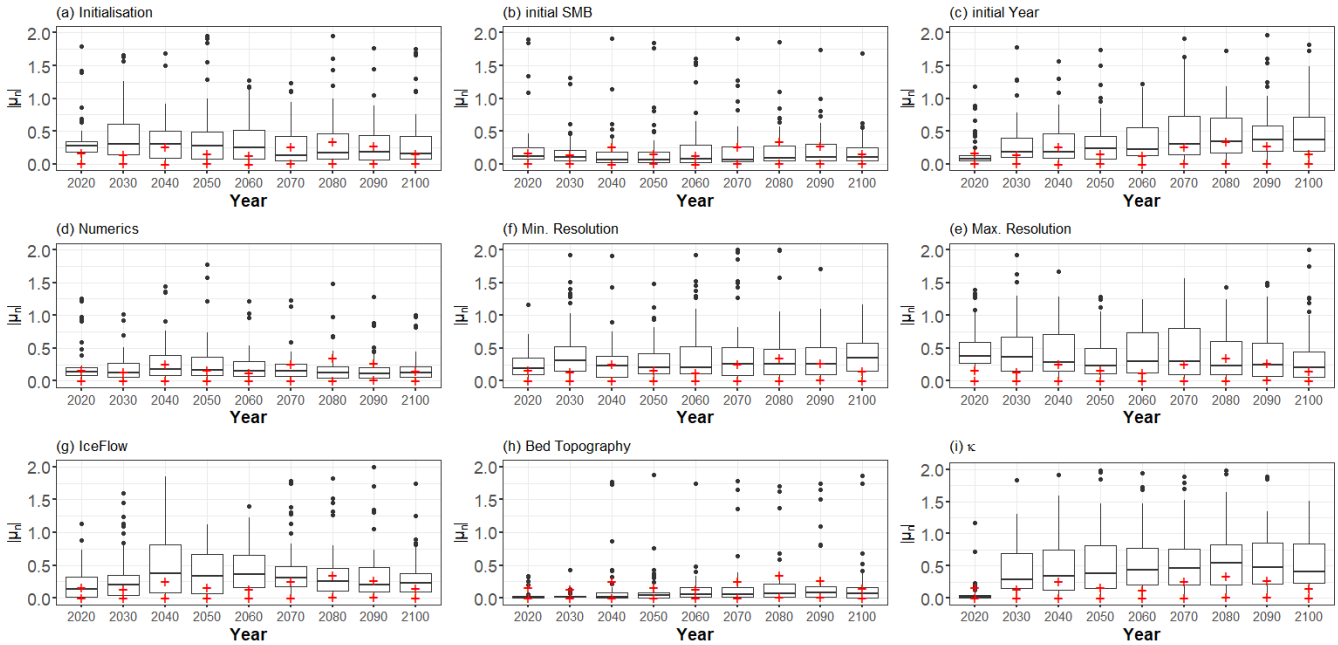

Figure 9: Statistics of $|\mu_n|$ summarised by a boxplot at each time instant for all members of the MIROC5,RCP8.5-forced GrIS MME. The lower and upper red cross is respectively the 1st and 3rd quartile of the cross-validation error RAE_n. If the boxplot does not overlap with the region defined by the interval between the red crosses, this indicates that the influence measured by $|\mu_n|$ can be considered significant with respect to the ML prediction error. For readability, the upper bound of the y-axis has been set up to 2.

5 Discussion

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 close *sl* values.

Second, aggregating all diagnostic results (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., 2016; 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 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 tagged as key contributors; though κ importance has relatively well understood physical basis, additional analysis should be carried out for the initial year;
- 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 parts of the parameter space in the view to confirm the identified trends (Figs. 7 and 8); 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.

These results were obtained by overcoming two major difficulties. The first one is related to the incomplete and unbalanced design of the numerical experiments (Sect. 4.1). here, applying more commonly-used statistical methods, namely the linear regression model or the ANOVA-based approach, would hardly be feasible. On the one hand, Sect. 4.2 clearly shows that the mathematical relationship between *sl* and the inputs is not necessarily linear, and more advanced regression techniques need to be used (like RF or XGB models). On the other hand, the considered design of experiments is incomplete and unbalanced (as shown in Sect. 2), which complicates the application of ANOVA. Ideally a full factorial design should be used to properly apply ANOVA: in our case, the design should then contain 3,200 experiments, i.e. far larger than the available experiments. Some solutions have been proposed in the literature (see e.g., Evin et al. (2019) and references therein), and an avenue for future work could focus on the comparison of ANOVA with our approach. The second difficulty is related to the

presence of statistical dependencies (as outlined in Sect. 4.1), which makes the interpretation of the individual effects less straightforward (a problem related to multicollinearity in the statistical community, e.g., Shrestha, 2020) and might even lead to wrong conclusions regarding uncertainty partitioning (see discussion by Do and Razavi (2020)). 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 inputs' dependence, which could serve as a strong basis for defining clear recommendations on how to treat it in the context of MME.

However, it should be underlined that the high performance of our approach is strongly dependent on two key prerequisites. First, the high predictive capability of the ML model should be carefully checked and confirmed as done in the GrIS case (Sect. 4.2). For this purpose, several aspects need further investigation in future work: (1) instead of selecting one single ML model, a combination of models could be proposed following e.g. the 'super-learner' method of van der Laan et al. (2007) or the model class reliance approach of Fisher et al. (2019); (2) finding the optimal hyperparameters' setting could benefit from more advanced search algorithms for optimization (Probst et al., 2019).

The second prerequisite is the careful selection of which input variables to include in the analysis. 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 the input variables. One advantage of SHAP approach is that 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. However, one disadvantage of this type of robustness analysis is the much higher computational cost (at least 9 times), which makes it difficult to implement for the whole MME results. This requires further research work related to the active research area of 'sensitivity of the sensitivity analysis' (e.g., Razavi et al. 2021).

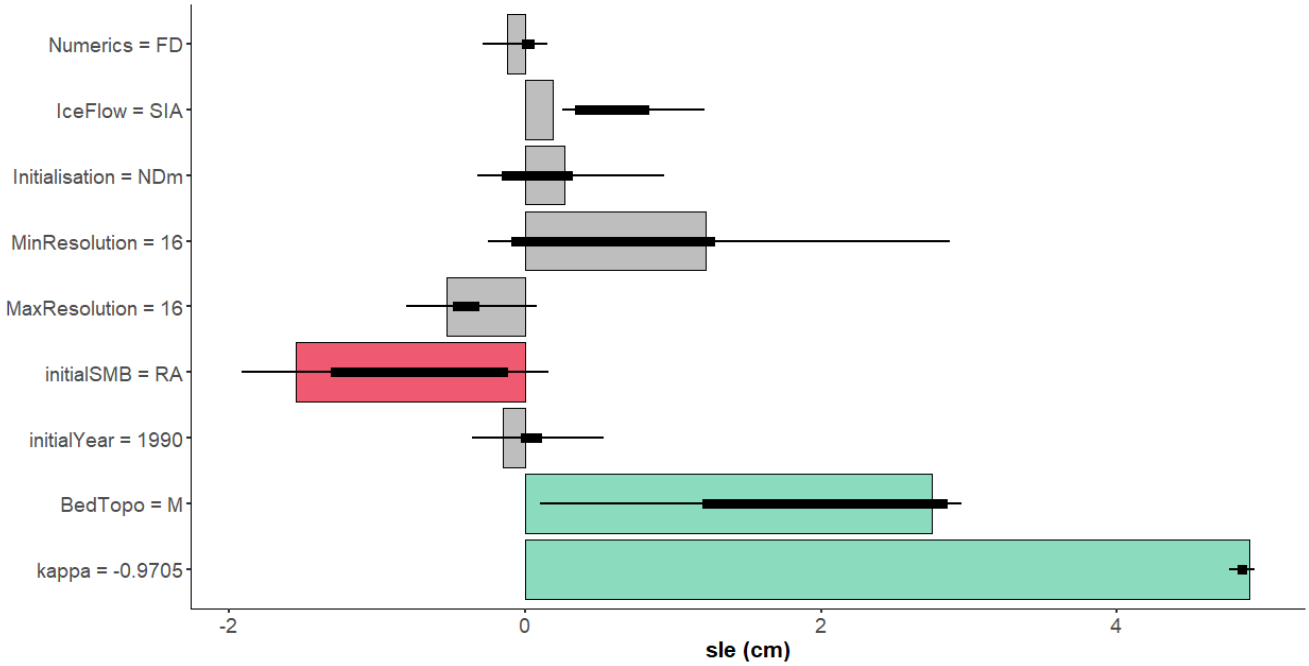

Figure 10: Robustness analysis of the local importance analysis for the largest simulated *sl* value in 2100 (Case (a) in Figure 6). The horizontal coloured bars correspond to the quantified contributions by including all input variables (results of Fig. 6a). The endpoints of the thick and thin horizontal 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.

6 Concluding remarks and further work

In this study, we described the use of the machine-learning-based SHapley Additive exPlanation (SHAP) approach to quantify the importance of modelling assumptions in sea-level projections produced in a MME study. The proposed approach was applied to a subset of the GrIS ensemble that is characterised by a limited number of experiments (50-100), an unbalanced design, and the presence of dependence between the inputs. Our results have shown the added value of the proposed approach to inform on the influence of the modelling assumptions at multiple levels: (*Level 1*) locally for particular instances of the modelling assumptions, (*Level 2*) on the model structure at a given prediction time, and (*Level 3*) globally over time. These results are intended for different potential users, namely the ice-sheet modelling community (individual modellers, or modelling groups in charge of the design of experiments), but also adaptation practitioners, who take decisions based on sea-level projections that rely on models such as those modelling the Greenland ice mass losses. Trust in these projections, and therefore accelerated coastal adaptation, can be enabled by the analyses described in this study allowing to better interpret the uncertainty range in projections. This study illustrates that performing such diagnoses rigorously requires advanced mathematical techniques.

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 AOGCM 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.

Second, it is the question of the global effects of the modelling assumptions that deserves particular intensified investigation. In addition to methodological work exploring advanced procedures such as SAGE (Shapley Additive Global importance, Covert et al., 2020) or variance-based approach used in the Uncertainty Quantification community (e.g. Iooss and Prieur, 2019), the key will be the developments of robust protocols to design balanced and complete numerical experiments. This partially resolved problem (see e.g. discussion by Aschwanden et al., 2021) could benefit from an increased inter-disciplinary cooperation as well.

Author contributions

JR designed the concept, set up the methods and undertook the statistical analyses. JR and HG defined the protocol of experiments. JR, RT, GLC, HG, GD analysed and interpreted the results. JR wrote the manuscript draft. JR, RT, GLC, HG, GD reviewed and edited the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Code/Data availability

The sea level dataset is the one compiled by Edwards et al. (2021)[1] (last access: 2 June 2022) from the original data of Goelzer et al., (2020) by selecting the experiments with column name *ice_source*='GrIS', *region*='ALL', *GCM*='MIROC5', *scenario*='RCP8.5', and with prior exclusion of experiments with NaN value of the retreat parameter. R scripts to reproduce the results of Sect. 4.3 corresponding to the three levels of analysis are provided by Rohmer (2022), and in particular, the different diagnostics for all MIRO5,RCP8.5-forced GrIS MME results (similar to Fig. 6). SHAP approach was implemented using R package *shapr* (Sellereite and Jullum, 2020). CTREE approach was implemented using R package *partykit* (Hothorn and Zeileis, 2005). ML model fitting was performed using R packages *ranger* (Wright and Ziegler, 2017) and *xgboost* (Chen et al., 2022).

Acknowledgements

For the ISMIP6 results used in this study, we thank the Climate and Cryosphere (CliC) effort, which provided support for ISMIP6 through sponsoring of workshops, hosting the ISMIP6 website and wiki, and promoted ISMIP6. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP5. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the CMIP data and providing access, the University at Buffalo for ISMIP6 data distribution and upload, and the multiple funding agencies who support CMIP5 and ESGF. We thank the ISMIP6 steering committee, the ISMIP6 model selection group and ISMIP6 dataset preparation group for their continuous engagement in defining ISMIP6. This is ISMIP6 contribution No XXX 'To be completed after acceptance'. This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 869304, PROTECT. HG has received funding from the Research Council of Norway under projects 270061, 295046 and

[1] https://raw.githubusercontent.com/tamsinedwards/emulandice/master/inst/extdata/20201106_SLE_SIMULATIONS.csv

324639. High-performance computing and storage resources were provided by Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway through projects NN8006K, NN8085K, NS8006K, NS8085K, NS9560K, NS9252K and NS5011K.

References

Aas, K., Jullum, M., and Løland, A.: Explaining individual predictions when features are dependent: More accurate approximations to Shapley values, Artificial Intelligence, 298, 103502, 2021.

Achen, C. H.: Intepreting and Using Regression, Sage Publications, Thousand Oaks, 1982.

Betancourt, C., Stomberg, T. T., Edrich, A. K., Patnala, A., Schultz, M. G., Roscher, R., et al.: Global, high-resolution mapping of tropospheric ozone–explainable machine learning and impact of uncertainties, Geoscientific Model Development Discussions, 1-36, 2022.

Aschwanden, A., Bartholomaus, T. C., Brinkerhoff, D. J., and Truffer, M.: Brief communication: A roadmap towards credible projections of ice sheet contribution to sea level, The Cryosphere, 15(12), 5705-5715, 2021.

Bamber, J. L., Griggs, J. A., Hurkmans, R. T. W. L., Dowdeswell, J. A., Gogineni, S. P., Howat, I., Mouginot, J., Paden, J., Palmer, S., Rignot, E., and Steinhage, D.: A new bed el- evation dataset for Greenland, The Cryosphere, 7, 499-510, https://doi.org/10.5194/tc-7-499-2013, 2013.

Barthel, A., Agosta, C., Little, C. M., Hattermann, T., Jourdain, N. C., Goelzer, H., Nowicki, S., Seroussi, H., Straneo, F., and Bracegirdle, T. J.: CMIP5 model selection for ISMIP6 ice sheet model forcing: Greenland and Antarctica, Cryosphere, 14, 855-879, https://doi.org/10.5194/tc-14-855-2020, 2020.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J.: Classification and regression trees, Wadsworth, California, 1984.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, 2001.

Bussmann, N., Giudici, P., Marinelli, D., and Papenbrock, J.: Explainable machine learning in credit risk management, Computational Economics, 57(1), 203-216, 2021.

Chen, T., and Guestrin, C.: Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794, 2016.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al.: Xgboost: extreme gradient boosting. R package version 1.6.0.1, available at: https://cran.r-project.org/web/packages/xgboost/index.html (last access 2 June 2022), 2022.

Covert, I., Lundberg, S. M., and Lee, S. I.: Understanding global feature contributions with additive importance measures, Advances in Neural Information Processing Systems, 33, 17212-17223, 2020.

Do, N. C., and Razavi, S.: Correlation effects? A major but often neglected component in sensitivity and uncertainty analysis, Water Resources Research, 56(3), e2019WR025436, 2020.

Edwards, T. L., Nowicki, S., Marzeion, B., Hock, R., Goelzer, H., Seroussi, H., et al.: Projected land ice contributions to twenty-first-century sea level rise, Nature, 593(7857), 74-82, 2021.

Evin, G., Hingray, B., Blanchet, J., Eckert, N., Morin, S., and Verfaillie, D.: Partitioning uncertainty components of an incomplete ensemble of climate projections using data augmentation, Journal of Climate, 32(8), 2423-2440, 2019.

Fisher, A., Rudin, C., and Dominici, F.: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously, J. Mach. Learn. Res., 20(177), 1-81, 2019.

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

Friedman, J.: Greedy function approximation: a gradient boosting machine, Annals of Statistics, 29(5):1189–1232, 2001.

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

Goelzer, H., Nowicki, S., Edwards, T., Beckley, M., Abe-Ouchi, A., Aschwanden, A., Calov, R., Gagliardini, O., Gillet-Chaulet, F., Golledge, N. R., Gregory, J., Greve, R., Humbert, A., Huybrechts, P., Kennedy, J. H., Larour, E., Lipscomb, W.'H., Le clec'h, S., Lee, V., Morlighem, M., Pattyn, F., Payne, A. J., Rodehacke, C., Rückamp, M., Saito, F., Schlegel, N., Seroussi, H., Shepherd, A., Sun, S., van de Wal, R., and Ziemen, F. A.: Design and results of the ice sheet model initialisation experiments initMIP-Greenland: an ISMIP6 intercomparison, The Cryosphere, 12, 1433-1460, https://doi.org/10.5194/tc-12-1433-2018, 2018.

Goelzer, H., Nowicki, S., Payne, A., Larour, E., Seroussi, H., Lipscomb, W. H., et al.: The future sea-level contribution of the Greenland ice sheet: a multi-model ensemble study of ISMIP6, The Cryosphere, 14(9), 3071-3096, 2020.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer: Berlin/Heidelberg, Germany, 2009.

Hawkins, E., Sutton R.: The potential to narrow uncertainty in regional climate predictions. Bulletin of the American Meteorological Society, 90(8), 1095–1107, 2009.

Hothorn, T., and Zeileis, A.: partykit: A modular toolkit for recursive partytioning in R, The Journal of Machine Learning Research, 16(1), 3905-3909, 2015.

Hothorn, T., Hornik, K., and Zeileis, A.: Unbiased Recursive Partitioning: A Conditional Inference Framework, Journal of Computational and Graphical Statistics, 15 (3), 651–74, 2006.

Iooss, B., and Prieur, C.: Shapley effects for sensitivity analysis with correlated inputs: comparisons with Sobol' indices, numerical estimation and applications, International Journal for Uncertainty Quantification, 9(5), 2019.

Jack, C. D., Jones, R., Burgin, L., and Daron, J.: Climate risk narratives: An iterative reflective process for co-producing and integrating climate knowledge, Climate Risk Management, 29, 100239, 2020.

Jothi, N., and Husain, W.: Predicting generalized anxiety disorder among women using Shapley value, Journal of infection and public health, 14(1), 103-108, 2021.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J.: Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning, in: Proceedings of the 2020 CHI conference on human factors in computing systems, 1-14, 2020.

Kopp, R. E., Gilmore, E. A., Little, C. M., Lorenzo-Trueba, J., Ramenzoni, V. C., and Sweet, W. V.: Usable science for managing the risks of sea-level rise, Earth's Future, 7(12), 1235-1269, 2019.

Lundberg, S. M., and Lee, S. I.: A unified approach to interpreting model predictions, in: Proceedings of the 31st international conference on neural information processing systems, 4768-4777, 2017.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al.: From local explanations to global understanding with explainable AI for trees, Nature machine intelligence, 2(1), 56-67, 2020.

Molnar, C., Casalicchio, G., and Bischl, B.: Interpretable machine learning–a brief history, state-of-the-art and challenges, in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Cham, 417-431, 2020.

Molnar, C.: Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.), Available at: christophm.github.io/interpretable-ml-book/ (last access 2 June 2022), 2022.

Morlighem, M., Williams, C., Rignot, E., An, L., Bamber, J., Chauche, N., et al.: BedMachine v3: Complete bed topography and ocean bathymetry mapping of Greenland from multi-beam radar sounding combined with mass conservation. Geophys. Res. Lett., 44, 2017.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B.: Definitions, methods, and applications in interpretable machine learning, Proceedings of the National Academy of Sciences, 116(44), 22071-22080, 2019.

Murphy, J. M., Sexton, D. M., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., and Stainforth, D. A.: Quantification of modelling uncertainties in a large ensemble of climate change simulations, Nature, 430(7001), 768-772, 2004.

Northrop, P. J., and Chandler, R. E.: Quantifying sources of uncertainty in projections of future climate, Journal of Climate, 27(23), 8793-8808, 2014.

Nowicki, S. M. J., Payne, A., Larour, E., Seroussi, H., Goelzer, H., Lipscomb, W., Gregory, J., Abe-Ouchi, A., and Shepherd, A.: Ice Sheet Model Intercomparison Project (ISMIP6) contribution to CMIP6, Geosci. Model Dev., 9, 4521-4545, https://doi.org/10.5194/gmd-9-4521-2016, 2016.

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.

Padarian, J., McBratney, A. B., and Minasny, B.: Game theory interpretation of digital soil mapping convolutional neural networks, Soil, 6(2), 389-397, 2020.

Probst, P., Wright, M. N., and Boulesteix, A. L.: Hyperparameters and tuning strategies for random forest, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301, 2019.

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.

Redelmeier, A., Jullum, M., and Aas, K.: Explaining predictive models with mixed features using Shapley values and conditional inference trees, in International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, Cham, 117-137, 2020.

Rohmer, J.: Local explanation SHAP approach applied to MIROC5,RCP8.5-forced multi-model ensemble study of GrIS future sea-level contributions [Data set], https://github.com/rohmerj/Interpretation-of-sea-level-projections, ('To be transformed to Zenodo'), 2022.

Shrestha, N.: Detecting multicollinearity in regression analysis, American Journal of Applied Mathematics and Statistics, 8(2), 39-42, 2020.

Sellereite, N., and Jullum, M.: shapr: An R-package for explaining machine learning models with dependence-aware Shapley values, Journal of Open Source Software, 5(46), 2027, 2020.

Seroussi, H., Nowicki, S., Payne, A. J., Goelzer, H., Lipscomb, W. H., Abe-Ouchi, A., et al.: ISMIP6 Antarctica: a multi-model ensemble of the Antarctic ice sheet evolution over the 21st century, The Cryosphere, 14(9), 3033-3070, 2020.

Shapley, L. S.: A value for n-person games, in: H. Kuhn, A. W. Tucker (Eds.), Contributions to the Theory of Games, Volume II, Annals of Mathematics Studies, Princeton University Press, Princeton, NJ, Ch. 17, 307-317, 1953.

Slater, D. A., Straneo, F., Felikson, D., Little, C. M., Goelzer, H., Fettweis, X., and Holte, J.: Estimating Greenland tidewater glacier retreat driven by submarine melting, The Cryosphere, 13, 2489–2509, 2019.

Slater, D. A., Felikson, D., Straneo, F., Goelzer, H., Little, C. M., Morlighem, M., Fettweis, X., and Nowicki, S.: Twenty-first century ocean forcing of the Greenland ice sheet for modelling of sea level contribution, The Cryosphere, 14, 985-1008, https://doi.org/10.5194/tc-14-985-2020, 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.

Štrumbelj, E., and Kononenko, I.: Explaining prediction models and individual predictions with feature contributions, Knowledge and information systems, 41(3), 647-665, 2014.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E.: Super learner, Statistical applications in genetics and molecular biology, 6(1), 2007.

Wieland, R., Lakes, T., and Nendel, C.: Using Shapley additive explanations to interpret extreme gradient boosting predictions of grassland degradation in Xilingol, China, Geoscientific Model Development, 14(3), 1493-1510, 2021.

Wright, M. N. and Ziegler, A.: ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R, J. Stat. Softw., 77, 1-17, 2017.

Yip, S., Ferro, C. A., Stephenson, D. B., and Hawkins, E.: A simple, coherent framework for partitioning uncertainty in climate predictions, Journal of Climate, 24(17), 4634-4643, 2011.

Appendix A Model characteristics

Table A1. Model characteristics used in the MIROC5,RCP8.5-forced GrIS MME considered in the study (adapted from Goelzer et al., 2020: Appendix A).

Model ID	Numerics	Ice flow	Initialisation	Initial year	Initial SMB	Velocity	Bed	Surface	GHF	Res min (km)	Res max (km)
AWI-ISSM1	FE	HO	DAv	1990	RA	J	M		G	1	7.5
AWI-ISSM2	FE	HO	DAv	1990	RA	J	M		G	1	7.5
AWI-ISSM3	FE	HO	DAv	1990	RA	J	M		G	0.75	7.5
BGC-BISICLES	FE	SSA	DAv	2000	HIR	RM	M			1	4.8
GSFC-ISSM	FE	SSA	DAv	2007	RA	J	M		SR	0.5	25
ILTS_PIK-SICOPOLIS1	FD	SIA	NDs	1990	ISMB	J	M	M	G	5	5
ILTS_PIK-SICOPOLIS2	FD	HYB	NDs	1990	ISMB	J	M	M	G	5	5
IMAU-IMAUICE1	FD	SIA	NDm	1990	RA		M		SR	16	16
IMAU-IMAUICE2	FD	SIA	NDm	1990	RA		M		SR	8	8
JPL-ISSM	FE	HYB	DAv	1979	MAR	RM	M		SR	0.25	15
JPL-ISSMPALEO	FE	SSA	DAv	1979	RA	RM	M		SR	3	30
LSCE-GRISLI	FD	HYB	DAs,i	1995	MAR		M	M	SR	5	5
MUN-GSM1	FD	HYB	NDm	1980	MAR		B		MIX	5	14
MUN-GSM2	FD	HYB	NDm	1980	MAR		B		MIX	5	14
NCAR-CISM	FE	HO	Das,i	1990	MAR		M	M	SR	4	4
UAF-PISM1	FD	HYB	NDs	2008	RA		M	M	SR	0.9	0.9
UAF-PISM2	FD	HYB	NDs	2008	RA		M	M	SR	0.9	0.9
UCIJPL-ISSM1	FE	HO	DAv	2007	RA	RM	M		SR	0.5	30
UCIJPL-ISSM2	FE	HO	DAv	2007	RA	RM	M		SR	0.2	20
VUB-GISM	FD	HO	DAs,i	1990	MAR		M	M	SR	5	5
VUW-PISM	FD	HYB	NDs	2000	RA		M		SR	2	2

The modelling assumptions colored in grey were not considered in the analysis, namely velocity type, surface/thickness, and geothermal heat flux GHF because they are not commonly shared across the different models.

Appendix B ML models and hyperparameters' definition

Let us first denote $sl^{i=1,\dots,n}$ the ith value of sea level change calculated relative to the ith vector of p input parameters' values $\boldsymbol{x}^{i=1,\dots,n} = \{x_1, x_2, \dots, x_p\}^{i=1,\dots,n}$ where n is the total number of experiments. In the following, we present the machine-learning ML models used in the study as well as their hyperparameters.

B.1 Linear (LIN) regression model

The linear (LIN) regression model is given by:

$$sl = \beta_0 + \sum_{j=1}^{p} \beta_j x_j, \tag{B1}$$

where the β_j are regression coefficients that are estimated using a least-square criterion minimization method.

B.2 Random Forest (RF) regression model

The Random Forest (RF) regression model is a non-parametric technique based on a combination (ensemble) of tree predictors (using regression tree, Breiman et al. 1984). Each tree in the ensemble (forest) is built based on the principle of recursive partitioning, which aims at finding an optimal partition of the input parameters' space by dividing it into L disjoint sets R_1, …, R_L to have homogeneous Y_i values in each set $R_{l=1,\dots,L}$ by minimizing a splitting criterion (for instance based on the sum of squared errors, see Breiman et al. 1984). The minimal number of observations in each partition is termed nodesize (denoted *ns*).

The RF model, as introduced by Breiman (2001), aggregates the different regression trees as follows: (1) random bootstrap sample from the training data and randomly select m_{try} variables at each split; (2) construct n_{tree} trees T($\boldsymbol{\alpha}$), where $\boldsymbol{\alpha}_t$ denotes the parameter vector based on which the tth tree is built; (3) aggregate the results from the prediction of each single tree to estimate the conditional mean of *sl* as:

$$\mathrm{E}(sl|\boldsymbol{X} = \boldsymbol{x}) = \sum_{j=1}^{n} w_j(\mathbf{x}) sl^j, \tag{B2}$$

where E is the mathematical expectation, and the weights w_j are defined as

$$w_j(\mathbf{x}) = \frac{\sum_{t=1}^{n_{\mathrm{tree}}} w_t(\mathbf{x}, \boldsymbol{\alpha}_t)}{n_{tree}} \text{ with } w_j(\mathbf{x}, \boldsymbol{\alpha}) = \frac{\mathrm{I}\{X_i \epsilon \mathrm{R}_{l(x,\alpha)}\}}{\#\{j: X_i \epsilon \mathrm{R}_{l(x,\alpha)}\}}, \tag{B3}$$

where I(A) is the indicator operator which equals 1 if A is true, 0 otherwise; $\mathrm{R}_{l(x,\alpha)}$ is the partition of the tree model with parameter $\boldsymbol{\alpha}$ which contains $\boldsymbol{x}$.

The RF hyperparameters considered in the study are *ns* and m_{try} which have shown to have a large impact on the RF performance (Probst et al., 2019). The number of n_{tree} was set up to a large value of 2,000 because of its smaller influence on the RF model performance (relative to *ns* and m_{try}).

B.3 Gradient tree boosting (XGB) regression model

Gradient tree boosting (Friedman, 2001) is a tree ensemble method like RF model but differs regarding how trees are built (gradient boosting builds one tree at a time), and how tree-based results are combined (gradient boosting combines results along the process).

Formally let us denote by $f_j(\mathbf{x}) = w_j(\mathbf{x}, \boldsymbol{\alpha})$ the jth tree model prediction. The set of tree models are learnt by minimizing the following regularized objective:

$$\sum_{i=1}^{n} l\left(sl_i, \widehat{sl_i}\right) + \sum_{t=1}^{n_{tree}} \Omega(f_t), \tag{B4}$$

where $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ with T the number of leaves in the t-th tree, and γ, and λ are two regularization parameters.

The first term l is a differentiable convex loss function that measures the difference between the prediction $\widehat{sl_i}$ and the true value sl_i. The second term Ω penalizes the complexity of the regression tree functions. Equation (B4) is solved through an additive training procedure by using a scalable implementation of Chen and Guestrin (2016) of tree boosting named "XGBoost". Among the different hyperparameters of this algorithm, we focus on:

- The maximum depth of the tree models, which corresponds to the number of nodes from the root down to the furthest leaf node. This hyperparameter controls the complexity of the tree model;
- The learning rate, which is a scaling factor applied to each tree when it is added to the current approximation. Low rate value means that the trained model is more robust to overfitting but slower to compute;
- The maximum number of iterations of the algorithm.