https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

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]IGE, University Grenoble Alpes, 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 by combining the game-theoretic approach known as 'SHAP' (SHapley Additive exPlanation) with machine-learning regression models. 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) the numerical implementation, (2) the initial conditions, and (3) the modelling of ice-sheet processes. 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 Global Climate Model (GCM) output. To cover the large spectrum of uncertainties that impact the outcomes of these numerical models, common sets of numerical experiments can be performed by considering 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-sheet of Greenland (Goelzer et al., 2020), and Antarctica (Seroussi et al., 2020).

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

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. For this purpose, the key is not only to deliver modelling results, but also to explain why the numerical model delivered these results (Molnar, 2022).

In this view, many studies have adopted a global approach by focusing on the quantification of the MME spread and on the identification of 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 here an alternative local approach that focuses on how a particular setting of the modelling assumptions (i.e. value of a given model parameter, a given ISM formulation, etc.) influences the 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 modeling assumptions leads to a certain prediction. As described by Štrumbelj and Kononenko (2014), if this 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 derive local importance 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

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

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, up to 50-100, 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 dependency 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 Data

The motivating test case is based on the MME study carried out by Goelzer et al. (2020) in the framework of the ISMIP6 initiative. In the following, we only summarize the main aspects 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 GCMs that produce climate projections according to given greenhouse gas forcing scenarios; (2) a Regional Climate Model (RCM) that locally downscales the GCM 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) that produce projections of ice mass changes and sea-level contributions. The ISMs are forced by surface mass balance 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 GCMs to the retreat of tide-water glaciers (Slater et al., 2019). The parameter that controls retreat is denoted κ and is used to sample uncertainty in the parameterisation.

As the primary objective of this work is to evaluate the relevance the 'SHAP' approach, we focus on a subset of the original GrIS MME study based on one GCM (MIROC5 forced under the most impactful climate scenario RCP8.5) for which the sensitivity to the parameter κ has been sampled, namely the experiments denoted *exp05*, *exp09* and *exp10* in Goelzer et al. 2020: Table 1. A total of 55 numerical experiments were extracted to analyse the time evolution of sea level changes with

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

95 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.

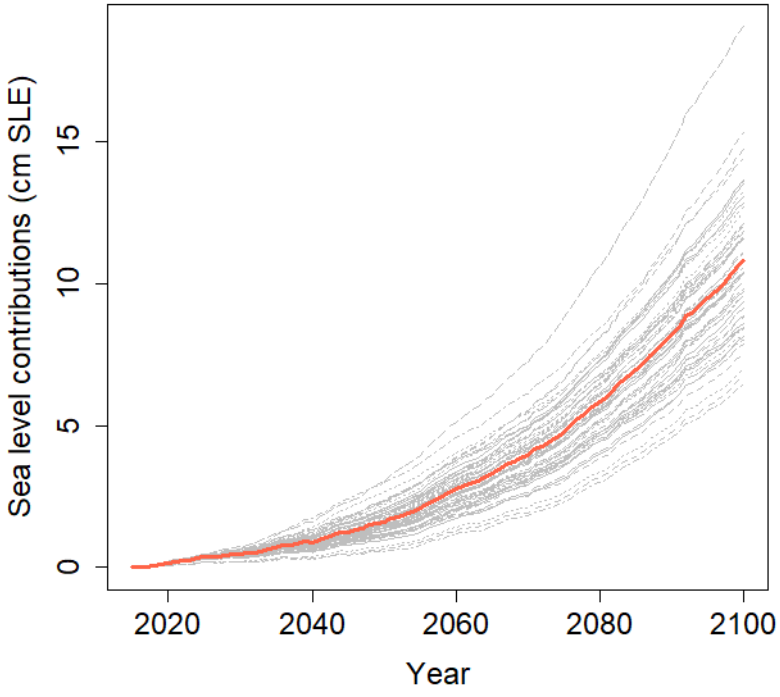

Figure 1: (a) Time evolution of the sea level contribution (with respect to 2015) from the Greenland ice-sheet (in cm sea-level equivalent, SLE). The results are the MIROC5,RCP8.5-forced MME of Goelzer et al . (2020). The red straight line is the temporal
100 **ensemble mean.**

The analysis is focused on six main modelling assumptions related to different aspects of the modelling chain (Table 1). 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
105 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).

110

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

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

Type	Modelling assumption	Symbol	Value range / Categories	Grouping of categories
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 resolution	*res*	From 0.25 to16 km	
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
Ice sheet processes	Type of ice flow	*iceFlow*	Shallow-ice approximation (*SIA*), shallow-shelf approximation (*SSA*), higher order (*HO*), *SIA* and *SSA* combined (*HYB*)	
Environmental forcing	value of the retreat parameter	κ	From -0.9705 to +0.0079 km.$(m^3.s^{-1})^{-0.4}$ °C	

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

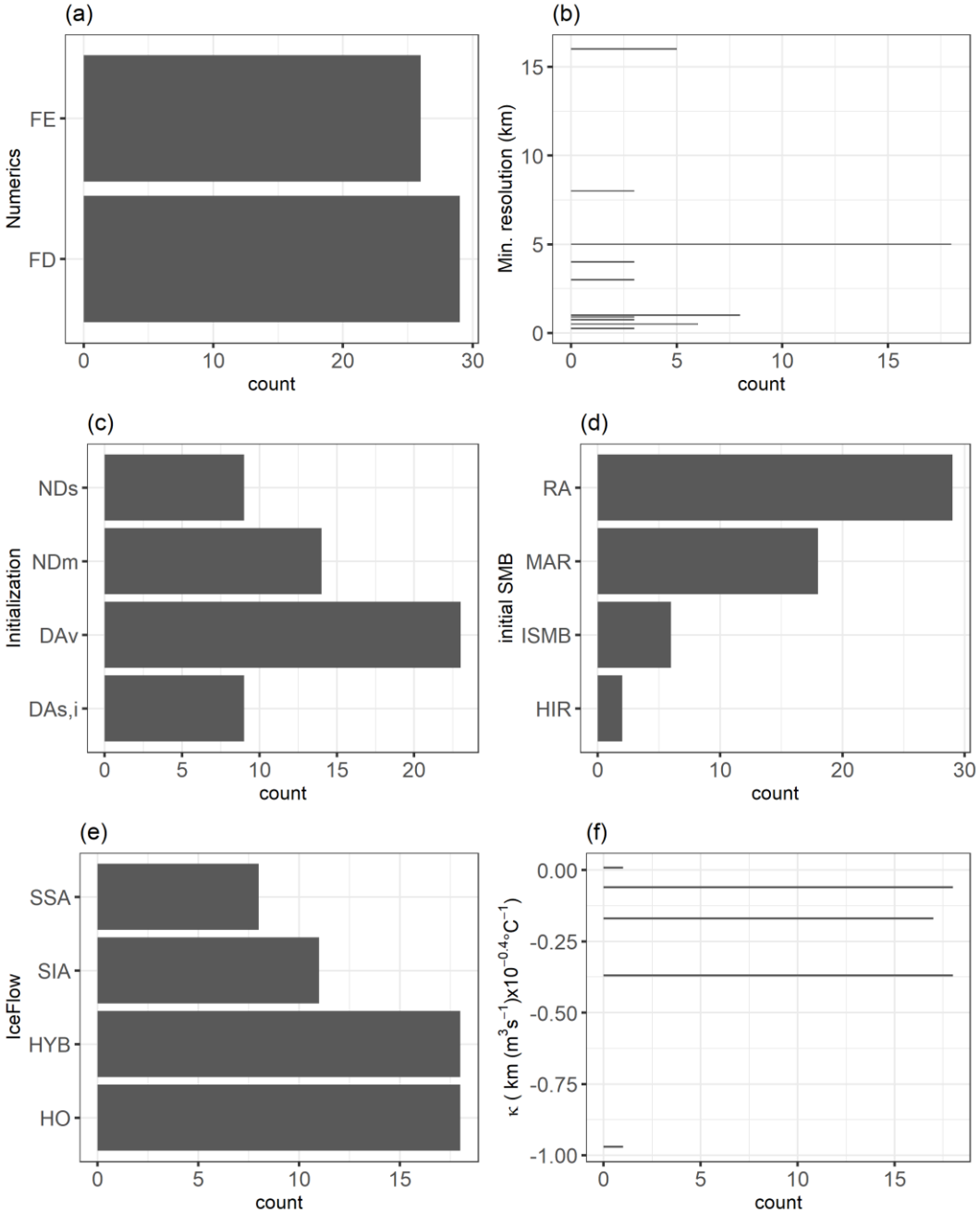

Figure 2: Count number of the MIROC5,RCP8.5-forced GrIS MME members with respect to the modelling assumptions: numerical method (a); minimum spatial resolution (b); initialisation type (c); initial SMB (d); ice flow (e); retreat parameter κ (f).

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 treated as continuous variables (for κ and resolution), and as categorical

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

variables (for the four other ones). Figure 2 shows that the design of experiments is unbalanced: some categories (like *RA* for instance, Fig. 2d) or with some values (like resolution at 5km, Fig. 2b) that 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. 2f), 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 $sl(t)$ the sea level change (with respect to a reference date) at a given time t. The chain of models described in Sect. 2 denoted f(.) takes up the different choices made for p different modelling assumptions as inputs (e.g. choice in initial SMB / ice flow formulation, value of the grid size, etc.). In our case p=6 (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 $\mathbf{x} = \{x_1, x_2, \dots, x_p\}$. We consider n different experiments; each of them associated to a particular $\mathbf{x}^{(i)}$. The MME results at a given time t are $\{sl^{(i)}(t), \mathbf{x}^{(i)}\}_{i=1,\dots,n}$ with $sl^{(i)}(t) = f(\mathbf{x}^{(i)})$. 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 ML model is denoted $\tilde{f}_{\boldsymbol{\theta}}(.)$ 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 $f(\mathbf{x}^*)$ for a given time horizon t by decomposing it as a sum of the inputs' contributions $\mu_i^*(t)$ (specific to $\mathbf{x}^*$) as follows:

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

where $\mu_0(t)$ (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 $f(\mathbf{x}^*)$. This means that the decomposition provides insights into the influence of the particular instance of the inputs $\mathbf{x}^*$ relative to $f(\mathbf{x}^*)$: (1) the absolute value of $\mu^*(t)$ informs on the magnitude of the influence at time t directly expressed in physical unit (for instance in centimetres for sea level), which eases the interpretation; (2) the sign of $\mu^*(t)$ indicates the direction of the contribution, i.e. whether the considered modelling assumption pushes the prediction higher or lower than the base value $\mu_0(t)$.

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

In order to quantify $\mu^*(t)$ 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 $\tilde{f}_{\boldsymbol{\theta}}(.)$ 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 (LIN), random forest (RF) regression method (Breiman, 2001), and Extreme Gradient Boosting (XGB) for regression (Chen and Guestrin 2016)). 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 k-fold cross validation procedure (Hastie et al., 2009). This validation can be performed by considering the different parametrisations of the ML methods, i.e. LIN, RF, and XGB models associated with different configurations of the hyperparameters $\boldsymbol{\theta}$. The ML model that performs the best is then 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. Once the ML model has been validated, this step aims to perform the additive decomposition (Eq. 1). 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 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 inputs' dependence by application of methods described in Sect. 3.4;

Step 4 Summarize 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 i^{th} 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 normalized version of μ^*, i.e. $\mu_n(t) = \mu^*(t)/(sl^*(t) - \mu_0(t))$. An application is provided in Sect. 4.3.3.

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

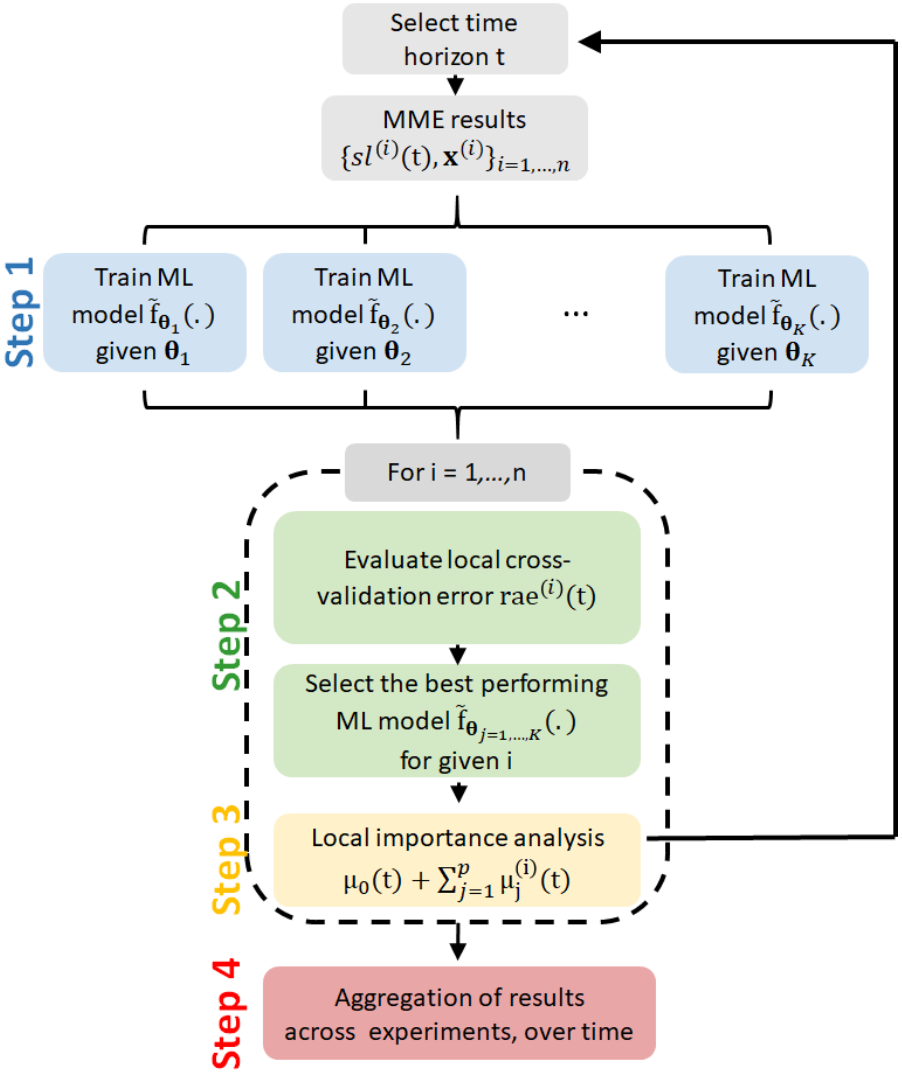

Figure 3: Schematic overview of the different steps of the procedure.

3.2 Predictive capability of the ML models

The objective is to assess the validity of replacing f(.) by $\tilde{f}_{\boldsymbol{\theta}}(.)$. 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 $\tilde{f}(.)$ 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 ***T***. Ideally, the analysis can be done by defining an independent test set ***T*** in addition to the MME results. In the

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

absence of such independent dataset, we preferably rely on a k-fold cross validation procedure (Hastie et al., 2009) that uses part of the available MME results to train the ML model, and a different part to test it. At a given time t, the procedure holds as follows.

- Step 1. The MME data is first randomly split into k roughly equal-sized parts. Using k=10 means that approximately 10% of the member results are retained in each part;
- Step 2. For the i[th] part, the ML model is fitted using the other k−1 parts of the data, and the prediction error measured for instance by difference $sl^{(i)}(\mathrm{t}) - \widehat{sl}^{(i)}(\mathrm{t})$ is calculated when predicting the ith part of the data;
- Step 3. The procedure is re-conducted for i= 1,2,...,k and the global performance indicators are calculated by combining the k estimates of the prediction error.

A widely-used performance indicator is the coefficient of determination $Q^2(\mathrm{t})$ evaluated at a given time horizon t as follows:

$$Q^2(\mathrm{t}) = 1 - \frac{\sum_{i\in \boldsymbol{T}}\left(sl^{(i)}(\mathrm{t}) - \widehat{sl}^{(i)}(\mathrm{t})\right)^2}{\sum_{i\in \boldsymbol{T}}\left(sl^{(i)}(\mathrm{t}) - \overline{sl}(\mathrm{t})\right)^2}, \quad (2)$$

where $\widehat{sl}^{(i)}(\mathrm{t})$ is the i[th] prediction of the model output $sl^{(i)}(\mathrm{t})$ at time t, and $\overline{sl}(\mathrm{t}) = \frac{1}{|\boldsymbol{T}|}\sum_{i\in \boldsymbol{T}} sl^{(i)}(\mathrm{t})$ is the average value for the test set at time t, and $|\boldsymbol{T}|$ is the size of the test set. A coefficient $Q^2(\mathrm{t})$ close to 1.0 indicates that the ML model is successful in matching the new observations that have not been used for the training at time t. Similar as for $Q^2(\mathrm{t})$, the mean absolute error $MAE=\frac{1}{|\boldsymbol{T}|}\sum_{i\in \boldsymbol{T}} |sl^{(i)}(\mathrm{t}) - \widehat{sl}^{(i)}(\mathrm{t})|$ over time can also be calculated.

Since we are interested in local explanations, the afore-described assessment of the global predictive capability should be complemented by some local performance indicators as well. In fact, high Q^2 values do not necessarily ensure that all predictions of f(.) are appropriately predicted by the ML model. For some cases, the discrepancies can be large (see some examples in Sect. 4.2). To increase our confidence in replacing f(.) by $\tilde{\mathrm{f}}_{\boldsymbol{\theta}}(.)$, we consider a series of different ML models (of different types and each of them with different settings of the hyperparameters $\boldsymbol{\theta}$.). We then select the ML model that leads to the highest local predictive quality for a given case i. The latter is measured by the relative absolute difference $\mathrm{rae}^{(i)}(\mathrm{t}) = \left|\frac{sl^{(i)}(\mathrm{t}) - \widehat{sl}^{(i)}(\mathrm{t})}{sl^{(i)}(\mathrm{t})}\right|$ (calculated from the cross-validation procedure). For a given case and at time t, the best performing ML model regarding $\mathrm{rae}^{(i)}(\mathrm{t})$ is then retained for the local explanation analysis described in Sect. 3.3.

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 and a residual degree of prediction error may still remain. This has implications for the interpretation of low $|\mu_j^*(\mathrm{t})|$ values. In theory, $|\mu_j^*(\mathrm{t})| = 0$ means that the j[th] 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

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

error. In the following, we propose to use $\mathrm{rae}^{(i)}(t)$ as an indicator of residual prediction error to assess whether some inputs have a negligible influence.

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 ith player get is defined by the Shapley value with respect to val(.):

$$\mu_i(t) = \frac{1}{k}\sum_{S \subseteq K \setminus \{i\}} \binom{k-1}{|S|}^{-1} (\mathrm{val}(S \cup \{i\}) - \mathrm{val}(S)), \quad (3)$$

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:

$$\mathrm{val}(S) = \mathrm{E}(\tilde{f}(\mathbf{x})|\mathbf{x}_S = \mathbf{x}_S^*), \quad (4)$$

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 $\mathrm{E}(f(\mathbf{x}))$. In this way, μ_i^* corresponds to the change in the expected model prediction (Eq. 5) when conditioning on that input and explains how to get from the global mean prediction. 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 $\sum_{i=1}^{p} \mu_i^* = f(\mathbf{x}^*) - \mathrm{E}(f(\mathbf{x}^*))$, 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).

In practice, the Shapley values are computed using 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. 4-5. For this purpose, we use the R package '*shapr*' (Sellereite and Jullum, 2020) with accounts for inputs' dependence (see Sect. 3.4).

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

3.4 Accounting for inputs' dependencies

In the case of dependence among the inputs, 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 expected value in Eqs. 4-5. 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 model the dependence structure of input variables that are of mixed types (i.e. continuous, discrete, ordinal, and categorical).

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 (Fig. 3) to the MIROC5,RCP8.5-forced GrIS MME described in Sect. 2. 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 and summarize the results to provide information on sensitivity at different levels (Steps 3-4, Sect. 4.3).

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 tree model is empty, which indicates the absence of (significant) dependence between this

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

parameter and the other modelling assumptions. The different tree models should be read by following the example of the top, 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).

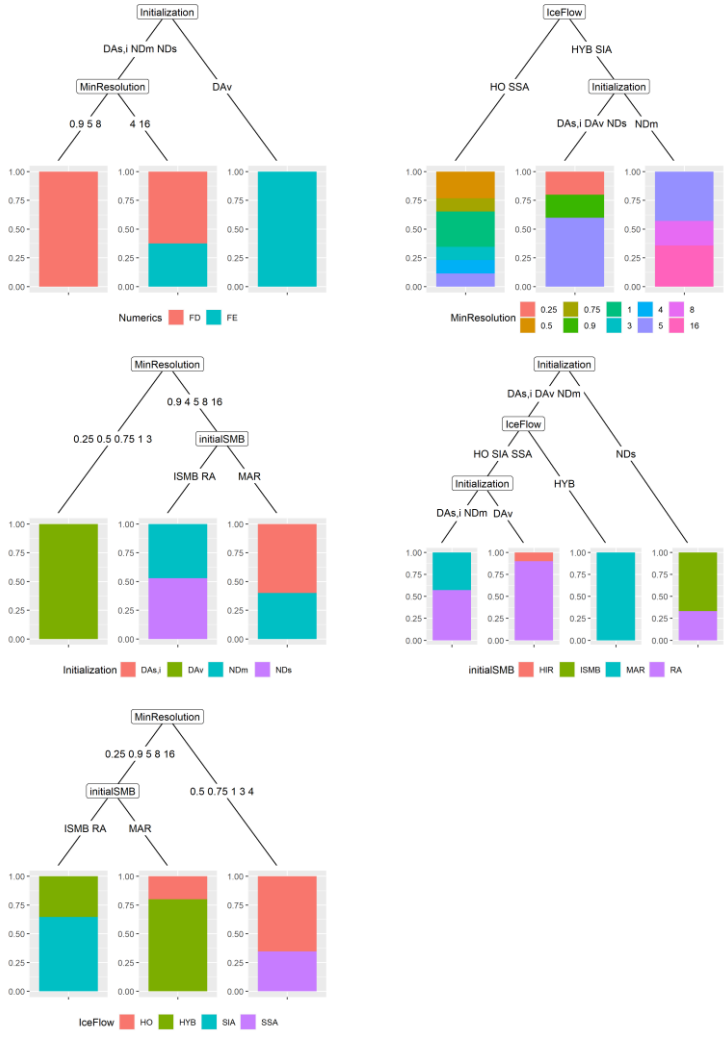

Figure 4: Tree models representing the dependence between the different modelling assumptions (indicated at the bottom). The bottom nodes (leaf nodes) provide the proportion of experiments given the modelling choices defined along the branches of the tree model. Each colour corresponds to a different category of the considered modelling assumption. For instance, the top, 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.

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

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*=1, 5 or 10, m_{try}= 1, 3 or 6 and n_{tree}=2,000;
- Thirty XGB models with hyperparameters maximum depth = 2,3,4,5 or 6, 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 k-fold cross validation approach (with k=10) by following the procedure of Sect. 3.2. Figure 5a depicts Q^2(t) for all considered ML models. Depending on the type of ML model (and corresponding parametrisation), the performance can reach satisfactory levels above 90% 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 perfectly approximates 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 RF model performs the best (outlined in orange 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. The retained predictions are outlined in blue in Fig. 5b.

In total, LIN, XGB and RF models are retained respectively 7%, 29% and 64% of the total number of experiments. After applying this procedure, the Q^2 criterion (shown in blue in Fig. 5a) reaches values very close to 1.0 (with a minimum value not lower than 0.98).

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

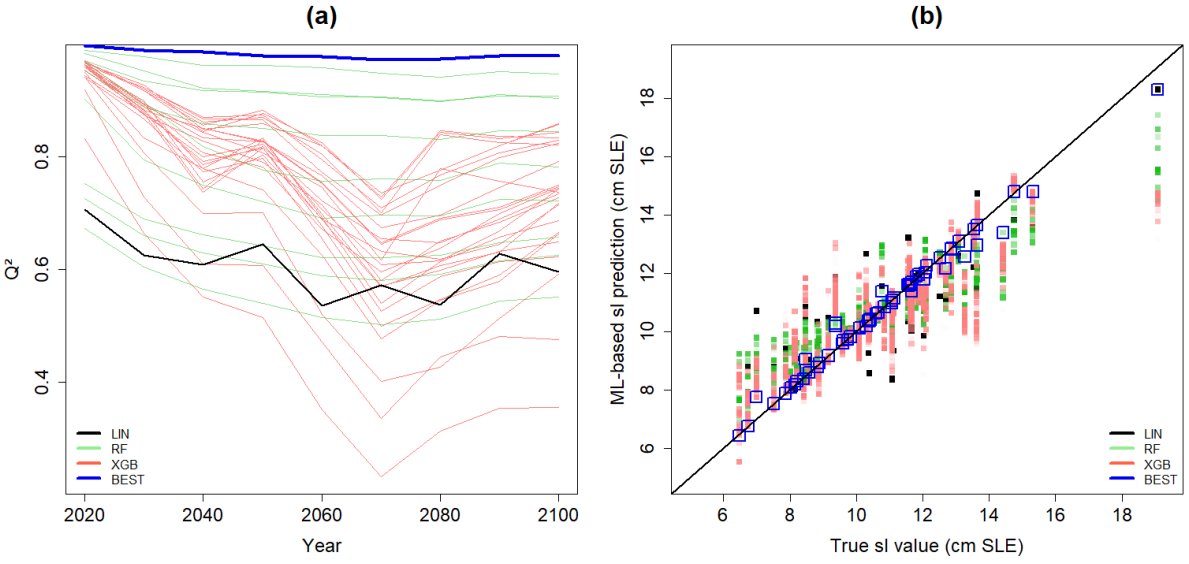

Figure 5: (a) Time evolution of the performance criterion Q^2 computed using a 10-fold cross validation procedure that assesses the predictive capability of all considered ML models with different parametrisations (RF models in green, XGB in red and LIN in black). The blue-coloured lines are related to the performance criterion after selecting the best performing ML model with respect to the relative absolute error; (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 with respect to the relative absolute error.

4.3 From local to global explanations

In this section, we first perform the local explanations 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 to deepen our understanding of the model structure for a given prediction time. Finally, Sect 4.3.3 summarizes all results over time (*Level 3* of the procedure) to provide a global insight in the sensitivity 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. 3-4) 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).

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

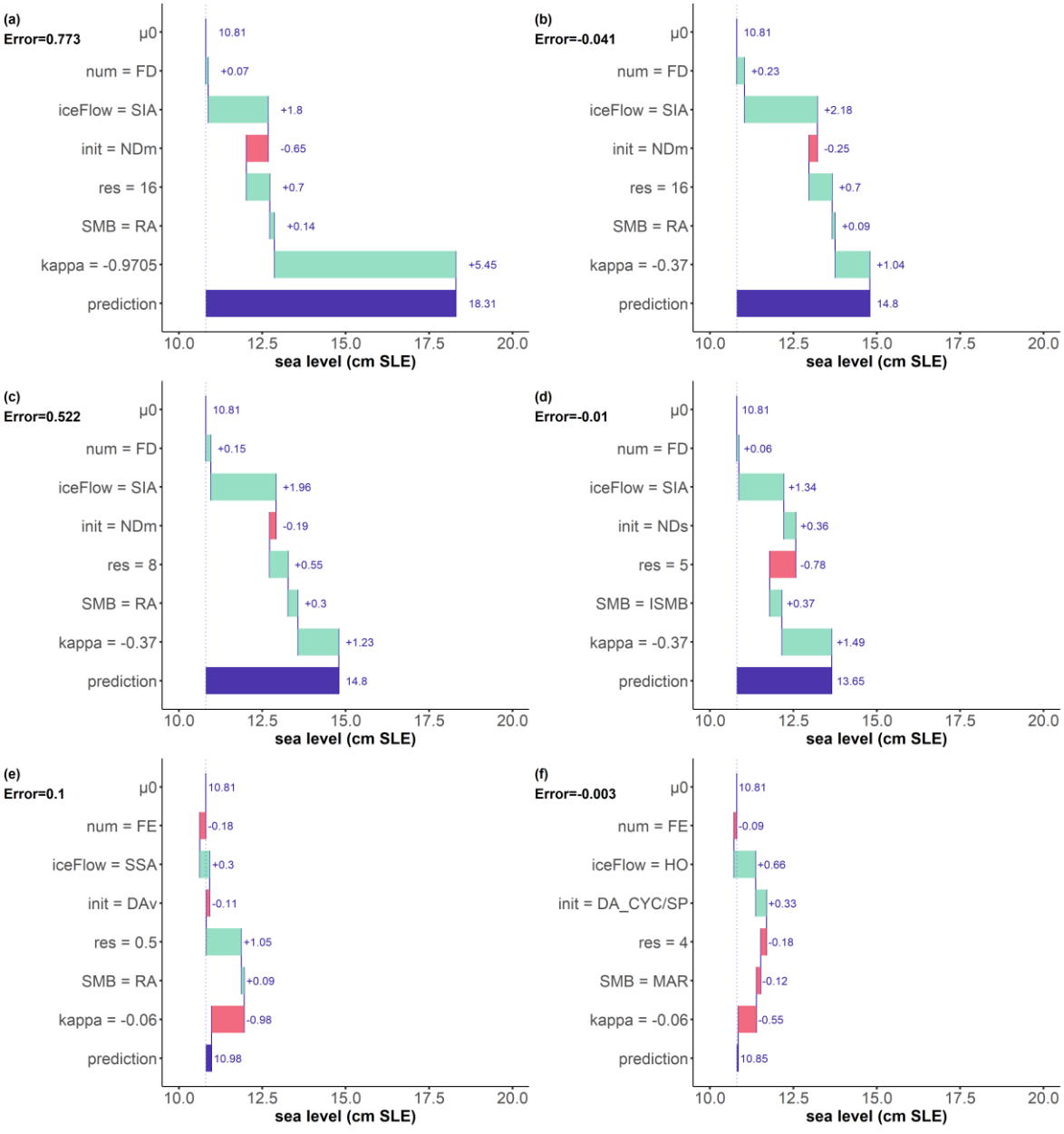

Figure 6: Diagnostic of particular ML-based *sl* predictions using SHAP for year 2100 considering six different settings of the modelling choices (indicated in the vertical axis). The horizontal blue bar corresponds to the ML-based *sl* prediction (the difference with the true value is indicated by the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

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

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

- Case (a) corresponds to the largest *sl* value (of 18.3cm). 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=+5.45cm, i.e. by >50% of μ_0. In this case, the second largest contributor to *sl* (with an influence of +1.8cm) is related to assuming an ice flow of type *SIA*. The other modelling choices all have absolute contributions <1 cm, i.e. less than the absolute difference between the ML-based prediction and the true value. This means that this low contribution is hardly distinguishable from the prediction error level, and these inputs, for this particular instance, can be treated of negligible influence;
- Case (b) (Fig. 6(b)) corresponds to the second largest *sl* value (of 14.8cm). 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. In this case, it is the choice in the ice flow type (here *SIA*) that contributes the largest to *sl*. The influence of κ drops here to low-to-moderate value;
- Case (c) shows that the reduction in the minimum grid size (from 16 to 8 km) has here only very low influence when considering the same setting than Case (b);
- Case (d) corresponds to the third largest *sl* value and illustrates that, despite the differences with Case (c) (i.e. initial SMB, initialisation type and minimum resolution), the contribution of ice flow's type and κ remain unchanged between both cases;
- Cases (e) and (f) illustrate that the modelling choices contribute differently to the prediction although the predicted vales are very close (here close to the ensemble mean of 10.8cm). In Case (e), it is κ and grid resolution that contribute the most (with compensated effect), whereas it is κ and ice flow's type in Case (f).

Such type of diagnostic can be performed for any MME results (they are all provided by Rohmer (2022) for year 2100).

4.3.2 Level 2. Model structure at a given prediction time

We explore how the contribution magnitude, 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).

Fig. 7a gives insights into the influence of κ and confirms its large influence (of several cm) for large absolute value. 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* (calculated from the cross-validation procedure, see Sect. 4.2). A clear trend is outlined by the smooth regression in Fig. 7a: κ influence decreases with increasing value in a quasi-linear manner (with slope of ~-8 cm per unit of retreat parameter). Figure 7a also provides indications where to perform additional numerical experiments to confirm κ influence, namely over then range -0.97 to -0.37 km.$(m^3.s^{-1})^{-0.4}$ °C (where the results are scarce).

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

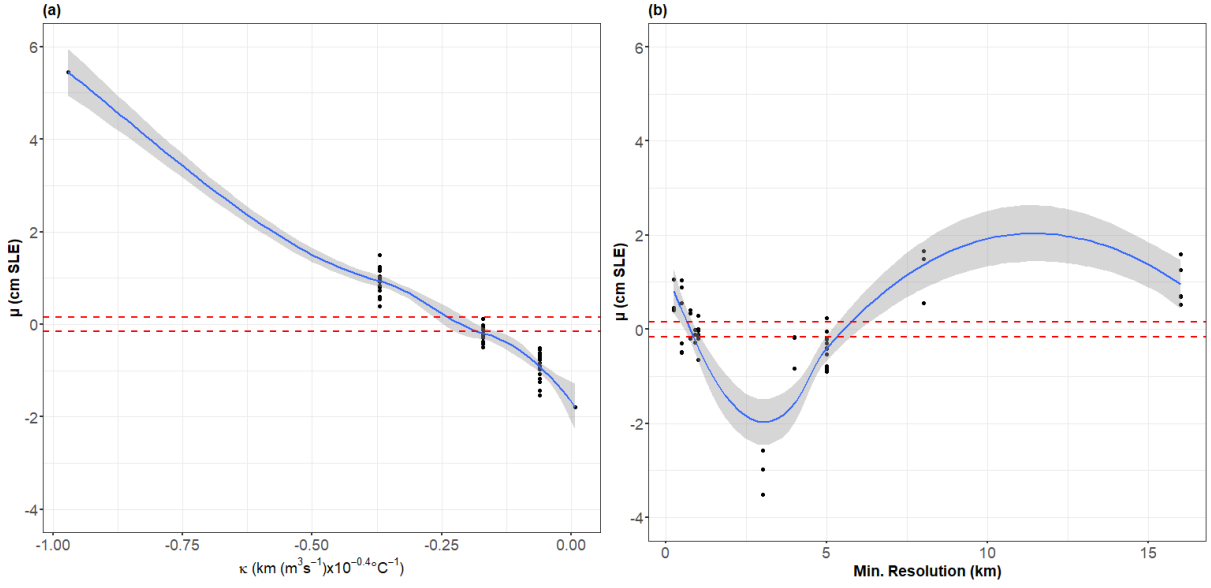

Figure 7: Application of SHAP additive explanation to all members of MIROC5,RCP8.5-forced GrIS MME for year 2100. Each panel provides μ (y-axis) as a function of the value of the retreat parameter κ (a), and of the minimum grid resolution (b). The horizontal dashed red lines indicate the limits defined by *MAE* calculated from the cross-validation procedure. 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.

Fig. 7b shows that the influence of the minimum grid size can be large up to ~3cm for values >5km, and minor provided that it is ≤2km. For grid size ranging 3-4km, some non-negligible negative contributions are outlined but this deserves further investigations because the trend modelled by the smooth regression is driven by only few MME results (5 in total). From a modelling perspective, this analysis suggests that there is clear interest in performing high resolution simulations (with minimum grid size≤2km): too coarse simulations might highly contribute to the final *sl* value.

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

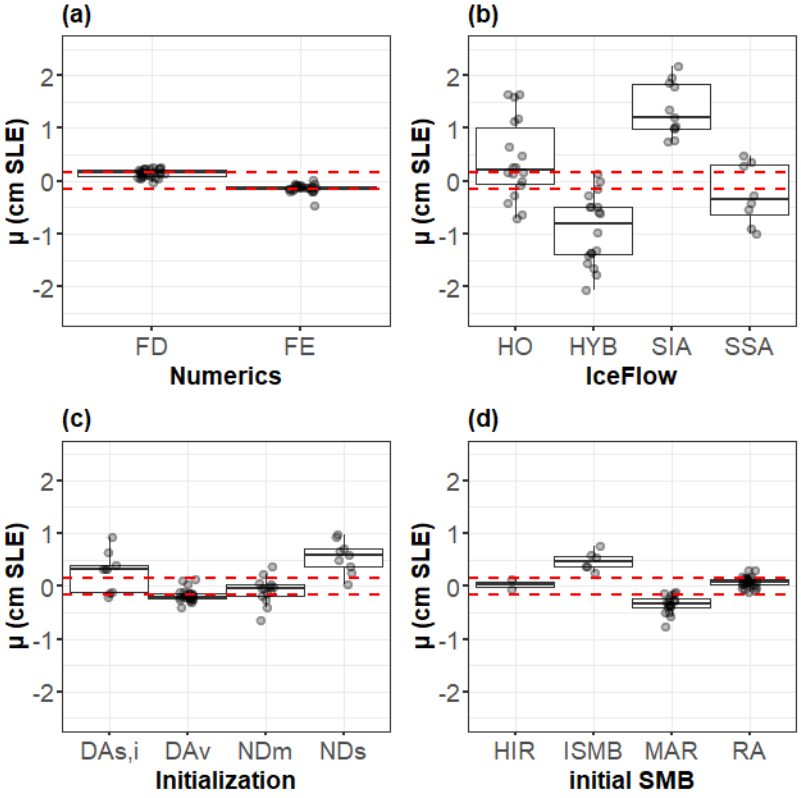

Figure 8: Application of SHAP additive explanation to all members of MIROC5,RCP8.5-forced GrIS MME for year 2100. Each panel provides the boxplots of μ values given the modelling choice for the numerical method (a), the ice flow (b), the initialisation (c) and the initial SMB (d). Each dot corresponds to a given MME member. The horizontal dashed red lines indicate the limits defined by *MAE* calculated from the cross-validation procedure.

Fig. 8 further suggests that the contribution of minimum grid size might mask the one of the other modelling choices: their μ values do not exceed 2.5cm, i.e. of the same order of magnitude than the one of the minimum grid size. The largest influence is here mostly attributable to the ice flow's choice, either of *SIA* or *HYB* type with positive or negative contribution: the corresponding boxplot in Fig. 8b is well above (below) zero. Finally, Fig. 8 also points out some modelling choices with negligible contribution, namely any type of numerical method *FD* or *FE* (Fig. 8a), *NDm* (and to a lesser extent, *DAv*) for initialisation (Fig. 8c), and *HIR* or *RA* for initial SMB (Fig. 8d).

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 $\mu_n(t) = \mu(t)/(sl(t) - \mu_0(t))$. To judge on the negligible level of the influence with respect to the ML

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

prediction error, we analyse the quartiles of $rae_n(t) = \left|\frac{sl(t) - \widehat{sl}(t)}{sl(t) - \mu_0(t)}\right|$. Fig. 9 shows that in the very short term (before 2030), all modelling assumptions, except κ, contribute to *sl* value (with a median value of $\mu_n(t)$ ranging between 0.25 and 0.50). After 2030/2040, the modelling assumptions related to ice sheet processes (ice flow's type and κ) are all major contributors to *sl* with a quasi-constant influence over time after this date. This time evolution suggests that there is a transitional time for the influence of these modelling assumptions to be noticeable in the *sl* values. In the medium / long term (after 2050), the type of numerical method and of initial SMB both show small (or even negligible) contributions to *sl* values, and it is the ice flow type, κ and the minimum grid size that are important for *sl* (in agreement with Fig. 8). Finally, we note that the moderate influence of the minimum grid size remains quasi-constant over time, hence suggesting that its influence is time-invariant, i.e. all modelled processes are affected by the spatial resolution in a similar way, independently of the prediction time.

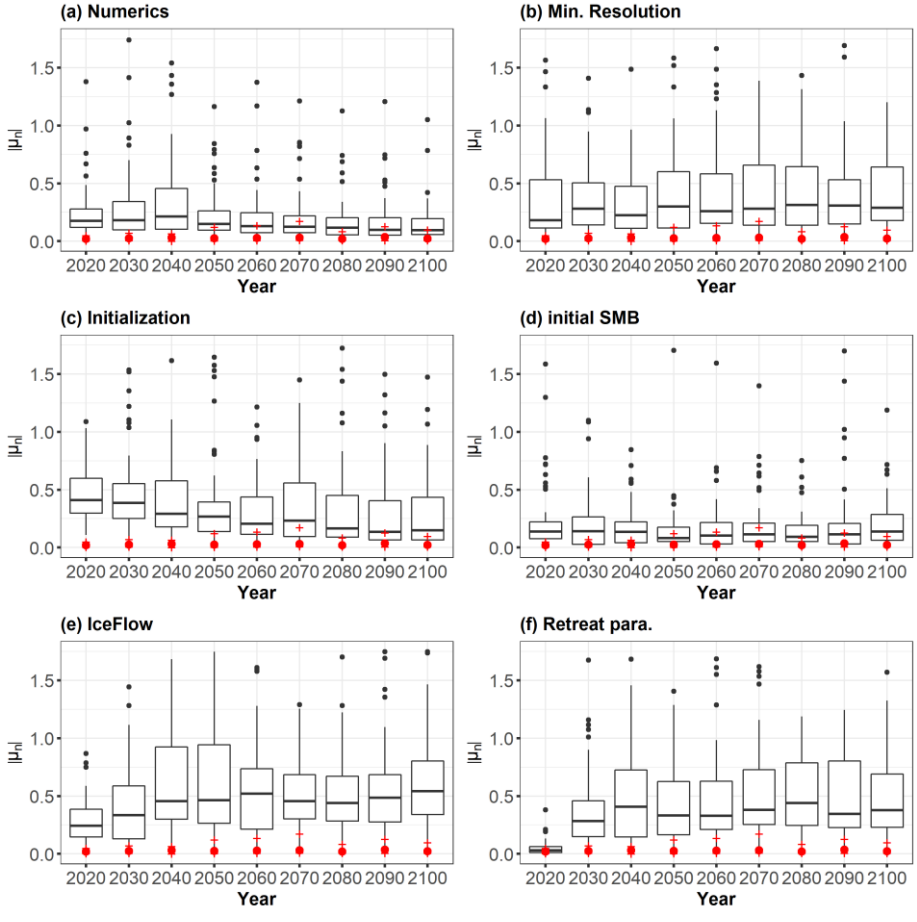

Figure 9: Time evolution of $|\mu_n|$ for all members of MIROC5,RCP8.5-forced GrIS MME. The red dot and crosses are respectively the median and the 1st and 3rd quartile of the cross-validation error $rae_n(t)$. For readability, the upper bound of the y-axis has been set up to 1.75.

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

5 Discussion

Improving the interpretability of sea level projections is a matter of high interest, because 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). To this end, we adopt the local attribution approach developed in the machine learning community, which allows providing narratives that can follow the example of the GrIS study (Fig. 6(a)): "the largest *sl* predicted value is of 18.3cm by 2100 and is mostly attributable (by a positive factor of 50% of the ensemble mean) to setting κ to its largest absolute value with only moderate influence of the other modelling assumptions". Such type of diagnostic can be performed for any MME results (they are all provided by Rohmer (2022) for 2100) and helps understanding and quantifying the impact of particular assumptions made by the modellers, i.e. by providing insights into the most and least impactful modelling choices. The aggregation of all these local sensitivity analyses further improves the understanding of the model structure (Level 2 and 3 of the proposed approach), which is helpful for guiding future model development as well as for the scientific interpretation.

In our case, our results confirm the need for sufficiently spatially resolved simulations: *sl* results are largely affected by setting up the minimum grid size to too high values (here at least 5km) regardless of the prediction time. In addition, large *sl* values are shown to be mostly attributable to large absolute κ values especially after 2050. The analysis also provides guidance for defining additional computer experiments, namely for 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, we show that some modelling choices have little impacts on the *sl* values; in particular, choosing a finite element or finite difference numerical scheme.

These results could hardly have been obtained by applying more common statistical methods (as described in the introduction), namely the linear regression model or the ANOVA-based approach. 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 with our approach. In addition, one major difficulty is related to the presence of statistical dependencies (as outlined in Sect. 4.1): this 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 alleviating this problem.

However, the key prerequisite for this performance is the high predictive capability of the ML model. Though the performance analysis showed satisfactory results in the GrIS case (Sect. 4.2), 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

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

'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).

6 Concluding remarks and further work

In this study we proposed the use of the machine-learning-based SHapley Additive exPlanation (SHAP) approach to quantify the importance of modelling assumptions regarding the sea-level projections produced in a MME study. The proposed approach is applicable to any MME study with a limited number of experiments (50-100), an unbalanced design, and the presence of dependence between the inputs. Results on a subset of the GrIS ensemble 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.

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 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 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.

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

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

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 869304, PROTECT.

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

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

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.

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.

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

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.

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.

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

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

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.

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.

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.

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

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], Zenodo. https://doi.org/10.5281/zenodo.6606487, (last access 2 June 2022), 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.

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.

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

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:

- velocity type, surface/thickness, and geothermal heat flux GHF are not commonly shared across the different models.
- the bed parametrisation was excluded because only two models are associated to different modelling choices.
- initial year was not considered because it is believed to have only minor impact.

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

Appendix B ML models and hyperparameters' definition

Let us first denote $sl^{i=1,\dots,n}$ the i^{th} value of sea level change calculated relative to the i^{th} 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:

$$sl = \beta_0 + \sum_{j=1}^{p} \beta_j x_j, \quad \text{(B1)}$$

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 t^{th} tree is built; (3) aggregate the results from the prediction of each single tree to estimate the conditional mean of *sl* as:

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

where E(.) is the mathematical expectation, and the weights W_j are defined as

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

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}).

https://doi.org/10.5194/egusphere-2022-435
Preprint. Discussion started: 16 June 2022

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:

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

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.