

Lessons for multi-model ensemble design drawn from emulator experiments: application to a large ensemble for future sea level contributions of the Greenland ice sheet

Jeremy Rohmer¹, Heiko Goelzer², Tamsin Edwards³, Goneri Le Cozannet¹, Gael Durand⁴

5 ¹BRGM, 3 av. C. Guillemin - 45060 Orléans Cedex 2 - France

²NORCE Norwegian Research Centre, Bjerknes Centre for Climate Research, Bergen, Norway

³Department of Geography, King's College London, Bush House, North East Wing, 40 Aldwych, London, WC2B 4BG, London, UK

⁴Univ. Grenoble Alpes, CNRS, IRD, Grenoble INP, IGE, 38000 Grenoble, France

10

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

Abstract.

Multi-model ensembles (MME) are key ingredients for future climate projections and the quantification of their uncertainties. Developing robust protocols to design balanced and complete computer experiments for MME is a matter of active research.

15 In this study, we take advantage of a large-size MME produced for Greenland ice sheet contributions to future sea level by 2100 to define a series of computer experiments that are closely related to practical MME design decisions: what is the added value of including specific set of members in the projections, i.e. either adding new models (Regional Climate Model, RCM, or Ice Sheet Model, ISM) or extending the range of some parameter values. By using these experiments to build a random-forest-based emulator, we assess the changes in the emulator's predictive performance and the emulator-based probabilistic
20 predictions for given temperature scenarios compared to the reference solution based on all members. For the considered MME, several aspects are outlined: (1) the highest impact of removing the most selected RCM, i.e., MAR, due to the large number of simulations available; (2) the high importance of including the SSP5-8.5 scenario for high temperature scenarios leading to under-estimations up to 30% of the considered percentiles; (3) the importance of having diverse ISM and RCM models leading to percentile absolute changes ranging between 10 and 20%; (4) the lesser importance of the choice in the
25 range of the Greenland tidewater glacier retreat parameter. We expect these recommendations to be informative for the design of next generations of MME, in particular for the next Ice Sheet Model Intercomparison Project ISMIP7 in preparation.

1 Introduction

Multi-model ensembles (MME) are key ingredients for future climate projections and the quantification of their uncertainties. They consist of sets of numerical experiments performed under common forcing conditions with different model designs (i.e.
30 different model formulations, input parameter values, initial conditions, etc.) to generate multiple realisations known as

ensemble members. This is the approach of Model Intercomparison Projects, MIPs, which are key for the understanding of past, present, and future climates and contribute to assessments from the Intergovernmental Panel on Climate Change (IPCC: e.g. Lee et al., 2021). In this study, we are interested in projected Greenland ice sheet contributions to sea level change this century, which are the subject of recent MME studies (Goelzer et al., 2018; 2020) within the Ice Sheet Model Intercomparison Project for CMIP6 (ISMIP6: Nowicki et al., 2016; 2020).

However, interpreting MME results is complicated by the choices made in their construction (e.g. Knutti et al., 2010). Ideally, each member of a MME should evenly span a representative and exhaustive set of plausible realisations of the combined sources of uncertainty, e.g. distinct climate models with different but plausible strategies for simulating the global climate (GCMs), equally represented by a single model run. However, members of a MME are often structurally similar, and the degree of their dependence is difficult to quantify (e.g. Merrifield et al., 2020). This difficulty is particularly emblematic of the Coupled Model Intercomparison Project (CMIP), coined an “ensemble of opportunity” (Tebaldi and Knutti, 2007) because it collects “best guesses” (Merrifield et al., 2020) from modelling groups with the capacity to participate. This capacity may range from substantial resources to develop climate models and perform relatively large ensembles through to the ability to perform only a small number of simulations with an existing version of a climate model. These disparities, combined with the high computational expense of climate models and the partial dependence of MME members, results in limited and unbalanced multi-model ensemble designs, in which various combinations of modelling choices and forcing conditions are either over-represented or missing in the MME, and a full sampling of modelling uncertainties is impossible to perform or even to define. Section 2.1 provides in the following an illustration for the MME considered in this study.

Emulators (also named surrogate models) have been proposed to address these limitations. An emulator is a fast statistical approximation of a computationally expensive numerical model, often building on machine learning techniques like linear-regression (Levermann et al., 2020), Gaussian process regression (Edwards et al., 2021), random forest regression (Rohmer et al., 2022), and deep learning-based methods (Van Katwyk et al., 2025). Their key advantage is that they can be used to predict with low computational cost the numerical model’s response at untried input values, to explore the uncertain input space far more thoroughly: potentially overcoming the incompleteness of ensemble designs and being used to produce probabilistic projections.

Some emulation studies have broadened this approach to represent entire MME at once, rather than individual models. One example in this field is provided by Edwards et al. (2021), who emulate ISMIP6 simulations for the Greenland and Antarctic ice sheets and multi-model glacier ensembles, driven by multi-model climate model ensemble simulations, to estimate land ice contributions to twenty-first-century sea level rise. Emulating an MME requires an assumption (and check) that the simulations are quasi-independent: i.e. that the differences induced by different model setups (in particular, initialisation) outweigh any similarities induced by common model structures. This was found by Edwards et al. (2021) to be the case for ice sheet and glacier MMEs. Another type of application is provided by Van Breedam et al. (2021) who used emulators to perform a large number of sensitivity tests with numerical simulations of ice sheet–climate interactions on a multi-million-year timescale.

65 In this study, we aim to explore how the results provided by an emulator can be informative for the design of an MME. Key design questions relate to the added value of including specific sets of experiments in the projections, i.e. either adding new models (e.g. new Regional Climate Model, RCM, new GCM, etc.) or extending the range of some parameter values (e.g., the Antarctic basal melt parameter or Greenland tidewater glacier retreat parameter described by Edwards et al. (2021)). To address these questions, we take advantage of a large MME of Greenland ice sheet contributions to sea level this century, based on
70 which we define a series of numerical experiments (referred to as emulator's experiments) that are closely related to practical MME design decisions. The evaluation of the emulator prediction capability with each of these experiments is used to provide information on the added value of including specific set of experiments.

The paper is organized as follows. We first describe the sea level numerical simulations as well as details of the statistical methods used to build the emulator and assess the different design questions (Section 2). In Section 3, we apply the experiments
75 and assess the influence of each design question. We discuss results in Section 4, and we draw lessons and guidance related to the MME design, and discuss the implications from a stakeholder's point of view. Finally, we conclude in Section 5.

2. Data and methods

2.1 Multi-model ensemble case study

80 We focus on the sea level contribution from the Greenland ice sheet (GrIS) in 2100 based on a new MME study performed for the European Union's Horizon 2020 project PROTECT (<http://protect-slr.eu>). Some modelling choices are taken from the protocols of the ISMIP6 initiative (Goelzer et al., 2020; in particular, the two main emissions scenarios, and the main model parameter explored). This MME has been designed as an extension of ISMIP6 MME through the inclusion of:

- a wider range of CMIP6 climate model output as well as more climate change scenarios (SSP126, SSP245, SSP585);
- 85 - the surface mass balance forcing from several RCMs, i.e. MAR, RACMO, and HIRHAM as well as a statistical downscaling approach of a given GCM;
- retreat forcing before 2015 that is calculated from reconstructions of past runoff and ocean thermal forcing, hence allowing for a consistent forcing of the models in past and future and to consider historical retreat of the outlet glaciers, which was an important source of mass loss after 1990.

90 We provide here a brief summary of the GrIS MME dataset and refer the interested reader to Goelzer et al. (2025) for further details, where appropriate. The full modelling chain for these projections combines: (1) a number of CMIP5 and CMIP6 GCMs that produce climate projections according to different emissions scenarios; (2) different RCMs, and their variants, that locally downscale the GCM forcing to the GrIS surface; (3) a range of ISM models that produce projections of ice mass changes and sea level contributions (initialised to reproduce the present-day state of the GrIS as best as possible, at a given initial year
95 sometime before the start of emissions scenarios in 2015). The ISMs are forced by surface mass balance (SMB) changes from the RCMs, added to their own reference SMB assumed during initialisation. Ocean forcing is integrated based on an empirically derived retreat parameterization that relates changes in meltwater runoff from the RCM and ocean temperature

changes from the GCM to the retreat of calving front positions (Slater et al., 2020, 2019). The parameter that controls retreat is denoted κ . It represents the sensitivity of the ocean forcing as a whole, and defines the sensitivity of the downscaling from global model to local ice sheet scale. Figure 1 shows the general approach used for forcing the ISMs and producing the projections. The MME design questions addressed in this study are related to the modelling choices made for each of the boxes outlined in Fig. 1.

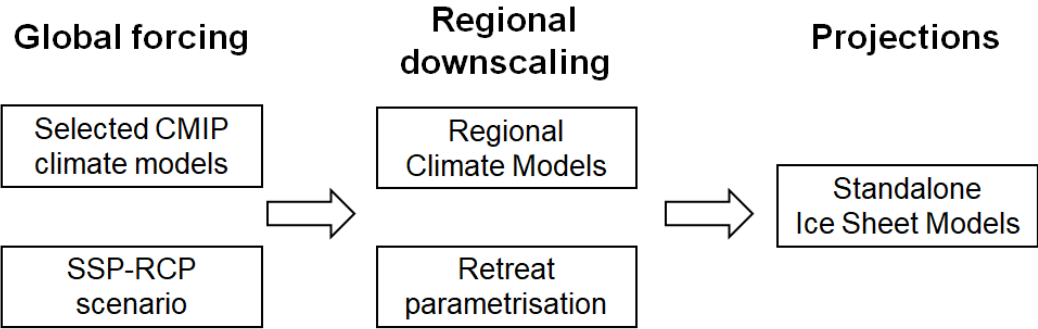


Figure 1: General forcing approach for Greenland ice sheet model projections. The questions relevant for the MME design (detailed in Table 2) are related to the modelling choices made for each of the boxes.

In what follows, we use the generic term ‘inputs’ to designate all the choices made throughout the modelling chain, i.e. the choices in the models used, the choices in the scenarios and the parameter values. The inputs are described in detail in Table 1. It should be noted that the two first inputs, i.e. the choice in the SSP-RCP scenario and in the GCM model, are not considered for the emulator construction described in Sect. 2.2. They are combined with a similar approach as Edwards et al. (2021), by relating each ‘SSP-RCP, GCM’ combination to the corresponding value of global annual mean surface air temperature change relative to 1995-2014, denoted GSAT.

Table 1: Inputs considered in the GrIS MME. The inputs from the double line are those used for the building of the RF emulator described in Sect. 2.2.

Type	Name	Type of variable	Value range / Categories
Future climate and societal conditions	SSP-RCP	Categorical	5 scenarios: three Shared Socio-economic Pathways (SSP1-2.6, SSP2-4.5, SSP5-8.5) and two Representative Concentration Pathways (RCP2.6, RCP8.5). The latter, older, scenarios are grouped with the nearest equivalent SSPs (RCP2.6 with SSP1-2.6; RCP8.5 and SSP5-8.5).
General Circulation Model	GCM	Categorical	15 global climate models: ACCESS1.3, CESM2, CESM2-Leo*, CESM2-WACCM, CNRM-CM6-1, CNRM-ESM2-1, CSIRO-Mk3.6.0, HadGEM2-ES, IPSL-CM5A-MR, IPSL-CM6A-LR, MIROC5, MPI-ESM1-2-HR, NorESM1-M, NorESM2-MM, UKESM1-0-LL-r1

Global mean temperature change 2015–2100	GSAT	Continuous	The joint influence of SSP-RCP and GCM is treated with a similar approach as Edwards et al. (2021), by relating each ‘SSP-RCP and GCM’ combination to the corresponding value of global annual mean surface air temperature change since 2015.
Ice Sheet Model	ISM	Categorical	4 models: CISM, Elmer/Ice, GISM, IMAUICE
Regional Climate Model	RCM	Categorical	6 model approaches: four versions of the RCM MAR (v3.9, v3.12, v3.13-e05, and v3.13-e55), one version of the RCM RACMO (v2.3p2), one version of the RCM HIRHAM (v5), and statistical downscaling (SDBN1).
Retreat parameter	κ	Continuous	From -0.9705 to +0.0070 km.(m ³ .s ⁻¹) ^{-0.4} °C
Minimal spatial resolution	res_min	Continuous	From 1 to 40 km
Sliding friction law	Sliding	Categorical	5 laws: Coulomb, Linear, Schoof, Weertman, Zoet-Iverson
Account for thermodynamics	thermodin.	Categorical	TRUE or FALSE
RCM used for initialisation	RCM_init	Categorical	4 model variants: IMAU-ITM, and MAR (v3.9, v3.11.5, and v3.12).
Type of initialisation method	Init	Categorical	Data assimilation based on velocities (DAv); nudging to ice mask (NDm); or nudging to surface elevation (NDs).
Number of years of the initialisation period	init_yrs	Continuous	From 20 to 240,000 years
Location of the surface elevation feedback	elev_feedback	Categorical	In the ice sheet model (with two formulations of the SMB-elevation gradient, X or B), or in the regional climate model

**CESM2-Leo is a variant pre-dating the official CESM2 release for CMIP6. It can be considered as another ensemble member of CESM2.*

One input setting, i.e., as a particular combination of inputs, defines a member of the MME. Formally, the inputs are either treated as continuous variables (e.g., for κ , minimum resolution), or as categorical variables (e.g., RCM or ISM choice). The considered MME comprises $n=1,343$ members, which are used to estimate the sea level contribution in 2100 (denoted slc expressed in meters sea level equivalent SLE) with respect to 2014. Figure 2 shows a probability density distribution of slc constructed directly using the members of the MME, which has a median value of 8.7 cm SLE and 17% and 83% quantiles of 4.1 and 18.9 cm; the latter being used to define the 66% credibility interval named “likely” following the IPCC terminology (Mastrandrea et al., 2010).

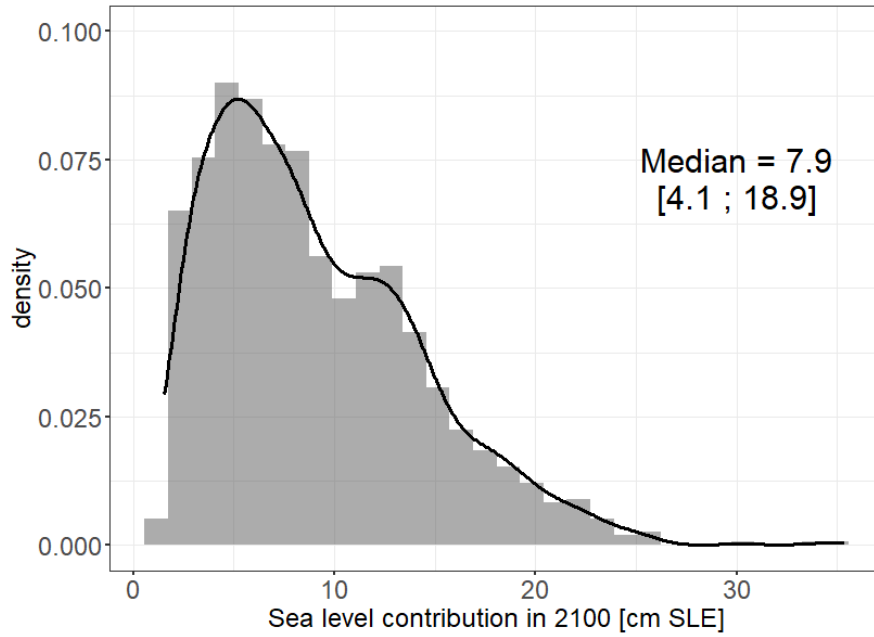
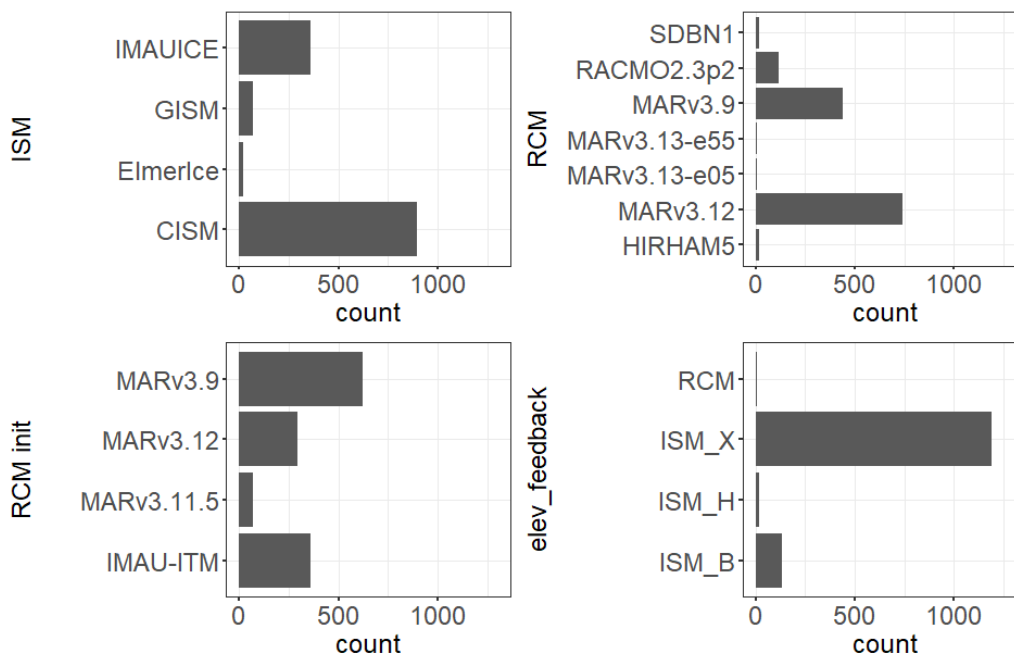


Figure 2: (a) Probability density function of the sea level contribution in 2100 (with respect to 2014) from the Greenland ice-sheet (in cm seal level equivalent, SLE) based on the raw MME ensemble data considered in this study. The black straight line provides the smoothed density function. The median value and the likely range are also indicated.

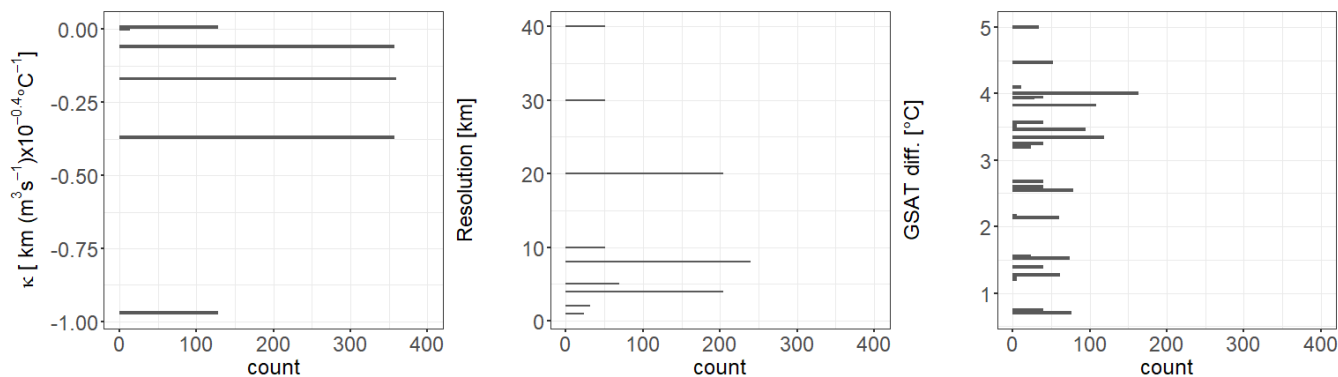
130

Figures 3 and 4 show the histograms for a selection of the continuous and categorical variables described in Table 1. For sake of space, we focus here on the 7 of 11 variables identified to have the highest importance with respect to *slc* (see Sect. 3 and Appendix C). Both Figures 3 and 4 show that the design of experiments is unbalanced: some categories (like *Elmer/Ice* model for instance for ISM in Fig. 3, top, left) or some values (like minimum resolution at 16km, Fig. 4, centre) are more frequent than others. The design is also incomplete with large gaps in the continuous class. This is for instance the case for κ between -0.9705 and -0.3700 $\text{km} \cdot (\text{m}^3 \cdot \text{s}^{-1})^{-0.4} \text{ } ^\circ\text{C}$ (Fig. 3, left), 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.

135



140 **Figure 3: Count number of the MME members with respect to the different inputs classified as “categorical” in Table 1: ISM (ice sheet model), RCM (regional climate model used for downscaling climate projections), RCM init (regional climate model used for initialisation climate), and elev_feedback (approach to representing the feedback between the ice sheet surface elevation and climate).**



145 **Figure 4: Count number of the MME members with respect to the different inputs classified as “continuous” in Table 1: κ (ice sheet tidewater glacier retreat parameter), minimum spatial resolution of the ice sheet model, and GSAT diff (global mean surface air temperature change during the driving global climate model simulation).**

150 **2.2 Setting up the emulator**

The objective is to predict *slc*, the sea level contribution at 2100 (with respect to 2014) from any values (configurations) of the different inputs (described in Table 1). The mathematical relationship between the inputs and *slc* is only partial and based on the MME results computed from the chain of numerical models described in Sect. 2.1. To overcome this difficulty, we replace the chain of numerical models by a machine-learning-based proxy (named emulator) built using the MME results. The advantage of using emulators is to make some predictions for input configurations that are not present in the original MME dataset at a low computation time cost. Among the different types of emulators (see e.g., Yoo et al. (2025) for a recent overview of different options), we focus in this study on the Random Forest (RF) regression model, as introduced by Breiman (2001). The interested reader can refer to Appendix A for further technical details. RF has shown high efficiency in diverse domains of application (sea level science, Tadesse et al. (2020); water resources, Tyrallis et al. (2019); flood assessments, Rohmer et al. (2018)), and more particularly for sea level projection studies (Hough & Wong, 2022; Rohmer et al., 2022; Turner et al., 2024). This emulator has the advantage of dealing, by construction, with different mixed types of input variables, categorical and continuous, which is a key aspect in our case (see Table 1).

The emulator remains however a statistical approximation in the sense that it uses only a limited number of numerical results, i.e. inputs-*slc* pairs (corresponding to the training data), to perform predictions given a "yet-unseen" inputs' configuration. Such an approximation introduces a new source of uncertainty referred to as "emulator uncertainty" as discussed by Storlie et al. (2009). To assess this type of uncertainty, we rely on the RF variant specifically developed by Meinshausen (2006) for predicting quantiles, i.e. the quantile RF model (qRF) as described in Appendix B. The advantage is that prediction intervals can be calculated at any level, which can be used to reflect the uncertainty of the RF emulator in emulator predictions.

In summary, the emulator provides a 'best estimate', corresponding to the conditional mean of the RF model, and prediction intervals constructed from the conditional quantiles of the qRF model. In what follows, we indifferently designate the emulator used as the "RF model".

2.3 Emulator experiments related to design questions

In this study, we address a series of questions described in Table 2 that are relevant for the design of MMEs. In general, the central concern is to investigate what is the added value of including a specific set of experiments in the projections. This could be subsets in already defined value range / categories, or subsets not currently categorised. For four different categories of inputs related to specific modelling choices (choice in SSP-RCP, choice in RCM, choice in ISM, and range of κ values), the design questions are formalised in Table 2. To assess the added value of including a specific set of experiments in the projections, we propose to construct RF emulators by leaving out specific results from the original MME without differentiating the members, i.e., by assuming that all members have the same weight in the ensemble. The last column of Table 2 translates the design questions into a specific emulator's experiment. Using a RF emulator trained with the complete original MME as a reference solution, we assess changes in two types of criteria: ~~changes in the MME characteristics,~~

performance of the RF emulator, and the probability estimates of *slc* in 2100 given future GSAT change scenarios, here chosen at 2°C (+/- 0.5°C), ~~3°C~~ or 4 °C (with a tolerance of +/- 0.5°C). The details of this assessment are explained in Sect. 2.4.

185 **Table 2: Design questions and corresponding emulator’s experiments. Modelling choices are evaluated based on the RF emulator performance and the probability estimate of *slc* in 2100 given GSAT at 2, or 4°C (+/- 0.5°C).**

Input	Question	Definition of the emulator’s experiment	Name of the experiment	Number of members*
SSP-RCP scenario	Does including a medium scenario SSP2-4.5 improve the results or is it enough to use the extreme scenarios SSP1-2.6 and SSP5-8.5?	A RF emulator is trained using only the results for SSP1-2.6 & SSP2-4.5, i.e. without SSP5-8.5	Without SSP5-8.5: ‘woSSP585’	418 (31%);
		SSP1-2.6 & SSP5-8.5, without SSP2-4.5;	‘woSSP245’	1,155 (86%)
		SSP2-4.5 & SSP5-8.5, without SSP1-2.6	‘woSSP126’	1,113 (83%)
RCM choice	What is the added value of including a new RCM, i.e. is it sufficient to focus on MAR regional climate model (Fettweis et al., 2017) only?	A RF emulator is built using only the results for MAR (regardless of the version: MARv3.12, MARv3.13-e05, MARv3.13-e55, or MARv3.9).	‘MAR’	1,197 (89%)
		Built using only the results for Regional Atmospheric Climate Model RACMO (Ettema et al., 2010), HIRHAM (Mottram et al., 2017), and the direct statistical downscaling of the GCM CESM2-WACCM (SDBN1).	‘woMAR’	146 (11%)
ISM choice	What is the added value of accounting for all ISM except for one?	A RF emulator is trained using only the results for the most selected ISM, namely the Community Ice Sheet Model (CISM; Lipscomb et al., 2019)	‘CISM’	894 (66.5%)
		Built without the results of CISM (experiment ‘woCISM’).	‘woCISM’	449 (33.5%)
Range of κ values	Should the design cover a large range of values, i.e. is it sufficient to focus on extreme values?	A RF emulator is built using the central value of -0.1700 and the endpoints, of -0.9705 and 0.007 km.(m ³ .s ⁻¹) ^{-0.4} °C only, i.e. without intermediate values.	‘Med. & Extr.’	615 (46%)
		Built only with central and medium values, from -0.37 to 0 km.(m ³ .s ⁻¹) ^{-0.4} °C.	‘Narrow’	1.087 (81%)

*% of the total number of members

2.4 Criteria for measuring the impact of the design questions

2.4.1 Emulator performance

The first criterion measures the decrease in the predictive performance of the emulator. It is assessed through a validation test exercise that consists in randomly selecting n_{test} test samples from the original MME, conducting the experiments described in Table 2, and estimating the *slc* error. In this study, we are more particularly interested in the ability of the emulator to perform well over a wide range of GSAT values. This is important in our case, because constraining the predictions to temperature constraints can help end-users to interpret the projections as illustrated by recent projections for France by Le Cozannet et al. (2025), although it should be noted that our GSAT definition does not strictly correspond to the global warming level (GWL) defined in AR6.

Therefore, instead of relying on the widely used cross validation procedure (Hastie et al., 2009), we propose an alternative validation procedure adapted to our objective as follows: (1) the GSATs are classified into a finite number of intervals, the ends of which are defined by the GSAT percentiles, with levels ranging from 0 to 100% with a fixed increase of 25%. This results in the following GSAT intervals, [0.705, 2.14°C], [2.14, 3.34°C], [3.34, 3.83°C], and [3.83, 5.00°C]; (2) for each interval, 50 samples are randomly selected. For one iteration of the procedure, a total of $n_{\text{test}}=200$ test samples are randomly selected. The procedure is repeated 25 times. We then compute three performance criteria (formally described in Appendix D), namely:

- the mean relative error, *RAE* and the coefficient of determination, Q^2 that both measures whether the RF emulator can predict *slc* with high accuracy given yet-unseen instances of the inputs. A high predictive capability is achieved for a *RAE* value close to zero and for a Q^2 value close to one;
- the continuous ranked probability score, denoted *CRPS*, as used for validating probabilistic weather forecast (Gneiting et al., 2005), that jointly quantifies the calibration of qRF probability distribution, i.e. the reliability of the estimation, and its sharpness (i.e. the concentration/dispersion of the probability distribution). The lower *CRPS*, the higher the quality of the qRF probabilistic predictions, with a lower limit of zero.

2.4.2 Emulator-based probabilistic predictions for a GSAT scenario

The second set of criteria measures the changes in the emulator-based probabilistic predictions, which are assessed through a Monte-Carlo random sampling procedure. For fixed GSAT change values (here chosen at 2°C and 4°C, with a tolerance of +/- 0.5°C), the input variables are randomly sampled by assuming a uniform discrete probability distribution for the categorical variables, and a uniform probability distribution for the continuous variables except for κ which is sampled as in (Edwards et al. 2021) from the smoothed version of the empirical density function by Slater et al. (2019). In addition, the emulator uncertainty is propagated by following the procedure based on the quantile RF emulator (Appendix B).

The emulator-based probabilistic results thus jointly reflect the impact of the uncertainty of the input variables and of the emulator uncertainty. The probabilistic predictions should however not be interpreted as calibrated uncertainty accounting for

model-observation misfits (e.g., Aschwanden and Brinkerhoff, 2022), and neither do they represent the *slc* probability distribution from the MME, because the uniform distribution over the input space is not representative of the MME itself. To measure the impact of the different design decisions on these probabilistic predictions, the second set of criteria corresponds to the changes in the median and to the endpoints of the 66% credibility interval, named ‘likely range’ following the IPCC terminology, defined here by the percentile at 17 and 83%, denoted Q17% and Q83%.

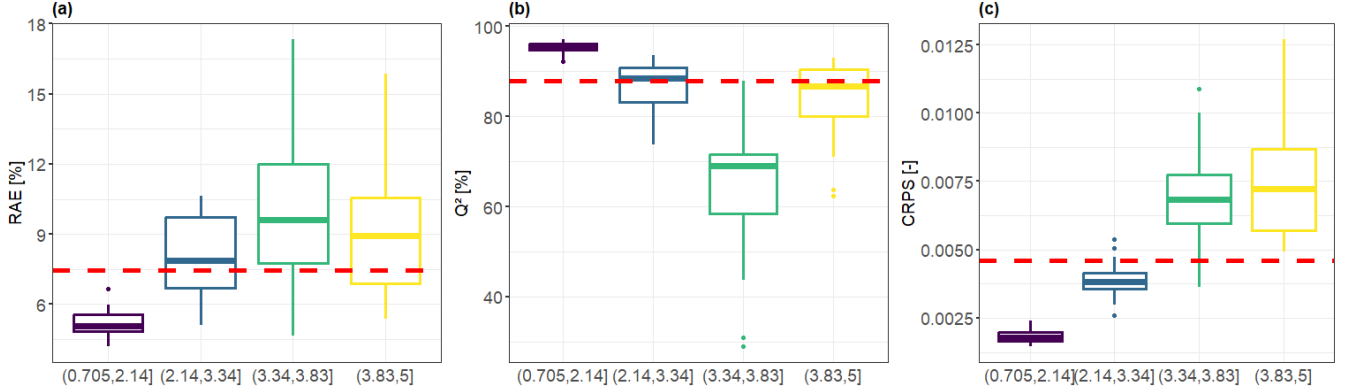


Figure 5: Boxplot of the RAE (a), Q^2 (b) and CRPS (c) performance indicator for different ranges of GSAT (indicated on the x-axis). The lower RAE and the closer Q^2 to one, the higher the emulator predictive capability. The lower CRPS, the higher quality of the emulator predictive probabilistic distribution. The horizontal red dashed line indicates the median value calculated over all validation tests defined through the repeated validation procedure described in Sect. 2.4.1 considering the whole range of GSAT.

3. Results

3.1 Emulator reference solution

We train a RF emulator to predict *slc* in 2100 using the results of the GrIS MME (see implementation details in Appendix A). A preliminary screening analysis was conducted (detailed in Appendix C), and showed that four predictor variables have no significant influence: the choice to account for thermodynamics, the choice in sliding law, the type of initialisation and the number of years for the initialisation phase. We therefore build the RF emulator using only 7 out of 11 possible input variables described in Sect. 2.

On this basis, we compute the reference solution for the criteria used to investigate the influence of the design questions. First, the RF model’s predictive performance is tested by applying the repeated validation procedure described in Sect. 2.4.1. The performance of the RF emulator shows satisfactory levels of predictability, with a median *RAE* value (calculated over all the validation tests defined by the repeated validation procedure) of no more than 8%, a median Q^2 value close to 90% and a median *CRPS* value close to zero as indicated by the dashed red horizontal line in Fig. 5. The examination of the performance depending on the GSAT interval of the test samples (coloured boxplots in Fig. 5) further shows that the highest performance is achieved for low GSAT below 2.14°C (dark blue boxplots in Fig. 5), and the worst performance is achieved for GSAT

between 3.34 and 3.83°C (green boxplot in Fig. 5). The performance of the other GSAT intervals, and in particular for the highest GSAT values above 3.83°C, can be considered satisfactory with a median *RAE* not larger than 9%, a median *Q*² value close to 90% and a median *CRPS* relatively low.

Second, the probability distribution of *slc* in 2100 relative to 2014 (Figure 6) is constructed using the Monte-Carlo-based procedure (with 10,000 random samples) described in Sect. 2.4.2 given GSAT change values fixed at 2°C and 4°C (+/-0.5°C). The choice of GSAT scenarios used here is supported by the afore-described analysis, which points out that the RF emulator should be used cautiously over the range of GSAT values around 3°C. The emulator-based probabilistic prediction results in a median value of respectively 6.1cm, and 13.5cm for *slc* with a likely range of [4.6; 7.4cm], and [10.4; 17.0cm].

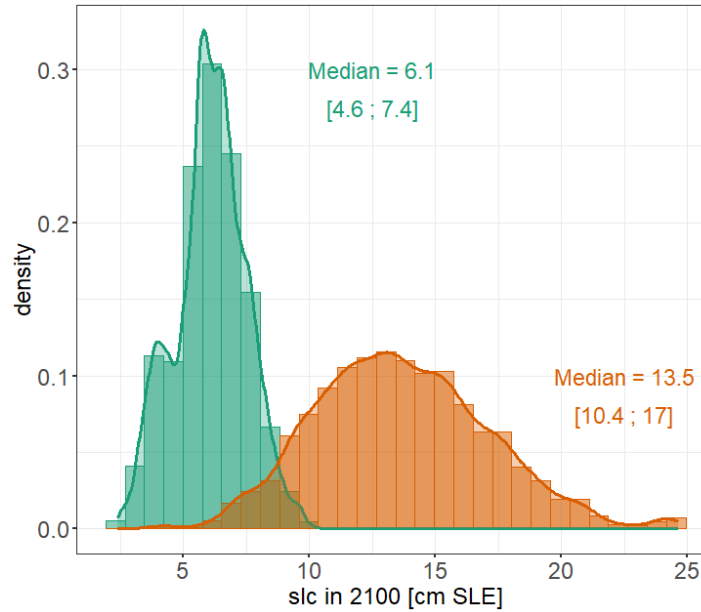


Figure 6: Emulator-based probabilistic predictions in the form of probability density function of *slc* in 2100 (with respect to 2014) constructed the Monte-Carlo-based procedure (with 10,000 random samples, see Sect. 2.4) for two GSAT change values of 2°C+/-0.5°C (green), and 4°C+/-0.5°C (orange). This results in a median value of respectively 6.7cm, and 13.5cm with a likely range of [4.6; 7.4cm], and [10.5; 17cm]. The straight line corresponds to the smoothed density function. The number and interval indicate the median value and the likely range. Note these probability density functions are derived using the conditional mean of the RF emulator (Appendix A) and do not include uncertainty arising from the emulator itself.

3.2 Impact of design decisions on the emulator performance

We analyse in Figure 7 the impact of design decisions with respect to the decrease of the RF predictive capability (measured by decrease of the relative differences of *RAE* and *CRPS* and the increase of the relative differences of *Q*²).

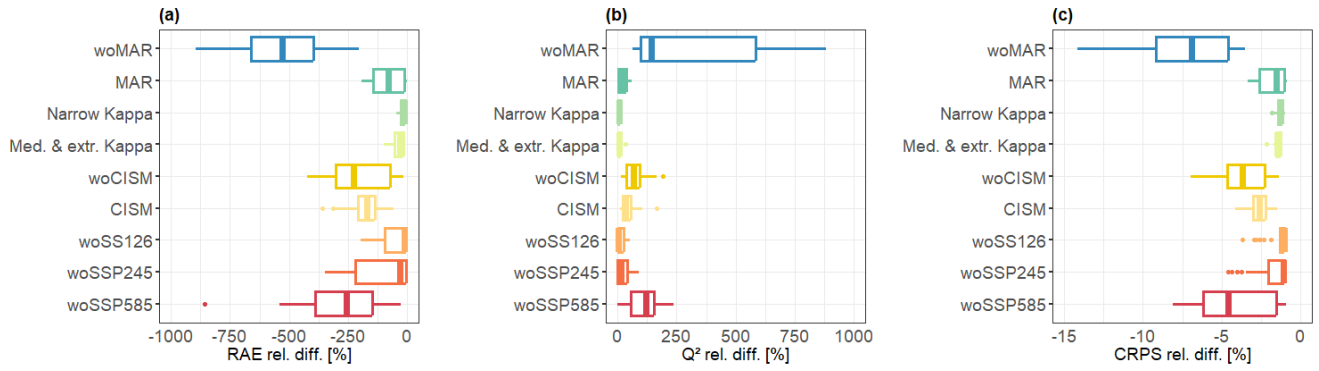
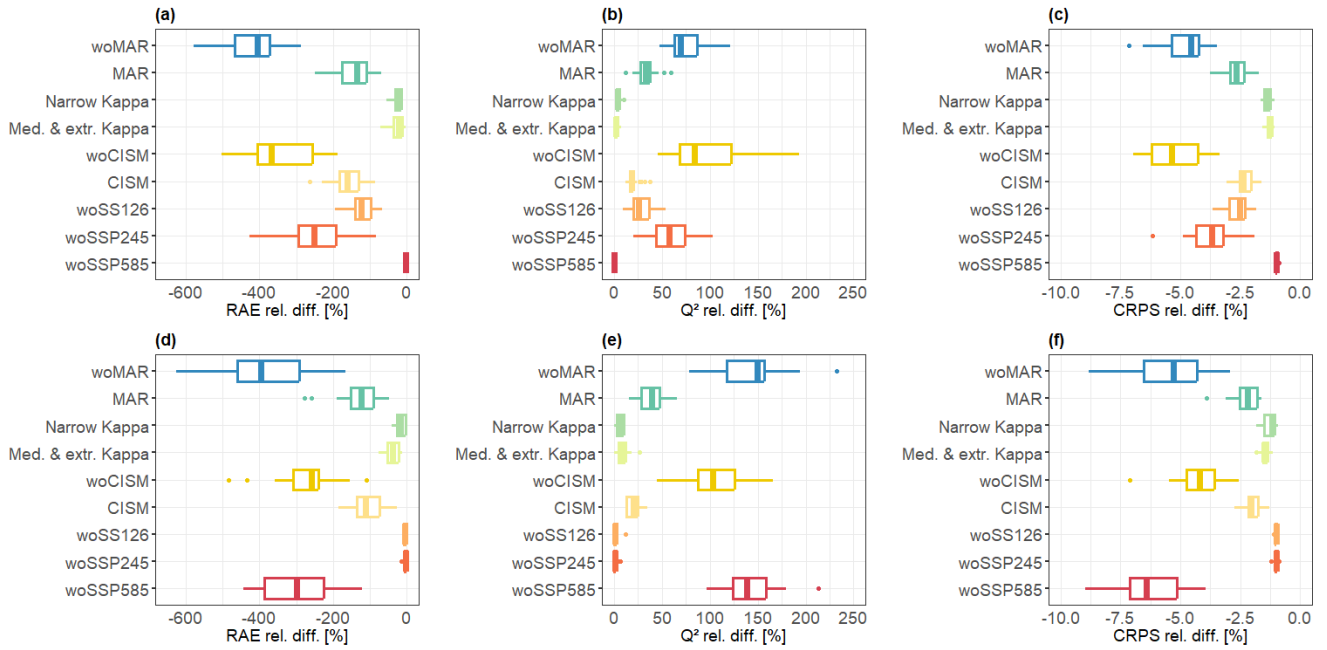


Figure 7: Relative difference (in %) of the performance criteria for RAE (a), Q^2 (b), and CRPS (c) calculated by applying the validation procedure described in Sect. 2.4.1 repeated 25 times.

Figure 7 shows that excluding MAR (experiment ‘woMAR’) has the largest impact whatever the performance criteria. This is also shown when considering a given GSAT interval in the validation procedure (Fig. 8 and Supplementary Materials S1).

This means that excluding MAR impacts both facets of the predictive capability of the emulator, i.e., the explained variance of the emulator Q^2 and the relative errors RAE. We also show that the quality of the prediction intervals, measured by the CRPS criterion, is also impacted over the whole range of GSAT values considered in this study. This result was expected due to the large decrease by almost 90% of the MME size (Table 1). Interestingly, restricting the analysis to one RCM model, here MAR, has a non-negligible impact. This suggests that integrating different RCMs in the MME is beneficial for the emulator performance although it should be recognised that the influence is by far of lower magnitude than that of the ‘woMAR’ experiment. The second most important driver of the emulator performance is the exclusion of extreme SSP scenario SSP5-8.5, i.e., experiment ‘woSSP585’ (red boxplot in Fig. 7) which induces a performance reduction of around half that of ‘woMAR’ and twice that of the third most important contributor, i.e., ‘woCISM’.



280 **Figure 8: Relative difference (in %) of the performance criteria considering the lowest GSAT values below 2.14°C (top) and the highest GSAT values above 3.83°C (bottom) for RAE (a, d), Q^2 (b, e), and CRPS (c, f).**

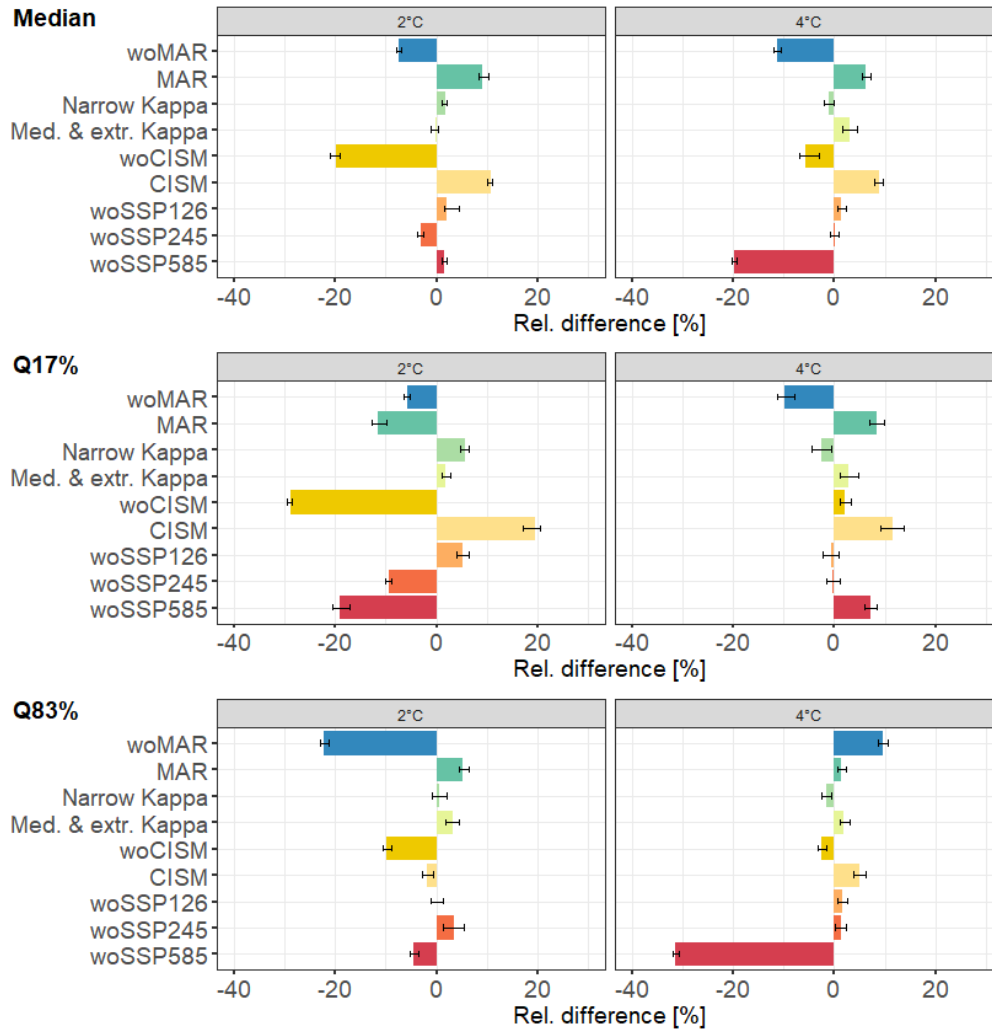
The analysis of Fig. 8 shows however that this high influence depends on the range of GSAT considered. On the one hand, for the highest GSAT values (Fig. 8, bottom), excluding the extreme SSP scenario SSP5-8.5 (experiment 'woSSP585') or excluding CISM (experiment 'woCISM') has the largest impact almost at the same level than that of 'woMAR' experiment; in particular in terms of Q^2 (Fig. 8e) and CRPS (Fig. 8f) relative differences. The almost similar decrease in the MME size of almost 70% explains this result (Table 1). On the other hand, for the lowest GSAT values (Fig. 8, top), excluding CISM has here the second largest impact on the performance. Again, it is interesting to note that restricting the analysis to a unique ISM model, here CISM, has a non-negligible impact on the emulator performance. The analysis of the other GSAT intervals (Supplementary Materials S1) even shows that the impact of this experiment can be as high as that of 'woSSP585'. For the lowest range of GSAT value, it is 'woSSP245', and to a lesser extent 'woSSP126' as well, that drives the most the performance decrease. This agrees well with the range of global warming levels that the simulations associated to these SSP scenarios cover. Finally, the experiments for κ appear not to affect much the performance; both experiments having the lowest influence.

3.3 Impact of design decisions on the emulator-based probabilistic predictions

295 In this section, we analyse the impact on the RF-based probabilistic predictions. Since the impact on the percentiles has more interest from the perspective of end-users, we primarily focus the analysis on the changes in the *slc* percentiles, Q17%, Q83% and in the median value in Fig. 9. The interested reader can refer to Supplementary materials S2 for an analysis of the whole

s/c probability distributions' changes. Here the results include estimates of uncertainty arising from the emulator itself. Figure 9 shows that, depending on the GSAT scenario, the percentiles are perturbed in different ways. Several observations can be made:

- Overall, the design decision for κ has only a minor impact on the order of the emulator uncertainty indicated by the width of the error-bars, regardless of the GSAT change and the considered percentile. This result agrees well with the analysis on the RF predictive capability in Sect. 3.2;
- Excluding MAR has a large impact as expected from Sect. 3.2 leading to absolute changes of the different percentile on the order of 10%. The largest influence is for the high percentile Q83%, which is under-estimated by more than 25% for the lowest GSAT scenario;
- Interestingly, it is not the most important driver of the emulator performance, i.e., experiment 'woMAR', that influences the most the considered percentiles, but the second drivers of the emulator performance, namely experiment 'woCISM' and 'woSSP585'. This is shown by the high under-estimation of the median on the order of -20% when considering 'woCISM' for the lowest GSAT scenario. This is also the case when considering 'woSSP585' for the highest GSAT scenario;
- The highest decrease of all tests is -30% for Q17% when considering 'woCISM' for the lowest GSAT scenario, and for Q83% when considering 'woSSP585' for the highest GSAT scenario. These results agree well with the analysis of the emulator performance (Fig. 8);
- Similarly as for the performance analysis, including in the MME a unique ISM, here CISM, or a unique RCM, here MAR, has a non-negligible influence leading to absolute changes on the same order of 10-15% (in average).



320 **Figure 9: Relative difference (in %) between the RF reference solution and the RF model trained when considering the experiments indicated in the y-axis (see Table 2 for full details) for the estimates of three *slc* percentiles in 2100 relative to 2014, the median and the quantile at 17% (Q17%) and at 83% (Q83%), considering two GSAT changes, 2°C (+/-0.5°C), and 4°C (+/-0.5°C). The endpoints of the error-bars correspond to the 5% and the 95% quantile calculated by applying 100 times the procedure described in Appendix B to reflect the emulator uncertainty.**

4. Synthesis and Discussion

4.1 Implications for MME design

325 Table 3 summarises the main results from the emulator’s experiments for each design question considering the MME of this study. In the following, we take the viewpoint of a MME designer, and derive the practical recommendations from these results.

Table 3. Summary of the results from the emulator’s experiments for each design question considering the MME of this study.

Input	Question	Results for the considered MME
SSP-RCP	Does including a medium scenario SSP2-4.5 improve the results or is it enough to use the end members SSP1-2.6 and SSP5-8.5?	Excluding the medium scenario has a small-to-moderate impact, and mainly affects the emulator performance for low GSAT values. The main reason is likely to be the importance of the end member scenarios, as shown by the experiment excluding SSP5-8.5, but also the relatively small number of medium scenario simulations in this ensemble and the strong linearity of the Greenland ice sheet response with global temperature in these models (other models could be more non-linear).
RCM choice	What is the added value of including new RCM, i.e. is it sufficient to focus on MAR regional climate model only?	This is the most impactful decision, whatever the GSAT scenario or the criteria, i.e., emulator performance or percentile assessment. This result is strongly linked to the high number of members using MAR.
ISM choice	What is the added value of accounting for all ISMs except for one?	Restricting the analysis to one unique ISM, here CISM, might lead to a moderate over-estimation of the probabilistic predictions. In addition, having different ISMs is beneficial (experiment ‘woCISM’), but not sufficient with high impact on the emulator performance and under-estimation of the percentile values.
Range of κ values	Should the design cover a large range of values, i.e. is it sufficient to focus on extreme values?	This decision is the least impactful relatively to the others in the considered MME. Results suggest that restricting to the ‘ <i>Medium and Extreme</i> ’ scenario is sufficient for this parameter, which has a relatively linear relationship with ice sheet response, though for other parameters this would not be known <i>a priori</i> and test simulations would be needed for multiple values to characterise whether the relationship was non-linear.

330

335

On the one hand, some conclusions were expected beforehand, namely the influence of emulator experiments leading to high decrease in the MME size between $\approx 70\%$ and $\approx 90\%$, i.e., ‘woMAR’, ‘woSSP585’, and ‘woCISM’. This decrease logically degrades the predictive capability since the RF is trained on a small dataset (Sect. 3.2). On the other hand, some other conclusions could not necessarily have been anticipated in detail more particularly the implications on the percentile assessment (Sect. 3.3). Here, ‘woMAR’ is not necessarily the highest contributor to the changes although it leads to the highest size decrease of nearly 90%. Depending on the GSAT scenario considered and on the targeted level of the percentile, the inclusion of SSP5-8.5 or that of CISM has strong implications although the induced decrease in MME size is lower on the

order of 70%. This suggests that this is not only a problem of ‘size’ but also a problem of type of information that is removed from the ensemble depending on the targeted objective, i.e., emulator performance or percentile assessment. As a complement, we analysed in Supplementary Materials S3 an indicator that measures the changes in the member distributions of the new MME subsets after application of the emulator experiment. This shows that the ‘woCISM’ experiment leads to the largest perturbation of the member distribution, i.e., by >50% largest that of ‘woMAR’. This suggests that the additional important factor is also the diversity among the members within the MME. This result helps to explain the greater contribution of ‘woCISM’ to the percentile (Sect. 3.3), although it must be recognised that the indicator introduced in Supplementary Materials S3 remains a first-order approximation of this diversity, and that further work should look into this aspect in more detail.

A very practical implication can be derived from the κ experiments: results indicate that restricting to the extreme and medium scenario is sufficient here because of the lesser impact between the two experiments, ‘*Med. & Extr.*’ or “*Narrow*”. This result is interpreted as being linked to a quasi-linear relationship between κ and *slc* as shown in Rohmer et al. (2022) using the MME of ISMIP6 for Greenland. This was confirmed by the analysis detailed in Supplementary Materials S3. In practice, this result implies that the number of scenarios explored in the MME can be limited to a three-scenario approach (low-medium-high value), i.e. the number of members can be reduced, thus reducing the number of long numerical simulations required.

Results for RCM and ISM choice can be seen as an additional justification for intensifying the model intercomparison efforts initiated in the past. They also support, to some extent, a posteriori, the choices that have been made for the construction of the MME considered here (based on that of Goelzer et al. (2020)). Restricting the analysis to one unique ISM, here CISM, leads to a clear over-estimation of the percentiles, which suggests that having a diversity of ISMs is ~~here~~ beneficial. This is in line with the initiative originally launched in ISMIP6 (Nowicki et al., 2016), which included coupled ISMs as well as stand-alone ISMs in CMIP for the first time. Similarly, the choice of restricting the MME to MAR is impactful for the emulator performance and the percentile assessment. This calls for intensifying the cooperative research efforts, potentially within a MIP, by extending this study to different RCM models, instead of MAR only, or investigating the relevance of using different versions of MAR (see Table 2). This also relates to the question of initialisation (and initial mass loss estimates) where the RCM choice is a key ingredient (e.g., Otsuka et al., 2023).

4.2 Implications from stakeholders’ point of view

Our work can help stakeholders in several ways. First, our study contributes to a better understanding of the contribution of Greenland ice sheet melt to sea level rise. According to the latest authoritative sea level projections developed by the IPCC (Fox-Kemper et al., 2022) the GIS contribution to sea-level rise is projected to reach 8cm [4cm; 13cm] (median [likely range]) by 2100 for the SSP2-4.5 scenario. This means Greenland has a sizeable share to the total global mean sea level rise and their uncertainties, which were estimated at 56cm [44cm; 76cm] for this scenario according to the same report. Second, our results support coastal adaptation practitioners in their decision-making. Our emulator experiments in Sect. 3.2 and 3.3 highlight how the different modelling choices affect differently the median or the upper tail (here measured by the Q83% percentile). This difference is important, because the literature on adaptation decision-making has clearly shown that knowing the median is

not sufficient for coastal adaptation practitioners managing long-living critical infrastructures or making strategic decisions for regions or countries (Hinkel et al., 2019). These practitioners need credible assessments of the uncertainties in ice mass losses in Greenland, including for the low probability scenarios corresponding to the tail of probabilistic projection. For example, France selected a unique climate scenario of 3°C GWL used in France within its 3rd development plan published in 2025. To define the associated sea-level scenarios to be mainstreamed in public policies, a detailed consideration of uncertainties is required to understand which security margins are taken (Le Cozannet et al., 2025). Thus, our study supports the need for improved experimental designs by making some practical recommendations, especially regarding the consideration of ISM, RCM and RCP8.5/SSP5-8.5 simulations.

Finally, the importance of SSP5-8.5, although expected, also underlines the fact that a wide range of emissions scenarios and climate simulations should continue to be considered in the future. The SSP5.8-5 scenario in this ensemble contains many simulations and covers a wide range of global warming levels at 2100. To represent plausible outcomes of failure of states to meet their own commitments, or political backlashes leading to climate policy setbacks (see recent discussion by Meinshausen et al., 2024), medium and medium-high emissions scenarios (e.g. radiative forcing reaching between 4.5 and 7.0W/m² in 2100) should continue to be used for simulations of climate impacts such as for the Greenland ice sheet, so that these do not rely too much on emulators interpolating from end member scenarios. Furthermore, the current design of the SSP3-7.0 involves very high aerosol emissions, so that the resulting simulations need to be considered carefully (Shiogama et al., 2023). Being able to use more intermediate climate simulations reaching radiative forcing between 4.5 and 7.0W/m² in 2100 is all the more important as another need is now emerging: projections of ice mass loss for specific levels of global warming relative to preindustrial (as in the IPCC: Fox-Kemper et al., 2022; see also the latest adaptation plan in France, Le Cozannet et al., 2025). Indeed, scenarios based on global warming levels can be potentially better understood by stakeholders than the SSP or RCP scenarios, and also allow users to better make the link with the climate objectives set out in the Paris agreement to stabilize climate change well below 2°C GWL. For all scenarios, including global warming levels, the development of probabilistic projections requires emulators, whose accuracy and precision can be improved by better experimental design.

5. Concluding remarks and further work

Developing robust protocols to design balanced and complete numerical experiments for MME is a matter of active research that has called multiple studies either for sea level projections via selection criteria (Barthel et al., 2020) or from an uncertainty assessment's perspective (Aschwanden et al., 2021), and more generally for regional impact assessment (Merrifield et al., 2023; Evin et al., 2019). In this study, we take advantage of a large MME produced for Greenland ice sheet contributions to future sea level to define a series of emulator's experiments that are closely related to practical MME design decisions. Our results confirm the high importance of including the SSP5-8.5 scenario as well as having diverse ISM and RCM models. Finally, the less impactful choice in this ensemble is the one in the sampling of the Greenland tidewater glacier retreat

parameter, because it has a relatively linear relationship with sea level contribution. These recommendations (detailed in Table 3) can be informative for the design of next generation MME for Greenland (ISMIP7: Nowicki et al., 2023).

Although the MME considered in this study covers a large spectrum of situations (multiple SSP scenarios, different RCMs and ISMs), with more than 1,000 members, a series of aspects need to be considered in the future to further increase the robustness of these results. First, our procedure should be tested on additional MMEs of interest to improve the transferability of our results, in particular for Antarctica (Seroussi et al., 2020), for multi-centennial projections (e.g., Seroussi et al., 2024), and for glaciers (Marzeion et al., 2020). These tests should also include new types of MMEs that are combined with calibration (e.g., Aschwanden and Brinkerhoff, 2022). They make it possible to circumvent an assumption in our study, namely that all members have the same weight, by taking into account the reliability of the different members (for example, low-resolution models compared with high-resolution models) or observational constraints, provided that good-quality data are available over a sufficiently long period in the past and that the numerical implementation of the ISMs is suitable for calibration. Second, our results are based on the use of an emulator, i.e., a statistical approximation of the ‘true’ chain of numerical models. The efforts made to nuance the results by including indicators of the emulator uncertainty should be strengthened in the future by considering multiple types of emulator models (e.g., Yoo et al., 2025) but also the impact of hyperparameter tuning (Bischl et al., 2023). Finally, our recommendations are derived, by construction, a posteriori, i.e., based on the available members of a large-size MME. Therefore, a third avenue here is to derive recommendations earlier on in the process, i.e. early during the construction of the MME design, through an iterative manner between phases of simulations (either test simulations to evaluate sensitivity to different inputs, or small exploratory ensembles that do not use the full computing/person/project time available) and emulator training – and re-training – with each phase, to identify key gaps (e.g. additional parameter interaction tests in Edwards et al., 2021). From a methodological perspective, robust tools may be found in the data valuation domain (Sim et al., 2022), which aims to study the worth of data in machine learning models based on similar methods as the ones used by Rohmer et al. (2022) in the context of sea level projections. Transposed to the MME context, these tools could be used in future studies to assess the impact of each member in the emulator’s predictions, i.e. the worth of each member. This type of result is expected to serve as guidance to the MME design in particular regarding the question of completeness and the necessity for balanced design.

Author contributions

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

Competing interests

The authors declare that they have no conflict of interest.

Code/Data availability

We provide the data and R scripts to run the experiments and analysing the results on the Github repository:

435 <https://github.com/rohmerj/MMEdesign>

Acknowledgements

We acknowledge the modelling work that constitutes the MME analysed in this study: the PROTECT Greenland ice sheet modelling and regional climate modelling groups, and the World Climate Research Programme and its Working Group on Coupled Modelling for coordinating and promoting CMIP5 and CMIP6. We thank the modelling groups for producing and
440 making available their model output and the Earth System Grid Federation (ESGF) for archiving the CMIP data and providing access. HG has received funding from the Research Council of Norway under project 324639 and had access to resources provided by Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway through projects NN8006K, NN8085K, NS8006K, NS8085K and NS5011K. This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 869304, PROTECT. The authors would
445 like to acknowledge the assistance of DeepL (<https://www.deepl.com/fr/translator>) in refining the language and grammar of this manuscript.

References

- Altmann, A., Tolosi, L., Sander, O. and Lengauer, T.: Permutation importance: a corrected feature importance measure, *Bioinformatics*, 26, 1340-1347, 2010.
- 450 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.
- Aschwanden, A., and Brinkerhoff, D. J.: Calibrated mass loss predictions for the Greenland ice sheet. *Geophysical Research Letters*, 49(19):e2022GL099058, 2022.
- Barthel, A., Agosta, C., Little, C. M., Hattermann, T., Jourdain, N. C., Goelzer, H., Nowicki, S., Seroussi, H., Straneo, F., and
- 455 Bracegirdle, T. J.: CMIP5 model selection for ISMIP6 ice sheet model forcing: Greenland and Antarctica, *Cryosphere*, 14, 855-879, doi:10.5194/tc-14-855-2020, 2020.
- Bergmeir, C. and Benítez, J. M.: On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191, 192-213, 2012.
- Bischi, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L.,
- 460 Deng, D., Lindauer, M.: Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2), e1484, 2023.
- Breiman, L.: Random forests, *Mach. Learn.*, 45, 5–32, 2001.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J.: *Classification and regression trees*, Wadsworth, California, 1984.
- Edwards, T. L., Nowicki, S., Marzeion, B., Hock, R., Goelzer, H., Seroussi, H., et al.: Projected land ice contributions to
- 465 twenty-first-century sea level rise, *Nature*, 593(7857), 74-82, 2021.
- Ettema, J., van den Broeke, M. R., van Meijgaard, E., van de Berg, W. J., Box, J. E., and Steffen, K.: Climate of the Greenland ice sheet using a high-resolution climate model – Part 1: Evaluation, *The Cryosphere*, 4, 511–527, doi:10.5194/tc-4-511-2010, 2010.
- Evin, G., Hingray, B., Blanchet, J., Eckert, N., Morin, S., and Verfaillie, D.: Partitioning uncertainty components of an
- 470 incomplete ensemble of climate projections using data augmentation, *Journal of Climate*, 32(8), 2423-2440, 2019.
- Fettweis, X., Box, J. E., Agosta, C., Amory, C., Kittel, C., Lang, C., van As, D., Machguth, H., and Gallée, H.: Reconstructions of the 1900–2015 Greenland ice sheet surface mass balance using the regional climate MAR model, *The Cryosphere*, 11, 1015–1033, doi:10.5194/tc-11-1015-2017, 2017.
- ~~Friedman, J. H.: Greedy function approximation: A gradient boosting machine, *Annals of statistics*, 29(5), 1189–1232, doi:10.1214/aos/1013203451, 2001.~~
- 475 ~~doi:10.1214/aos/1013203451, 2001.~~
- Fox-Kemper, B., Hewitt, H.T., Xiao, C., Aðalgeirsdóttir, G., Drijfhout, S. S., Edwards, T. L., Golledge, N. R., Hemer, M., Kopp, R. E., Krinner, G., Mix, A., Notz, D., Nowicki, S., Nurhati, I. S., Ruiz, L., Sallée, J.-B., Slangen, A. B. A., and Yu, Y.: Ocean, Cryosphere and Sea Level Change. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* [Masson-Delmotte, V., P. Zhai,

- 480 A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)], Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1211–1362, doi:10.1017/9781009157896.011, 2021.
- Gneiting T., Raftery A. E., Westveld III A. H., and Goldman T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review* 133(5): 1098–1118, 2005.
- 485 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-](https://doi.org/10.5194/tc-12-1433-2018)
- 490 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.
- Goelzer, H., Berends, C. J., Boberg, F., van den Broeke, M., Durand, G., Edwards, T., Fettweis, X., Gillet-Chaulet, F., Glaude, Q., Huybrechts, P., Le clec'h, S., Mottram, R., Noel, B., Olesen, M., Rahlves, C., Rohmer, J., van de Wal, R. S. W.: Extending
- 495 the range and reach of physically-based Greenland ice sheet sea-level projections. Preprint egusphere-2025-3098, 2025.
- Hastie, T., Tibshirani, R., and Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer: Berlin/Heidelberg, Germany, 2009.
- Hinkel, J., Church, J.A., Gregory, J.M., Lambert, E., Le Cozannet, G., Lowe, J., McInnes, K.L., Nicholls, R.J., van Der Pol, T.D. and Van De Wal, R.: Meeting user needs for sea level rise information: A decision analysis perspective, *Earth's Future*,
- 500 7(3), 320-337, 2019.
- Hough, A. and Wong, T.E.: Analysis of the evolution of parametric drivers of high-end sea-level hazards, *Advances in Statistical Climatology, Meteorology and Oceanography*, 8(1), 117-134, 2022.
- Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G.: Challenges in combining projections from multiple climate models, *J. Climate*, 23, 2739–2758, 2010.
- 505 Le Cozannet, G., Thiéblemont, R., Rohmer, J., and Capderrey, C.: Sea-level scenarios aligned with the 3rd adaptation plan in France. *Comptes Rendus. Géoscience*, 357(G1), 105-123, 2025.
- Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J. P., Engelbrecht, F., Fischer, E., Fyfe, J. C., Jones, C., Maycock, A., Mutemi, J., Ndiaye, O., Panickal, S., and Zhou, T.: Future Global Climate: Scenario-Based Projections and NearTerm Information. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment*
- 510 *Report of the Intergovernmental Panel on Climate Change* [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 553–672, doi:10.1017/9781009157896.006, 2021.

- Levermann, A., Winkelmann, R., Albrecht, T., Goelzer, H., Golledge, N. R., Greve, R., Huybrechts, P., Jordan, J., Leguy, G.,
515 Martin, D., et al.: Projecting Antarctica's contribution to future sea level rise from basal ice shelf melt using linear response
functions of 16 ice sheet 600 models (LARMIP-2), *Earth System Dynamics*, 11, 35–76, <https://doi.org/10.1175/JCLI-D-23-0580.1>, 2020.
- Lipscomb, W. H., Price, S. F., Hoffman, M. J., Leguy, G. R., Bennett, A. R., Bradley, S. L., Evans, K. J., Fyke, J. G., Kennedy,
J. H., Perego, M., Ranken, D. M., Sacks, W. J., Salinger, A. G., Vargo, L. J., and Worley, P. H.: Description and evaluation of
520 the Community Ice Sheet Model (CISM) v2.1, *Geosci. Model Dev.*, 12, 387–424, doi:10.5194/gmd-12-387-2019, 2019.
- Mastrandrea, M. D., Field, C. B., Stocker, T. F., Edenhofer, O., Ebi, K. L., Frame, D. J., Held, H., Kriegler, E., Mach, K. J.,
Matschoss, P. R., Plattner, G.-K., Yohe, G. W., and Zwiers, A. F. W.: Guidance note for lead authors of the IPCC fifth
assessment report on consistent treatment of uncertainties. Intergovernmental Panel on Climate Change (IPCC), 2010.
Available at <https://www.ipcc.ch>
- 525 Marzeion, B., Hock, R., Anderson, B., Bliss, A., Champollion, N., Fujita, K., et al.: Partitioning the uncertainty of ensemble
projections of global glacier mass change. *Earth's Future*, 8(7), e2019EF001470, 2020.
- Meinshausen, N.: Quantile regression forests. *Journal of Machine Learning Research*, 7(35): 983–999, 2006.
- Meinshausen, M., Schleussner, C. F., Beyer, K., Bodeker, G., Boucher, O., Canadell, J. G., et al.: A perspective on the next
generation of Earth system model scenarios: towards representative emission pathways (REPs), *Geoscientific Model*
530 *Development Discussions*, 2023, 1-40, 2023.
- Merrifield, A. L., Brunner, L., Lorenz, R., Medhaug, I., and Knutti, R.: An investigation of weighting schemes suitable for
incorporating large ensembles into multi-model ensembles, *Earth System Dynamics*, 11(3), 807-834, 2020.
- Merrifield, A. L., Brunner, L., Lorenz, R., Humphrey, V., and Knutti, R.: Climate model Selection by Independence,
Performance, and Spread (ClimSIPS v1. 0.1) for regional applications, *Geoscientific Model Development*, 16(16), 4715-4747,
535 2023.
- 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,
doi:10.5194/gmd-9-4521-2016, 2016.
- Nowicki, S., Payne, A. J., Goelzer, H., Seroussi, H., Lipscomb, W. H., Abe-Ouchi, A., Agosta, C., Alexander, P., Asay-Davis,
540 X. S., Barthel, A., Bracegirdle, T. J., Cullather, R., Felikson, D., Fettweis, X., Gregory, J., Hatterman, T., Jourdain, N. C.,
Kuipers Munneke, P., Larour, E., Little, C. M., Morlinghem, M., Nias, I., Shepherd, A., Simon, E., Slater, D., Smith, R.,
Straneo, F., Trusel, L. D., van den Broeke, M. R., and van de Wal, R.: Experimental protocol for sealevel projections from
ISMIP6 standalone ice sheet models, *Cryosphere*, 14, 2331–2368, doi:10.5194/tc-14-2331-2020, 2020.
- Nowicki, S., Payne, A. J., Seroussi, H. L., Goelzer, H., Smith, R. S., Asay-Davis, X., et al.: Plans for the Ice Sheet Model
545 Intercomparison Project for CMIP7, In *AGU Fall Meeting Abstracts*, vol. 2023, pp. G52A-07, 2023.
- Otosaka, I. N., Shepherd, A., Ivins, E. R., Schlegel, N. J., Amory, C., van den Broeke, M. R. et al.: Mass balance of the
Greenland and Antarctic ice sheets from 1992 to 2020. *Earth Syst. Sci. Data*, 15(4), 1597–1616, 2023.

- 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.
- 550 Rohmer, J., Thieblemont, R., Le Cozannet, G., Goelzer, H., & Durand, G.: Improving interpretation of sea-level projections through a machine-learning-based local explanation approach, *The Cryosphere*, 16(11), 4637-4657, 2022.
- Rohmer, J., Idier, D., Paris, F., Pedreros, R., & Louisor, J.: Casting light on forcing and breaching scenarios that lead to marine inundation: Combining numerical simulations with a random-forest classification approach. *Environmental Modelling & Software*, 104, 64-80, 2018.
- 555 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.
- Seroussi, H., Pelle, T., Lipscomb, W. H., Abe-Ouchi, A., Albrecht, T., Alvarez-Solas, J., et al.: Evolution of the Antarctic Ice Sheet over the next three centuries from an ISMIP6 model ensemble, *Earth's Future*, 12(9), e2024EF004561, 2024.
- Shiogama, H., Fujimori, S., Hasegawa, T., Hayashi, M., Hirabayashi, Y., Ogura, T., Iizumi, T., Takahashi, K. and Takemura, T.: Important distinctiveness of SSP3-7.0 for use in impact assessments. *Nature Climate Change*, 13(12), pp.1276-1278, 2023.
- 560 <https://doi.org/10.1038/s41558-023-01883-2>
- Sim, R. H. L., Xu, X., and Low, B. K. H.: Data Valuation in Machine Learning: "Ingredients", Strategies, and Open Challenges, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence Survey Track*, 5607-5614, doi:10.24963/ijcai.2022/782, 2022.
- 565 Silverman, B. W.: *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall, ISBN 9780412246203, 1986.
- 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, doi:10.5194/tc-14-985-2020, 2020.
- 570 Storlie, C. B., Swiler, L. P., Helton, J. C., and Sallaberry, C. J.: Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models. *Reliability Engineering & System Safety*, 94(11), 1735-1763, 2009.
- 575 Tadesse, M., Wahl, T., and Cid, A.: Data-Driven Modeling of Global Storm Surges, *Frontiers in Marine Science*, 7, 260, <https://doi.org/10.3389/fmars.2020.00260>, 2020.
- Tebaldi, C. and Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections, *Philos. T. Roy. Soc. A*, 365, 2053-2075, doi:10.1098/rsta.2007.2076, 2007.
- Turner, F., Rougier, J., Edwards, T., Coulon, V., and Klose, A. K.: Building probabilistic projections of the Antarctic contribution to global sea level rise using a random forests emulator, *EGU General Assembly 2024*, Vienna, Austria, 14-19 Apr 2024, EGU24-16820, doi:10.5194/egusphere-egu24-16820, 2024.
- 580

- Tyralis, H., Papacharalampous, G., and Langousis, A.: A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources, *Water*, 11, 910, <https://doi.org/10.3390/w11050910>, 2019.
- 585 Van Breedam, J., Huybrechts, P., and Crucifix, M.: A Gaussian process emulator for simulating ice sheet–climate interactions on a multi-million-year timescale: CLISEMv1.0, *Geosci. Model Dev.*, 14, 6373–6401, doi:10.5194/gmd-14-6373-2021, 2021.
- Van Katwyk, P., Fox-Kemper, B., Nowicki, S., Seroussi, H., & Bergen, K. J. (2025). ISEFlow v1. 0: A Flow-Based Neural Network Emulator for Improved Sea Level Projections and Uncertainty Quantification. *EGUsphere*, 2025, 1-32.
- Wright, M. N., and König, I. R.: Splitting on categorical predictors in random forests. *PeerJ*, 7, e6339, 2019.
- 590 Yoo, M., Gopalan, G., Hoffman, M., Coulson, S., Han, H. K., Wikle, C. K., & Hillebrand, T.: Emulation with uncertainty quantification of regional sea-level change caused by the Antarctic Ice Sheet. *Journal of Geophysical Research: Machine Learning and Computation*, 2(2), e2024JH000349, 2025.

Appendix A Random Forest regression model

Let us first denote $slc^{i=1,\dots,n}$ the i^{th} value of sea level contribution calculated relative to the i^{th} vector of p input parameters' values $\mathbf{x}^{i=1,\dots,n} = \{x_1, x_2, \dots, x_p\}^{i=1,\dots,n}$ where n is the total number of experiments. 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). By nature, tree models can deal with mixed types of variables, categorical or continuous. 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 $R_1, \dots, R_L$ to have homogeneous Y_i values in each set $R_{l=1,\dots,L}$ by minimizing a splitting criterion, which is chosen in this study as the sum of squared errors (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 $T(\alpha)$, where α 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 slc as:

$$E(slc|\mathbf{X} = \mathbf{x}(s)) = \sum_{j=1}^n w_j(\mathbf{x}(s))slc_j, \quad (A1)$$

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

$$w_j(\mathbf{x}(s)) = \frac{\sum_{t=1}^{n_{tree}} w_j(\mathbf{x}(s), \alpha_t)}{n_{tree}}, \text{ with } w_j(\mathbf{x}(s), \alpha) = \frac{I_{\{X_j \in R_{l(x,\alpha)}\}}}{\#\{i: X_i \in R_{l(x,\alpha)}\}}, \quad (A2)$$

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

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 1,000 because of its smaller influence on the RF model performance (relative to ns and m_{try}). To select values for these parameters, we rely on an approach based on a 10-fold cross validation exercise (Hastie et al., 2009), which consists in varying ns from 1 to 10, and m_{try} from 1 to 7, and in selecting the most optimal combination with respect to cross-validation predictive error. The number of random trees is fixed at 1,000; preliminary tests having showed that this latter parameter has little influence provided that it is large enough.

An additional difficulty of our study is the presence of a large number of categorical variables with large number of levels (unordered values). The partitioning algorithm described above tends to favour categorical predictors with many levels (Hastie et al. (2009): chapter 9.2.4). To alleviate this problem, we rely on the computationally efficient algorithm proposed by Wright et al. (2019) based on ordering the levels *a priori*, here by their mean response.

Appendix B Accounting for the emulator uncertainty

The RF method described in Appendix A is very flexible and can be adapted to predict quantiles, which can be used to assess the RF emulator uncertainty. To do so, we rely on the quantile regression forest (qRF) model, which was originally developed by Meinshausen (2006), who proposed to estimate the conditional quantile $q_\tau(y|\mathbf{x})$ at level τ as

$$q^\tau(slc|\mathbf{x}) = \inf(\mathbf{x}: F_{slc|\mathbf{x}}(slc|\mathbf{x}) \geq \tau), \quad (\text{B1})$$

$$\text{where } F_{slc|\mathbf{x}}(y|\mathbf{x}) = \sum_{j=1}^n w_j(\mathbf{x}) I_{\{slc_j \leq y\}}, \quad (\text{B2})$$

where the weights are calculated in the same manner as for the regression RF model (described in Appendix A). The major difference with the formulation for regression RF is that the qRF model computes a weighed empirical cumulative distribution function of slc for each partition instead of computing a weighted average value.

The quantiles computed using the qRF model can directly be used to define the prediction intervals at any level α : $[q^{(1-\alpha)/2}(slc|\mathbf{x}^*); q^{(\alpha+1)/2}(slc|\mathbf{x}^*)]$, which can be used to reflect the RF emulator uncertainty when providing the emulator predictions.

When performing the probabilistic predictions (Sect. 2.4.2), the emulator uncertainty is propagated in addition to the uncertainty of the different input variables based on the following procedure:

(Step 1) Draw N random realizations of the input variables $\tilde{\mathbf{x}}$;

(Step 2.1) Draw N random number $\tilde{u}$ between 0 and 1 by assuming a uniform random distribution;

(Step 2.2) Compute the N values $\tilde{slc} = q^{\tilde{u}}(slc|\tilde{\mathbf{x}})$ given $\tilde{u}$ and $\tilde{\mathbf{x}}$ using the qRF model;

(Step 2.3) Compute the quantile $Q_{\tilde{u}}^\alpha$ at the chosen level α from the set of N values of $\tilde{slc}$;

(Step 3) Repeat n times Steps 2.1 to 2.3. At Step 2.2, $\tilde{slc}$ are calculated for the same set of random input variables $\tilde{\mathbf{x}}$ defined at

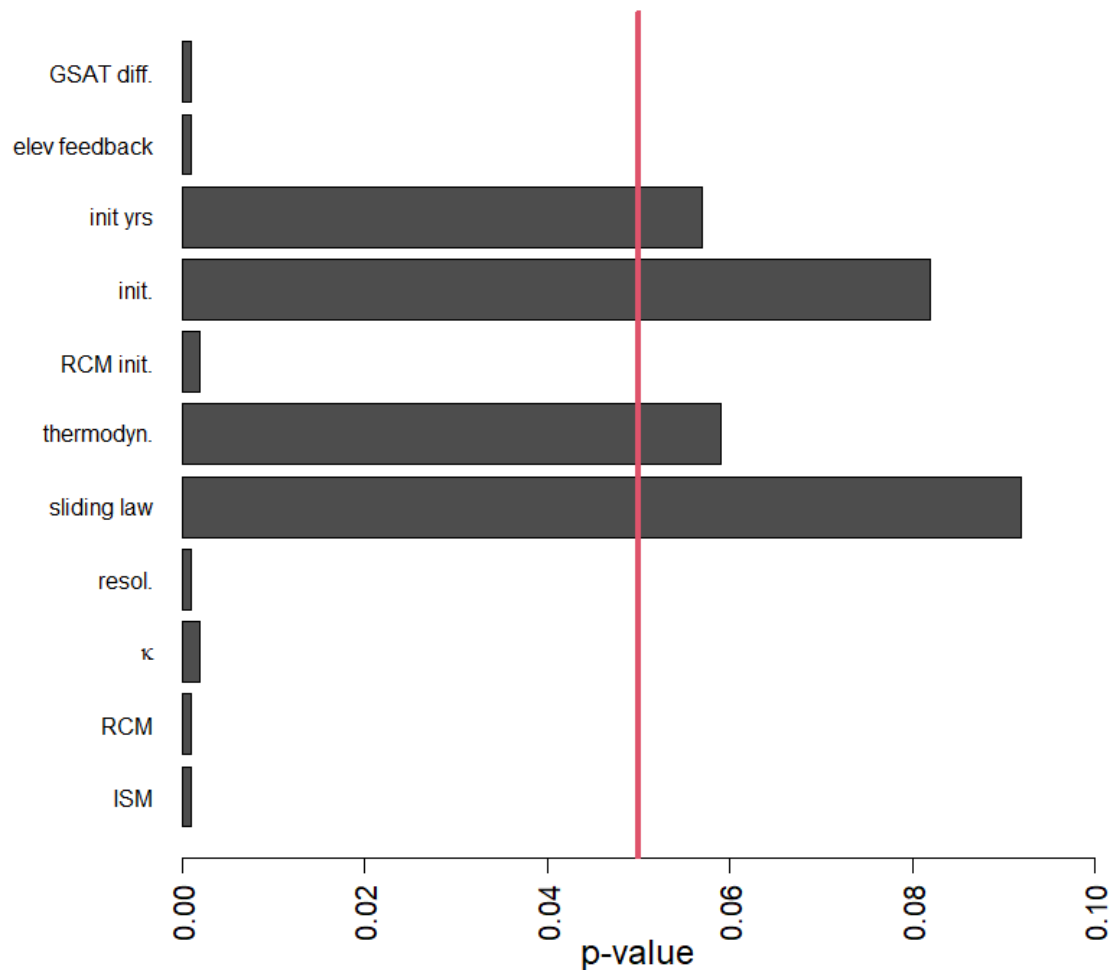
Step 1, but each time a newly randomly generated set of levels $\tilde{u}$ is used based on Step 2.1. This means that, at Step 2.3, the newly calculated quantiles $Q_{\tilde{u}}^\alpha$ vary for each of the repetitions.

The output of the procedure is a set of n quantile values $(Q_{\tilde{u}(1)}^\alpha, Q_{\tilde{u}(2)}^\alpha, \dots, Q_{\tilde{u}(n)}^\alpha)$. The variability among these values reflects the emulator uncertainty and can be summarized by the $\tau\%$ confidence interval with lower and upper bounds defined by the $(1-\tau)/2$, and the $(1+\tau)/2$ quantile of $Q_{\tilde{u}}^\alpha$. In this study, we choose $N=10,000$, $n=100$ and $\tau=90\%$.

Appendix C Screening analysis

We rely on the hypothesis testing of Altmann et al. (2010). To identify the significant predictor variables, the null hypothesis “no association between slc and the corresponding predictor variable” is tested. The corresponding p-value is evaluated by (1) computing the probability distribution of the importance measure of each predictor variable through multiple replications (here 1,000) of permuting slc ; (2) training a RF model; and (3) computing the permutation-based variable importance. When the p-value is below a given significance threshold (typically of 5%), it indicates that the null hypothesis should be rejected, i.e., the considered predictor variable has a significant influence on slc . Figure C1 shows that four predictor variables have non-

significant influence with p-values above 5%, namely the choice in the account for thermodynamics, the choice in the sliding law, the type of initialisation and the number of years for initialisation phase.



655 **Figure C1: Screening analysis showing the p-values of the RF variable importance-based test of independence of Altmann et al. (2010). The vertical red line indicates the significance threshold at 5%. When the p-value is below 5%, it indicates that the null hypothesis should be rejected, i.e., the considered variable has a significant influence, and should retained in the RF construction.**

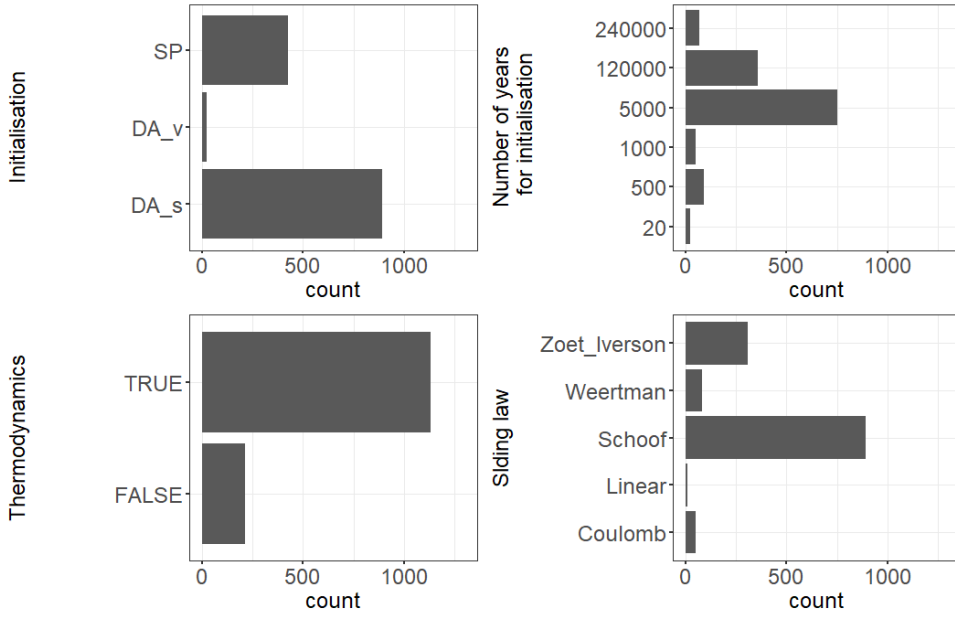


Figure C2: Count number of the MME members with respect to the variables identified as non-influential.

660 Appendix D Formal definition of the performance criteria

Let us consider the slc prediction error, i.e. $e^{(i)} = slc^{(i)} - \widehat{slc}^{(i)}$ for each test sample $i=1, \dots, n_{\text{test}}$. We consider two performance indicators to measure the emulator's predictive capability as follows:

- the relative absolute error, $RAE = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left| \frac{e^{(i)}}{slc^{(i)}} \right|$ (quoted as a percentage);
- the coefficient of determination, $Q^2 = 1 - \frac{\sum_{i=1}^{n_{\text{test}}} (e^{(i)})^2}{\sum_{i=1}^{n_{\text{test}}} (slc^{(i)} - \overline{slc})^2}$ where $\overline{slc}$ is the average value of slc calculated over the

665 test set.

In addition, we consider a third performance indicator, i.e., the continuous rank probability score $CRPS$, that jointly quantifies the calibration of qRF probability distribution, i.e. the reliability of the estimation, and its sharpness (i.e. the concentration/dispersion of the probability distribution). To evaluate the $CRPS$ score, the formulation based on quantiles (Berrisch and Ziel (2024): Eq. 2) is used:

$$670 \quad CRPS = 2 \int_0^1 B(q^\tau(slc|\mathbf{x}^*), slc^{true}) d\tau \approx \frac{2}{P} \sum_{\tau \in \Gamma} B(q^\tau(slc|\mathbf{x}^*), slc^{true}),$$

where the term $B(q^\tau(slc|\mathbf{x}^*), slc^{true})$ is defined as:

$$\begin{cases} (1 - \tau)(q^\tau(slc|\mathbf{x}^*) - slc^{true}) & \text{if } slc^{true} < q^\tau(slc|\mathbf{x}^*) \\ \tau(slc^{true} - q^\tau(slc|\mathbf{x}^*)) & \text{if } slc^{true} \geq q^\tau(slc|\mathbf{x}^*) \end{cases},$$

where slc^{true} is the true value of the sea level contribution, and where the quantiles $q^{\tau}(slc|\mathbf{x}^*)$ are evaluated using the trained qRF model at given instance of the input variables $\mathbf{x}^*$ for an equidistant dense grid of quantile levels $(\tau_1, \dots, \tau_p)$ with $\tau_1 < \tau_{i+1}$ and $\tau_{i+1} - \tau_i = 1/P$. In this study, we consider level $\tau_1=5\%$ and $\tau_p =95\%$ with $1/P=5\%$.

Appendix E List of acronyms / abbreviations

Name	Description
AR	Assessment Report
CMIP	Coupled Model Intercomparison Project
CRPS	Continuous Ranked Probability Score
GCM	Global Climate Model
GrIS	Greenland Ice-Sheet
GSAT	Global Surface Atmosphere Temperature
GWL	Global Warming Level
IPCC	Intergovernmental Panel on Climate Change
ISM	Ice-Sheet Model
ISMIP	Inter-Sectoral Impact Model Intercomparison Project
qRF	Quantile Random Forest
MME	Multi-model ensemble
RAE	Relative Absolue Error
RCM	Regional Climate Model
RCP	Representative Concentration Pathway
RF	Random Forest
<i>slc</i>	Sea level contribution
SMB	Surface Mass Balance
SSP	Shared Socio-economic Pathways