Annals of Glaciology

Article

Cite this article: Luo X, Nowicki S (2025) Model weighting for ISMIP6-Greenland based on observations and similarity among models. *Annals of Glaciology* **66**, e14, 1–16. https://doi.org/10.1017/aog.2025.10010

Received: 22 November 2024
Revised: 23 April 2025
Accepted: 16 May 2025

Keywords:
Greenland; ISMIP6; model weighting; quality; similarity

Corresponding author: Xiao Luo;
Email: xiaoluo@buffalo.edu

cambridge.org/aog

Model weighting for ISMIP6-Greenland based on observations and similarity among models

Xiao Luo[1] and Sophie Nowicki[1,2]

[1]Department of Earth Sciences, College of Arts and Sciences, University at Buffalo, State University of New York, Buffalo, NY, USA and [2]RENEW Institute, College of Arts and Sciences, University at Buffalo, State University of New York, Buffalo, NY, USA.

Abstract

The Ice Sheet Model Intercomparison Project for CMIP6 (ISMIP6) resulted in many ice-sheet simulations from multiple ice-sheet models. To date, no model weighting studies have analyzed or quantified the model performance, possible duplication of the ISMIP6 ice-sheet models and the effect on mass loss projections. In this study, we adopt a model weighting scheme for the ISMIP6-Greenland that accounts for both model performance compared to observation and model similarity due to possible duplication. We choose ice velocity and thickness for the measurement of model performance, and we use all suitable variables to compute similarity indexes. We update the sea level rise contribution from ISMIP6-Greenland by the end of this century with the weights, and we find that, although the multi-model mean is not considerably shifted (mostly within $\pm$1cm), the model spreads are reduced by 10–30% after applying the model weights. The magnitude of reduction varies largely among experiments and types of model weights applied. In general, we find that the model weighting scheme is skillful in producing model weights that effectively and reasonably quantify the model performance and inter-dependency, which can potentially benefit the future phase of the Ice Sheet Model Intercomparison Project, i.e. ISMIP7.

1. Introduction

Greenland ice-sheet mass loss has shown a large contribution to global sea level rise in the past decades (Shepherd and others, 2020) and will continue to play an important role in future sea level rise (Hofer and others, 2020; Fox-Kemper and others, 2021). The Ice Sheet Model Intercomparison Project for CMIP6 (ISMIP6) (Nowicki and others, 2016, 2020) serves as an important estimator for ice-sheet evolution in the future, showing considerable spreads by the end of the 21st century for the Greenland ice sheet (Goelzer and others, 2020; Payne and others, 2021). Following the approach taken by previous ice-sheet community efforts, such as Sea-level Response to Ice Sheet Evolution (SeaRISE; Bindschadler and others, 2013; Nowicki and others, 2013), the analysis of the ISMIP6 ice-sheet model ensemble has adopted a 'one model one vote' strategy that assigns equal weights to each model.

The issues of assigning equal weights have been extensively discussed in climate modeling literature (Knutti, 2010; Knutti and others, 2010, 2017; Masson and Knutti, 2011; Pennell and Reichler, 2011), but not thoroughly explored in the ice-sheet modeling realm. There are existing studies that used calibration strategy to generate performance scores for ice-sheet model ensemble (Gladstone and others, 2012; Ritz and others, 2015; DeConto and Pollard, 2016; Nias and others, 2019, 2023; Brinkerhoff and others, 2021; Aschwanden and Brinkerhoff, 2022; Felikson and others, 2023; Jager and others, 2024), but this approach has not yet been applied to the ISMIP6 ensemble. Furthermore, as these studies mostly calibrate on single model ensembles, the model inter-dependence is not a pertinent topic. The single model ensemble members are essentially different realizations branching from the same model with different model parameters, while the ISMIP6 models are different ice-sheet models.

Similar to the issues faced by climate models, the assignment of equal weights to ice-sheet models has the following assumptions: (i) each ice-sheet model has an equal performance of capturing the present-day ice-sheet state (e.g. ice thickness, surface ice velocity, and temperature) and projecting the ice-sheet evolution, and (ii) all models in the ensemble are independently developed without any duplications or exchanges of modeling ideas, codes, and subcomponents. For the first assumption, ice-sheet models do not perform equally even at their initial states, which can be shown by the model errors of ice velocity and thickness compared to observation documented in SeaRISE (Bindschadler and others, 2013; Nowicki and others, 2013) and ISMIP6-Greenland (Goelzer and others, 2020). Also, it is highly unlikely that the models will all have equal performance in projecting the ice sheet into the future, which is shown by the great differences in sea level projections in 2100 reported by ISMIP6-Greenland. The uncertainty range of contributions is almost the same magnitude as the ensemble mean (Goelzer and others, 2020; Payne and others, 2021).

As for the model inter-dependence, it is unreasonable to assume the models are completely independent. Commonly, the ice-sheet models have similar initialization processes (e.g. data assimilation using the same velocity or thickness) and similar physical laws. If these similar simulations are counted repetitively, the unweighted ensemble will be biased toward the repeated projections. This is true for ISMIP6-Greenland, as multiple submissions come from the same ice-sheet model. For example, the Ice-sheet and Sea-level System Model (ISSM; Larour and others, 2012) has been used by multiple modeling groups, and in some cases, a group might have submitted several simulations. The three ISSM submissions from Alfred Wegener Institute (AWI) have mostly identical configurations. The main differences are that they are run with slightly different horizontal resolutions over the fast-flowing regions (Goelzer and others, 2020). Therefore, they are expected to produce similar simulations (Rückamp and others, 2020). In contrast, the ISSM simulations from other groups (e.g. JPL_ISSMPALEO) might have larger differences resulting from different initialization techniques and modeling choices. These submissions are all considered independent models in Goelzer and others (2020) and Payne and others (2021), motivating our interests to account for not just model performance but also interdependence to better interpret the information provided by ISMIP6.

For the model weighting strategy, it is challenging to define the correct metrics (or diagnostics/variables) to measure the model performance because it is difficult to have a proper definition of the general skills of models (Knutti and others, 2017). When the simulations are scored based on different metrics, each model could outperform the other models on one metric but underperform on another. No decisive conclusions may be properly drawn in terms of which metric is the 'best one' or 'most appropriate one'. The choices of diagnostics might have considerable influences on the performance weighting, as demonstrated by both climate model weighting (Lorenz and others, 2018) and a recent study using Bayesian calibration to a single ice-sheet model (Felikson and others, 2023). The latter study assigned weights using ice velocity, dynamic thickness change, and mass balance separately as diagnostics for calibration and showed largely different posterior distributions. Also, for the same diagnostic, such as ice surface velocity, complexity may arise from the initialization step of some ice-sheet models that used data assimilation via velocity. If the same observation is chosen for measuring the performance, the weighting scheme will likely favor the models that used the same observation for the data assimilation, while the other models may be scored lower.

Yet, it is not trivial to assign weights based on some metrics, as shown in the practices in the climate modeling community for extracting credible and reliable information from multi-model ensembles. The Climate model Weighting by Independence and Performance project (ClimWIP; Brunner and others, 2019; Merrifield and others, 2020), following the previous work of Sanderson and others (2015) and Knutti and others (2017), has assigned weights to CMIP6 models that lead to a reduction of the model spreads. A recent ClimWIP study (Brunner and others, 2020) utilized an updated version of this model weighting strategy and found a reduction of model spreads and generally lower global temperature rise under both weak and strong climate scenarios. This reduction is because several climate models with strong warming received low weights. In this study, we utilize the ClimWIP model weighting framework to assign model weights to the ice-sheet models that participated in ISMIP6-Greenland and we investigate if the model weighting shifts sea level projections and to what extent.

2. Data and methods

2.1. Model weighting scheme

The ClimWIP model weighting scheme is generally applicable for process-based model weighting, as it accounts for both the performance of model simulation and the similarity of a certain model in a multi-model ensemble due to possible duplication of codes and components. In short, a certain model is weighted by both skill and independence. Essentially, the weight of a model w_i is determined via two parameters: (1) the weight of quality w_q, which measures how close the simulations are to the observed values, and (2) the weight of uniqueness w_u, which scores higher if the model demonstrates uniqueness compared with other models in the ensemble. The total weight w_i is then evaluated as the multiplication of these two quantities, as shown in Eq. (1):

$$w_i = w_q \times w_u = \exp\left[-\left(\frac{D_{i(obs)}}{D_q}\right)^2\right] \times \frac{1}{1+\sum_{j\neq i}^{m}\exp\left[-\left(\frac{D_{ij}}{D_u}\right)^2\right]} \quad (1)$$

Whereas $D_{i(obs)}$ is the distance of the ith model from the true observation, D_{ij} is the model similarity between a model pair. Following the method described by Sanderson and others (2017), we evaluate both $D_{i(obs)}$ and D_{ij} as root mean square errors (RMSEs). The pairwise distances are calculated for each variable and then linearly combined to formulate the distance matrices. For each variable, only the grid cells where all models and observations have valid values are retained to compute the distances, and other grids are removed. The ice-sheet mask of each model will evolve in future projections as some marginal grids are retreated and marked as ice-free. However, the time-varying ice masks in the projection do not have influences here because we use only the snapshots of ice thickness and velocity in 2015, as described in more detail in Section 2.2.

We note that any form of model weighting scheme is not an absolute but somehow subjective choice regarding the conversion from model distances to model weights. The form in Eq. (1) is also empirically defined, and it meets certain important criteria: (1) for quality weighting, the model gets increasingly higher weights as it approaches the observation and infinitely converges to zero when the model–observation distance gets farther; (2) likewise, for similarity weighting, a similarity index $\exp[-(\frac{D_{ij}}{D_u})^2]$ is computed for each model-to-model pair and they are summed up, and the value of a model to itself is 1 because D_{ij} is zero. The reciprocal form taken by the similarity weight ensures that the weight will approach 1 if the distance of one model to every other model is small, and it will converge to 0 if the summed distance gets larger. D_u and D_q are free parameters representing the radius of uniqueness and the radius of quality, respectively. They are used to scale the distances $D_{i(obs)}$ and D_{ij}. D_u and D_q are quantified as percentiles of the mean of the inter-model distances. We explore the choices of these two parameters in later sections.

Observations are necessary to quantify the weights of quality, but they are not needed to measure the weights of uniqueness (interchangeable with 'weights of similarity'). Therefore, we use the variables described in Section 2.2 for quality weighting, and we use the simulated quantities mentioned in Section 2.3 to produce similarity weights. Finally, all fields are normalized before they are used to compute the weights. The normalization is the classic 'z-score' standardization so that the normalized data has a mean of 0 and a standard deviation (std dev.) of 1.

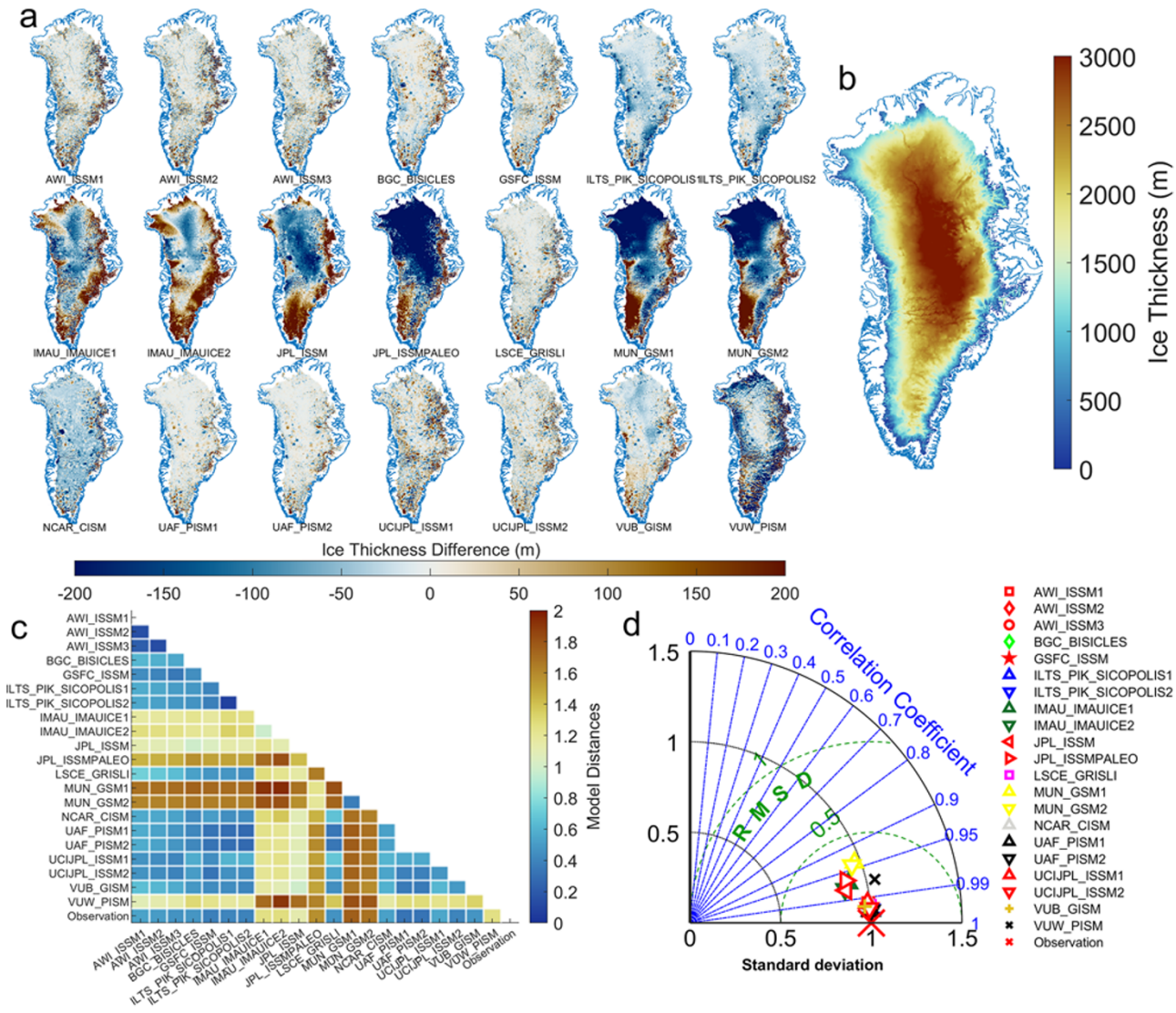

Figure 1. (a) The differences of simulated ice-sheet thickness by ISMIP6-Greenland ice-sheet models and (b) observed ice thickness magnitude obtained from BedMachine datasets (Morlighem and others, 2017, 2020). The differences between model pairs and model–observation are visualized with (c) heatmap and (d) Taylor diagram. Note that the values shown in (c) and (d) are normalized, and the red cross in (d) corresponds to the observation shown in (b).

2.2. Simulation and observational data for quality weighting

In this study, we limit our focus to the same variables as used in Goelzer and others (2020) to assign performance weights to models, i.e. ice velocity and thickness. The simulated ice thickness and ice velocity fields are used directly from the outputs of ISMIP6-Greenland (Goelzer and others, 2020), and we plot the differences with observations in Figs 1a and 2a. We use the snapshots of ice velocity and thickness at the beginning of the control projection ('ctrl_proj' hereafter) to compare with the observational datasets. We note that not all ice-sheet models report surface ice velocity; for instance, the ice-sheet models BGC_BISICLES, IMAU_IMAUICE1, and IMAU_IMAUICE2 only provide vertically averaged ice velocity fields, as these three models are run under either Shallow-Ice Approximation or Shallow-Shelf Approximation. To maintain consistency with Goelzer and others (2020), which compares the vertically averaged velocity with observation, we use the same mean velocity to compare with observation. The ice thickness observation (Fig. 1b) is obtained from BedMachine datasets (Morlighem and others, 2017, 2020). The observed surface ice velocity fields (Fig. 2b) are obtained from MEaSUREs Greenland Ice Sheet Velocity Map from InSAR Data, Version 2 (Joughin and others, 2015), for the Greenland ice sheet. The coverage of satellite mosaics varies from year to year, particularly in the southeast of the Greenland ice sheet. We use the 2016–2017 mosaics for Greenland that were produced mostly from Sentinel-1A/1B data and provided almost complete coverage over the southeast region (Joughin and others, 2015). The same period is used for the ISMIP6-Greenland for the model weighting regarding surface velocity.

For Figs 1 and 2, we use both heatmaps and Taylor diagrams (Taylor, 2001) to visualize the inter-model distances as well as model–observation differences. In a heatmap, each cell shows the distance between either a model pair or model and observation (last row in Fig. 1c). In a Taylor diagram, the RMSE, the std dev. and the Pearson correlation between models and observation are shown. Note that the values shown in both heatmaps and Taylor diagrams are normalized, and the red crosses in the Taylor diagrams (Figs 1d and 2d) correspond to the observations shown

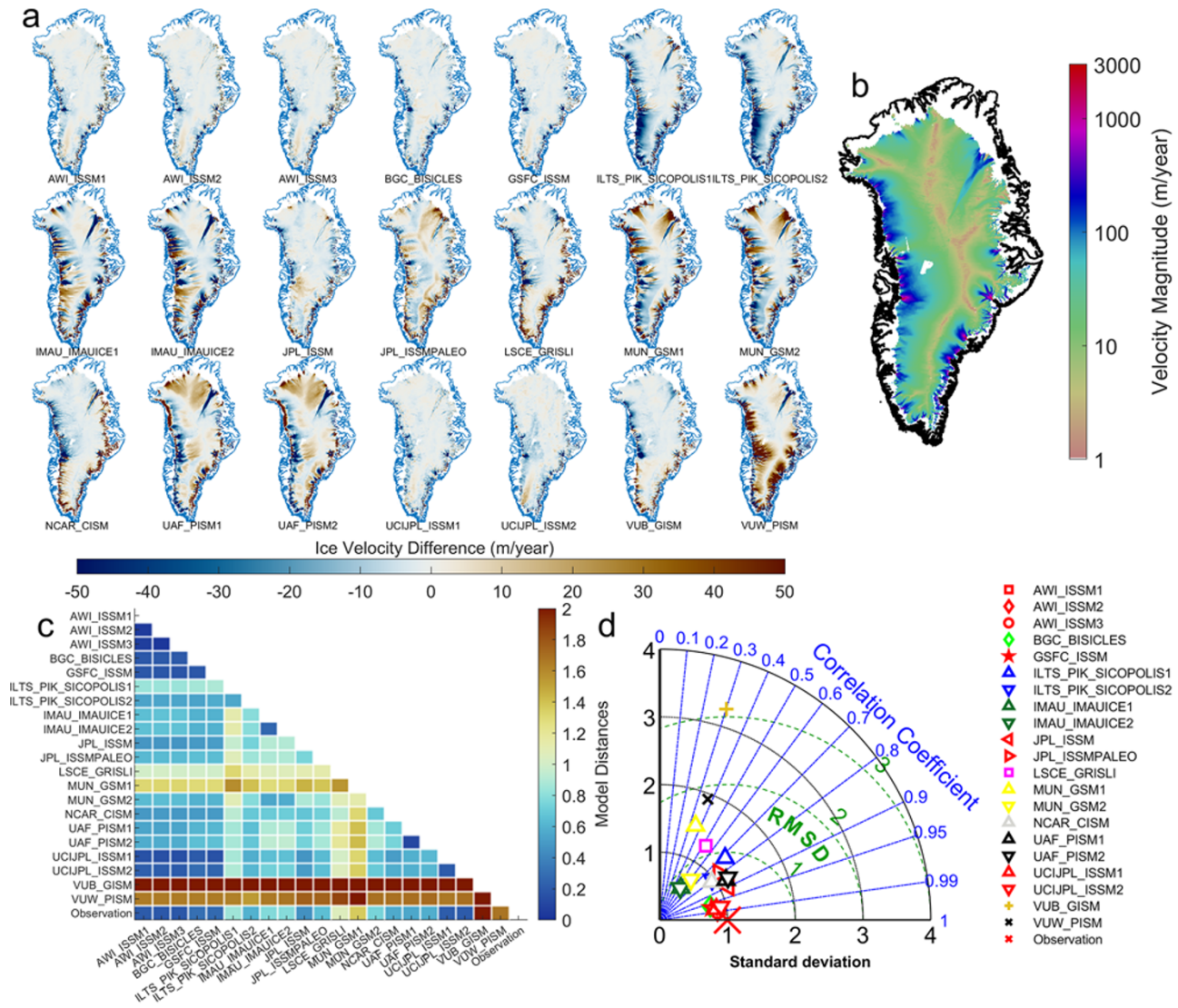

Figure 2. (a) The differences of simulated ice-sheet velocity by ISMIP6-Greenland ice-sheet models and (b) observed ice velocity obtained from MEaSUREs Greenland Ice Sheet Velocity Map from InSAR Data, Version 2 (Joughin and others, 2015). The differences between model pairs and model–observation are visualized with (c) heatmap and (d) Taylor diagram. Note that the values shown in (c) and (d) are normalized, and the red cross in (d) corresponds to the observation shown in (b).

in Figs 1b and 2b. From the model–observation distances shown in the last row in Fig. 1c, three models (MUN_GSM1, MUN_GSM2 and JPL_ISSMPAELO) show the largest differences between simulations and observations. This can be anticipated from Fig. 1a as well.

The relationships between model pairs are captured in the heatmap (Fig. 1c) as well. For instance, the three ISSM submissions from AWI show small differences between each other. This is expected because these three submissions start from the same initial ice-sheet state, and their major differences are mainly the minimum horizontal resolutions and mesh grids. From the Taylor diagram of the ice thickness (Fig. 1d), it can be seen that all models have high correlations (from 0.95 to 0.99) with observation, meaning that all models capture the spatial features of thickness well. In addition, the distances from models to observation are also shown as radial distances from models to the reference point in the Taylor diagrams.

For the ice velocity comparison, there are some models that show large differences between simulated ice velocity and observations (Fig. 2c), e.g. VUB_GISM and VUW_PISM. However, the heatmap scales by all distances and it is difficult to convey any useful information for the rest of the models (other than VUB_GISM and VUW_PISM), as their distances seem to be similar. On the other hand, the Taylor diagram highlights more on the model–observation differences and gives a clearer view of which models are farther from observation and which are closer. We notice that most ISSM submissions match the observation quite well except JPL_ISSMPALEO. This is because it used a longer period of interglacial spin-up, while others used data assimilation of the ice velocity that matches better with present-day observation, such as AWI_ISSMs and GSFC_ISSM. Note that the ISSM submissions from JPL used different velocity datasets (Rignot and Mouginot, 2012) for data assimilation; therefore, this might add to the distance between model and observation because we use velocity data from Joughin and others (2015) as observation. However, larger differences are observed for other models; therefore, we do not expect much difference if a different ice velocity dataset is used for quality weighting.

Table 1. ISMIP6 variables used to generate similarity weighting

	Variable name	Variable	Units
1	acabf	Surface mass balance flux	kg m^{-2} s^{-1}
2	dlithkdt	Ice thickness imbalance	m s^{-1}
3	hfgeoubed	Geothermal heat flux	W m^{-2}
4	litempbotgr	Basal temperature beneath grounded ice sheet	kg m^{-2} s^{-1}
5	litemptop	Surface temperature	K
6	lithk	Ice thickness	m
7	orog	Surface elevation	m
8	topg	Bedrock elevation	m
9	xvelbase	Basal velocity in the *x*-direction	m s^{-1}
10	xvelmean	Mean velocity in the *x*-direction	m s^{-1}
11	yvelbase	Basal velocity in the *y*-direction	m s^{-1}
12	yvelmean	Mean velocity in the *y*-direction	m s^{-1}

In summary, models demonstrate different performances compared to different observations. For example, JPL_ISSMPALEO and MUN_GSM show large differences between simulated and observed ice thickness but smaller differences in velocity, while some models (e.g. VUB_GISM and VUW_PISM) show the opposite. This highlights the value of using both variables for quality weighting for the remainder of this study. Finally, we note that the purpose of the correlation coefficients shown in the Taylor diagrams (Figs 1d and 2d) is to provide another angle to view the differences between simulation and observation. In the actual model weighting, the correlation is not a part of the calculation according to Eq. (1).

2.3. ISMIP6 data for similarity weighting

Since similarity weighting does not require a complete set of corresponding observational data for each field considered, we use as many fields as we can to calculate the model similarity weights. We also exclude the variables with more than five models that do not have simulations in the first year of 'ctrl_proj'. This gives a reasonable size of variables that we can use for similarity weighting. The fields used to produce similarity weights are tabulated in Table 1. The other variables in the data repository are not considered because there are more than five models that do not report them.

We take each variable tabulated in Table 1 directly from the ISMIP6-Greenland dataset and reshape the field into a vector. The distances between each pair of models are evaluated as RMSE. The inter-model distances are then shown as heatmaps utilizing data from different times: initial state (Fig. 3a), 2100 in exp05 (Fig. 3b), 2100 in all experiments excluding control and control projection (Fig. 3c) and averaged values of initial state and 2100 in all experiments (Fig. 3d). The reason for testing model similarity beyond the initial state of the ice sheet is that, despite its large influence on the future ice-sheet evolution as shown in Goelzer and others (2018), similar initial conditions do not always guarantee that a pair of models are truly similar, even if they stem from the same modeling group. Although they might have been initialized in an identical fashion, they can respond differently to climate forcing due to different modeling choices in the projections, eventually leading to different sea level rise projections. Therefore, we use both the initial condition of the ice sheet in the first year of 'ctrl_proj' experiment and the last year of different experiments to measure the proximity of models.

For model distances shown in Fig. 3b, we use exp05 for all models except for UAF_PISM2, which did not participate in exp05. Therefore, we use its expc01 submission instead. For the heatmap in Fig. 3c, all experiments are considered to generate the model distances, and we do not need to make substitutions for UAF_PISM. In the case when not all models participated in an experiment, then the distances are only computed using the available models. Finally, the model distances are averaged over each experiment. Figure 3d shows the average values of Figs 3a and 3c.

We observe that the three submissions from AWI_ISSM have the smallest model distances to each other in the ensemble, which can be seen from both the initial state (Fig. 3a) and future projections (Figs 3b and 3c). JPL_ISSM, JPL_ISSMPALEO, MUN_GISM (1 and 2), UAF_PISM (1 and 2) and VUW_PISM show the largest model differences with other models for both initial state (Fig. 3a) and future projections in exp05 (Fig. 3b). VUW_PISM is less distinct when all experiments are considered (Fig. 3c). We also note that the model distances are smaller when using future projections in exp05 (Fig. 3b) and much smaller when using all future projections (Fig. 3c) compared to using the initial state only (Fig. 3a). This motivates the utilization of both initial state and future projection for the similarity measurement. The usage of only one experiment may not fully capture model behaviors, because ice-sheet models may respond similarly to one set of climate forcing but differently to another. For instance, one ice-sheet model may be sensitive to high atmospheric forcing but not oceanic forcing, and the ISMIP6 project was designed to sample their different responses to climate forcing (table 3 in Nowicki and others, 2020). In addition, ice-sheet models may have different ocean forcing strategies (standard or open experiments) such as UAF_PISM1 and UAF_PISM2. Therefore, it is meaningful to explore the complete set of experiments for model similarity. On the other hand, using only future projections to infer the similarity weighting could also be problematic. For instance, two models could have distinctly different initial conditions, parameterizations and all other components, yet they may still have similar projections for the future. This does not mean they should be given less weights because they are still different models. The intention of similarity weighting is to avoid double-counting of similar models, but not to give less weights to the models that happen to project similar sea level rise. In fact, if two different models simulate similar future ice-sheet configurations, there should be higher confidence in this future (i.e. larger weights) rather than lower.

3. Results

3.1. Similarity weighting

As mentioned in Section 2, the radius of uniqueness D_u that represents the nearest distance of one model is a free parameter; therefore, we show the similarity weights of each model with varying uniqueness radii expressed as percentiles of the mean of the intermodal distances in Fig. 4. The similarity weights are separately computed using the initial condition (Fig. 4a), year 2100 in exp05 (Fig. 4b) and year 2100 in all experiments (Fig. 4c). The averaged values from Figs 4a and 4c are shown in Fig. 4d.

From the approach in Knutti and others (2017), it has been shown that it is ideal to select a D_u that produces nearly $1/N$ of similarity weights for the same model with different variants (N is the number of variants submitted to ISMIP6). For example, if there are eight submissions of ISSM, then each of them receiving 1/8 of weight is an ideal case; whereas SICOPOLIS receiving a weight of 0.5 is ideal as it has two submissions. However, there does not seem to be an optimal value of D_u that yields approximately $1/N$ of similarity weights for all models (Fig. 4). In addition,

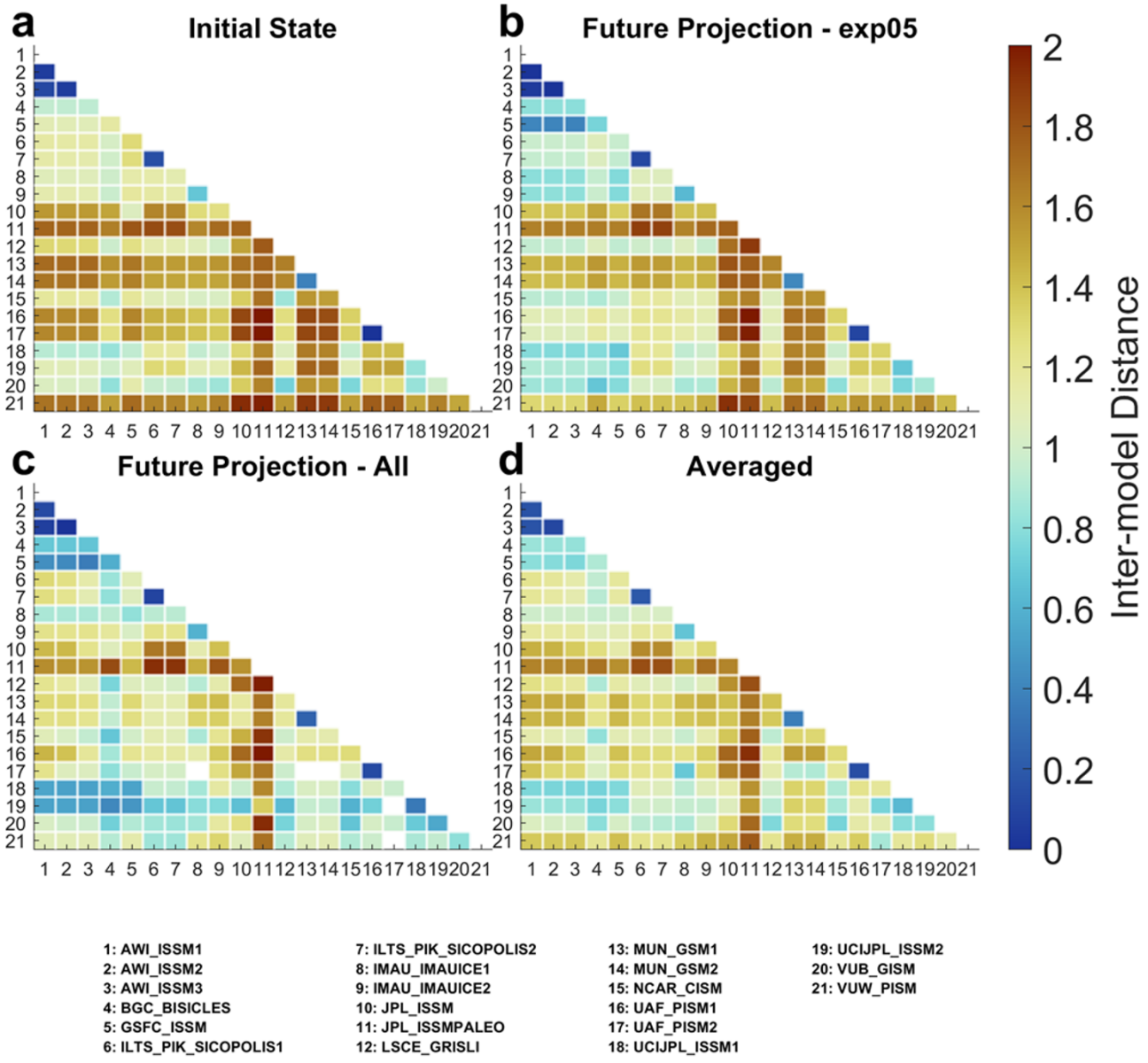

Figure 3. A heatmap representation of the inter-model distances of ISMIP6-Greenland models using the variables listed in Table 1 from (a) initial condition, (b) 2100 in exp05 (with UAF_PISM1 using expc01) and (c) 2100 in all experiments. The averaged inter-model distance of (a) and (c) is shown in (d).

there are many different physical parameterizations and initialization methods for the same ice-sheet model that could lead to different model simulations. For example, ISSM simulated by AWI is quite distinct from the simulations by JPL. Indeed, from all panels in Fig. 4, the three ISSM simulations of AWI receive low weights of similarity due to their similar simulations among themselves and also with other models, while the two ISSM submissions from JPL are distinct from all other models resulting in higher similarity weights.

In this study, we pick an intermediate value of 50% of the mean of inter-model distance as the similarity radius D_u (Fig. 4). We note that this choice of D_u is not a strict choice but rather an empirical selection. With D_u smaller than this value, most models are given similar weights and there is essentially little weighting effect. With greater radius, the few models that are most unique receive the most weights. This results in little weighting effect on all other models, because the rest of them are almost equally downweighted. The two variants of UAF_PISM have almost identical initial conditions, so they would have received much lower weights using 'ctrl_proj' only (Fig. 4a). However, UAF_PISM1 shows a considerably different response (compared to all other models including UAF_PISM2), which leads to higher weights (Fig. 4c). This highlights the value of using both initial conditions and future responses to measure model behavior and their similarities (Fig. 4d). We notice that the differences between UAF_PISM1 and UAF_PISM2 are not obvious using exp05 only (Fig. 4b) because they respond similarly to the climate. Eventually, we use the 50% of the mean inter-model distance as D_u shown in Fig. 4d.

We perform the same practice as in Brunner and others (2020) to examine the validity of similarity weighting via a hierarchical clustering approach using the initial state, and the results are shown

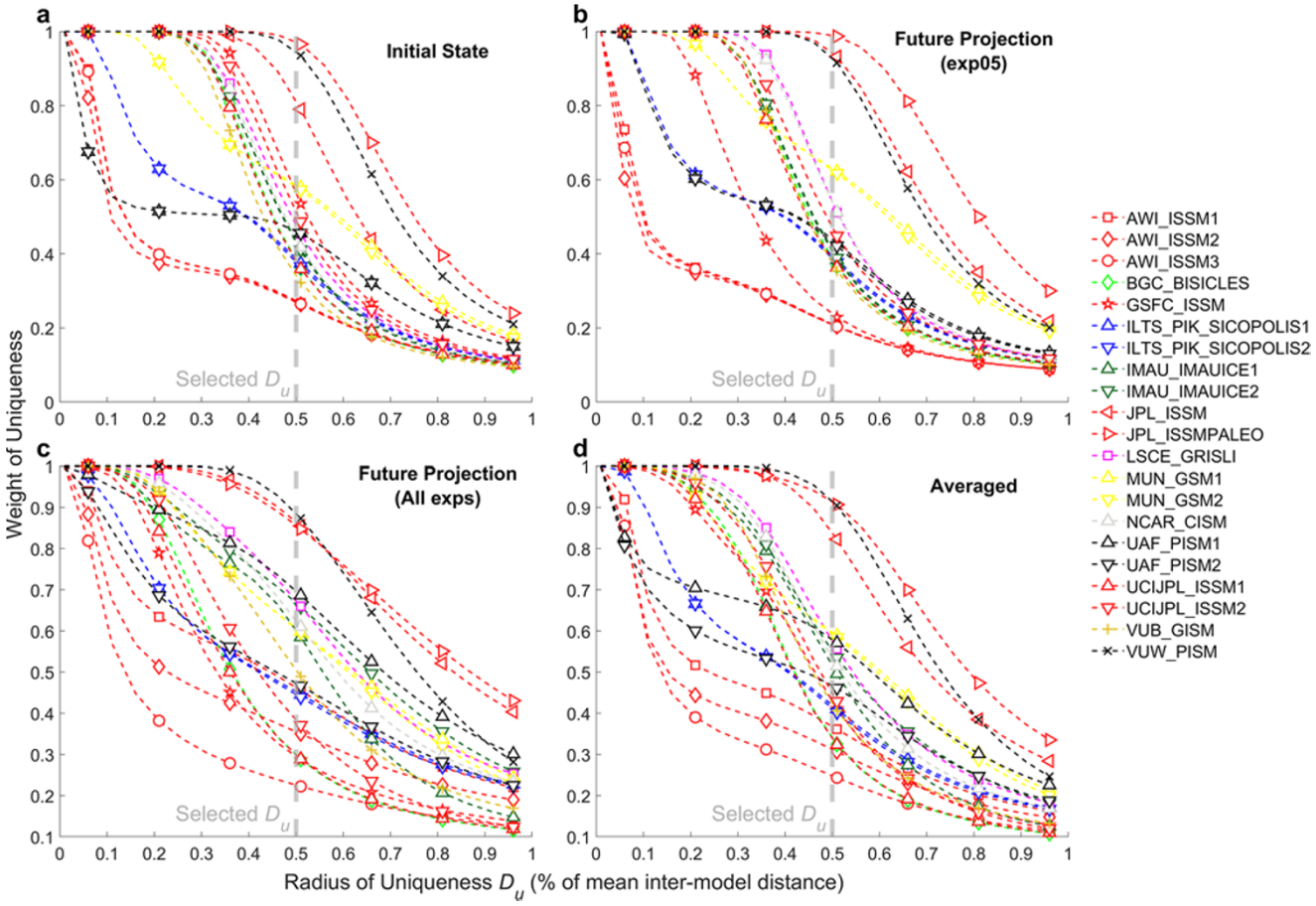

Figure 4. Weight of similarity of ISMIP6-Greenland models with a varying radius of uniqueness D_u measured as percentiles of mean inter-model distances using (a) the initial state, (b) 2100 in exp05 experiments and (c) 2100 in all experiments. The average weight of similarity is shown in (d), which is used as the final similarity weight under selected $D_u = 0.5$ times the mean of inter-model distance.

in Fig. 5. The hierarchical clustering technique automatically sorts similar models into the same family and formulates a complete family tree. When the distances (or cut-offs) are large (beginning from the leftmost side), all models are sorted into the same family. The models are gradually sorted out to different branches in the tree when the cut-off is decreased until each model is its own family (the rightmost side of the family tree).

The 'clusterdata' function in MATLAB is used to perform the hierarchical clustering algorithm and create the dendrogram as shown in Fig. 5. The algorithm first calculates the pairwise distances between each model pair, and a binary tree is then formulated starting by grouping the closest two points (in this case, two models) into one cluster. New clusters are gradually formed, with only two clusters joined at each time. Eventually, all the models are grouped into the same cluster. In Fig. 5, the algorithm proceeds from right to left, but we view the dendrogram from left to right to better tell which models are branched out first.

JPL_ISSMPALEO is the first one to formulate an independent branch. Other unique models that show considerably different initial conditions (see Section 2), including VUW_PISM, MUN_GSM1 and MUN_GISM2, are also rapidly branched out. The vertical line shows the same distance as in Fig. 4. We can observe that, under this cut-off, the unique models that scored high in Fig. 4 (such as JPL_ISSMPALEO) are clearly distinguished from the others. In contrast, the similar models that scored lower are grouped into the same model family. For instance, the three ISSM simulations from AWI are grouped together, and the same applies to UAF_PISM (1 and 2), ILTS_PIK_SICOPOLIS (1 and 2) and IMAU_IMAUICE (1 and 2). GSFC_ISSM and JPL_ISSM are sorted into the same family but not with the other ISSMs from JPL and UCIJPL, indicating that our choice of similarity radius can still highlight the differences among these ISSM models.

3.2. Quality weighting

The choice of the radius of quality D_q is also a free parameter, and we use similar tests as shown in Knutti and others (2017). Ideally, the chosen radius of quality should result in a decrease of RMSE between the weighted ensemble mean and observation (compared to the unweighted ensemble), because we want the weighted ensemble to be closer to observation. Therefore, we show the fraction of RMSE of the weighted ensemble (between model and observation) compared to the unweighted ensemble with varying quality radii measured as percentiles of the mean of model–observation distances (Fig. 6). For both the velocity and the multivariate cases, strong quality weighting with a narrow radius results in the largest decrease of RMSE, that is, when the skill radius equals around 20% of the mean of model–observation distances.

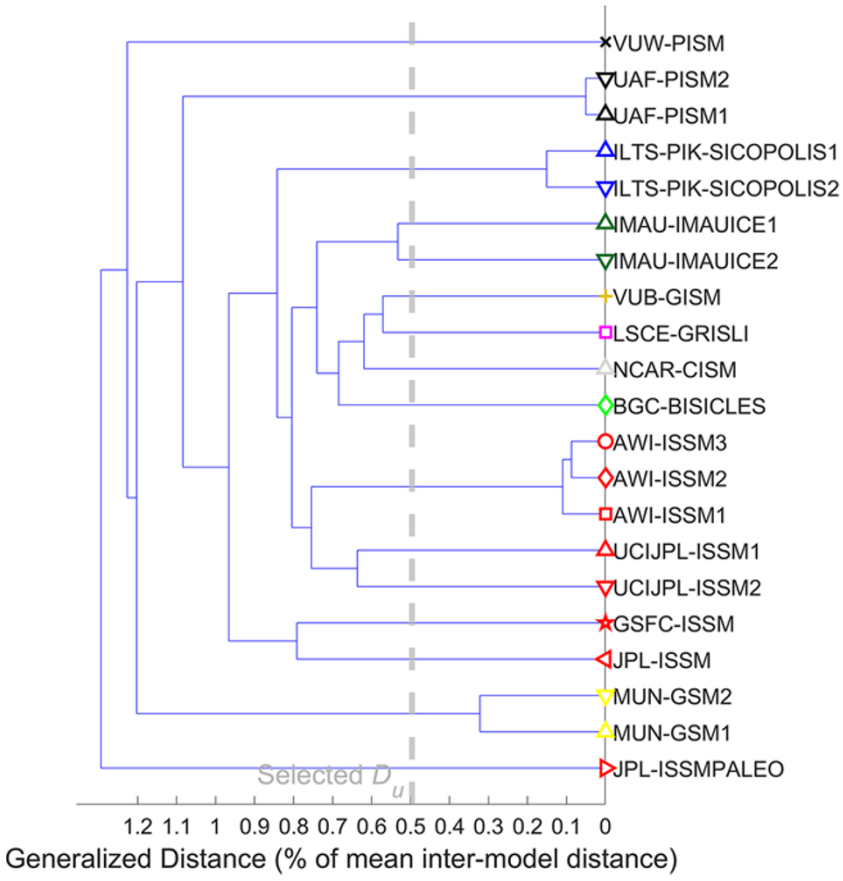

Figure 5. Hierarchical clustering of ISMIP6-Greenland models using the initial conditions of control projections in 2015. The vertical line indicates the selected similarity radius, which is D_u =0.5 times the mean of inter-model distance.

However, the decrease of the present-day bias does not guarantee the weighted ensemble will have better performance in predicting the future. Thus, we also perform an out-of-sample test as a cross-validation, and we show the results shown in Fig. 6b.

Out-of-sample test is an additional examination extending beyond Fig. 6a, which is based on present-day observation. Relying completely on Fig. 6a may potentially lead to over-fitting the weighting scheme to present-day observations and picking up on idiosyncrasies of the observational data (O'Loughlin, 2024). The out-of-sample test then provides a way 'to test the skill of the method, flag overfitting to natural variability or certain data sets, and guide the choice of parameters and metrics' (Knutti and others, 2017, p. 1915), in this case, the choice of the quality parameter D_q. Since we obviously do not have observational data in the future to work with, e.g. ice thickness in 2100, then the out-of-sample test uses each simulated projection in 2100 as 'truth' to validate the weighting choices. This test is mostly considered necessary to pass and we should have more confidence in the weighting choices with this test than without (Knutti and others, 2017).

The velocity and ice thickness of the 'exp05' projection at the end of the 21st century are used to conduct the out-of-sample test. Each model projection in 2100 is iteratively treated as the truth. The distances from the remaining models to this 'truth' are computed, which are also measured as RMSEs, and then summed up. During this procedure, the obvious family members of each model are removed when it is treated as the 'truth' model. For a certain model, if the distance to another model is smaller than its distance to observation, then this model is considered as its family member. The family models are indicated as black cells in Fig. 6c. The diagonal is all removed since the model itself is its family model.

The fractions of RMSE of the weighted ensemble to the unweighted (Fig. 6b) suggest that extreme quality weighting with a small quality radius does not necessarily reduce the bias in future projection. Assigning excessive weights to only a few models (i.e. using a small quality radius), which have a close agreement with the present-day observation, does not guarantee a reduction of future projection bias, especially for ice thickness change (blue-dotted curve in Fig. 6b). Note that Fig. 6a is constructed using the initial state in 2015 only, while Fig. 6b uses the last year in the 'exp05' projection, which is 2100. Eventually, we chose the quality radius D_q as equal to the mean model–observation distance. This choice reduces both the present-day ensemble distance to observation and the ensemble bias of future projection.

3.3. Model weighting results

We show the final weighting results (Fig. 7) in the same fashion as Sanderson and others (2015). We test the weighting scheme under both univariate quality weighting (Fig. 7a using velocity only and Fig. 7b using ice thickness only) and combined weights (Fig. 7c). Obvious clustering behavior is observed for most models, for instance, most ISSM variants receive similar model weights except for the submissions from JPL, indicating the JPL submissions are quite different than the rest of the ISSMs. The ISSM submissions from AWI, GSFC, UCIJPL and BGC_BISICLES used data assimilation for initialization that resulted in close agreement with the present-day ice-sheet state, and therefore these models receive higher quality weights for both velocity and thickness. However, since these models that chose data assimilation as initialization have similar initial conditions, they receive lower similarity

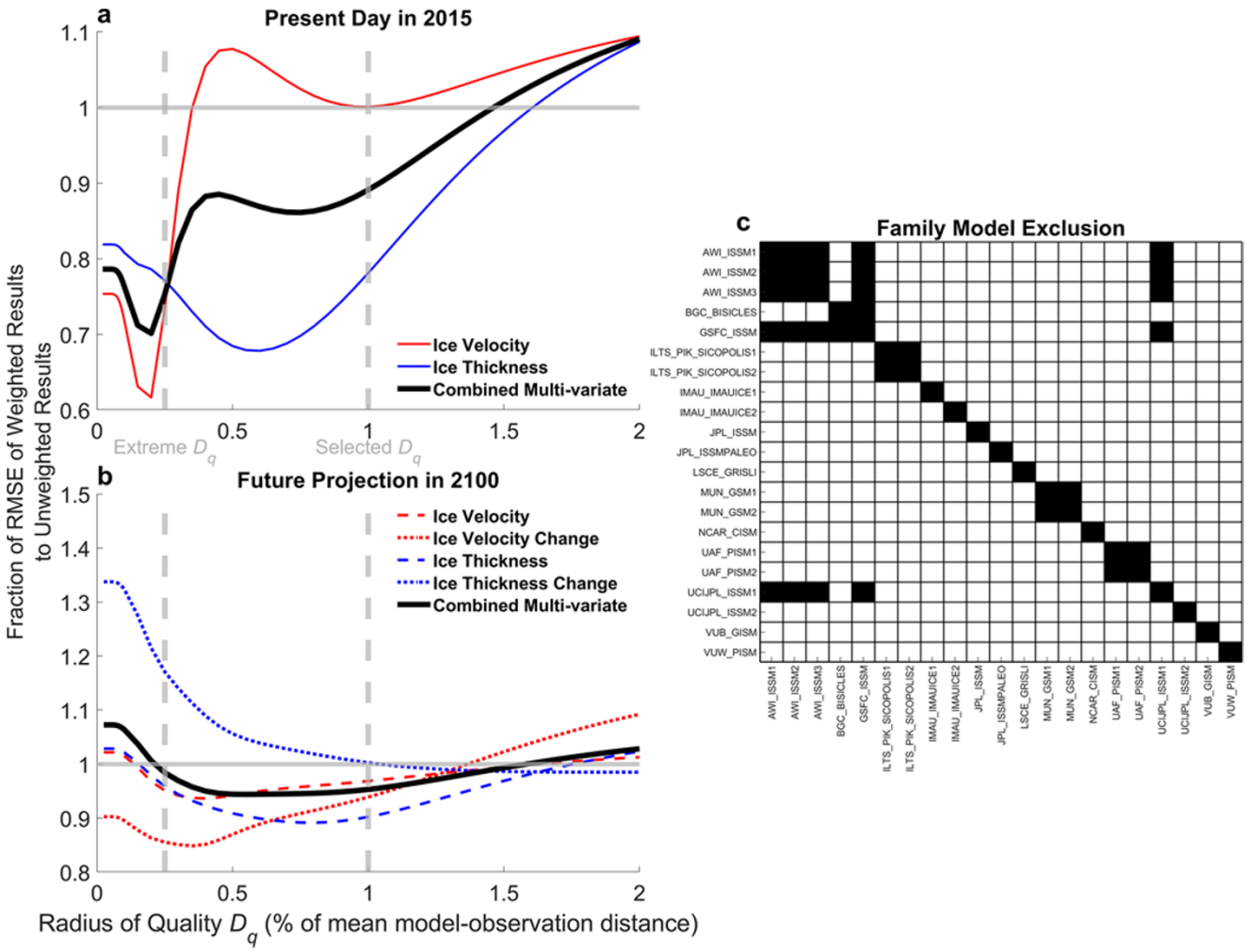

Figure 6. The fraction of RMSE of weighted to unweighted results with a varying radius of quality D_q measured as percentiles of mean model–observation distances using (a) the data in 2015 and (b) 'exp05' experiment in 2100. The magnitude of ice velocity and thickness (dashed curves) and their changes from 2015 to 2100 (dotted curves) as well as the mean of the above (solid curve) are shown in (b). (c) The models that are excluded from the out-of-sample test in (b) when each model is treated as the truth. The black cells in (c) represent the excluded family models for each model. Finally, note that (a) is constructed using the initial state in 2015 of 'ctrl_proj' only, while (b) is using the last year in the 'exp05' projection, which is 2100.

weights (x-axes in Fig. 7). SICOPOLIS, IMAUICE and GSM models, each having two submissions, also show little differences in model weights (for each pair), indicating each pair of submissions is quite similar. The PISM model simulated by the VUW group is clearly distinguishable from UAF_PISM1 and UAF_PISM2.

The same model may also receive different weights when either velocity or ice thickness is used to measure their performance. For instance, PISM submissions from UAF receive high weights for thickness but low weights for velocity. These two models use long inter-glacial spin-ups and keep the ice surface close to observation using a flux correction method (Aschwanden and others, 2016). This approach results in a close similarity between the simulated thickness and the observed thickness, while this is not the case for the velocity. This highlights the need to use both velocity and thickness to measure the model performance instead of univariate as some models do not have equal performance regarding different variables. In contrast, VUW_PISM receives low weights for both velocity and thickness because it uses a different initialization method than the submissions from UAF. VUW_PISM did not use flux correction, leading to its lower thickness weights. We notice that JPL_ISSM, JPL_ISSMPALEO, MUN_GSM1 and MUN_GSM2 receive very low weights for quality weighting using ice thickness. However, this does not mean they differ drastically from the observed thickness field over the Greenland ice sheet (Fig. 2), but they are comparatively less close to the observation than other models. This is by design of Sanderson's method (Eq. (1)). These models used long interglacial spin-up that leads to less constrained ice geometry, and MUN_GSM1 and MUN_GSM2 used different bedrock (Bamber and others, 2001) compared to the other models that used BedMachine (Morlighem and others, 2017). Finally, we note that the models simulated by different groups may or may not be similar to each other, indicating it is worthwhile to treat each submission as an independent model.

3.4. Weighted ice-sheet projections

The weights shown earlier in Section 3.3 are then used to produce the weighted sea level projections by the ISMIP6-Greenland models (Table 2 and Figs 8–11). We pick the experiments shown in figure 12 of Goelzer and others (2020) to demonstrate the weighting

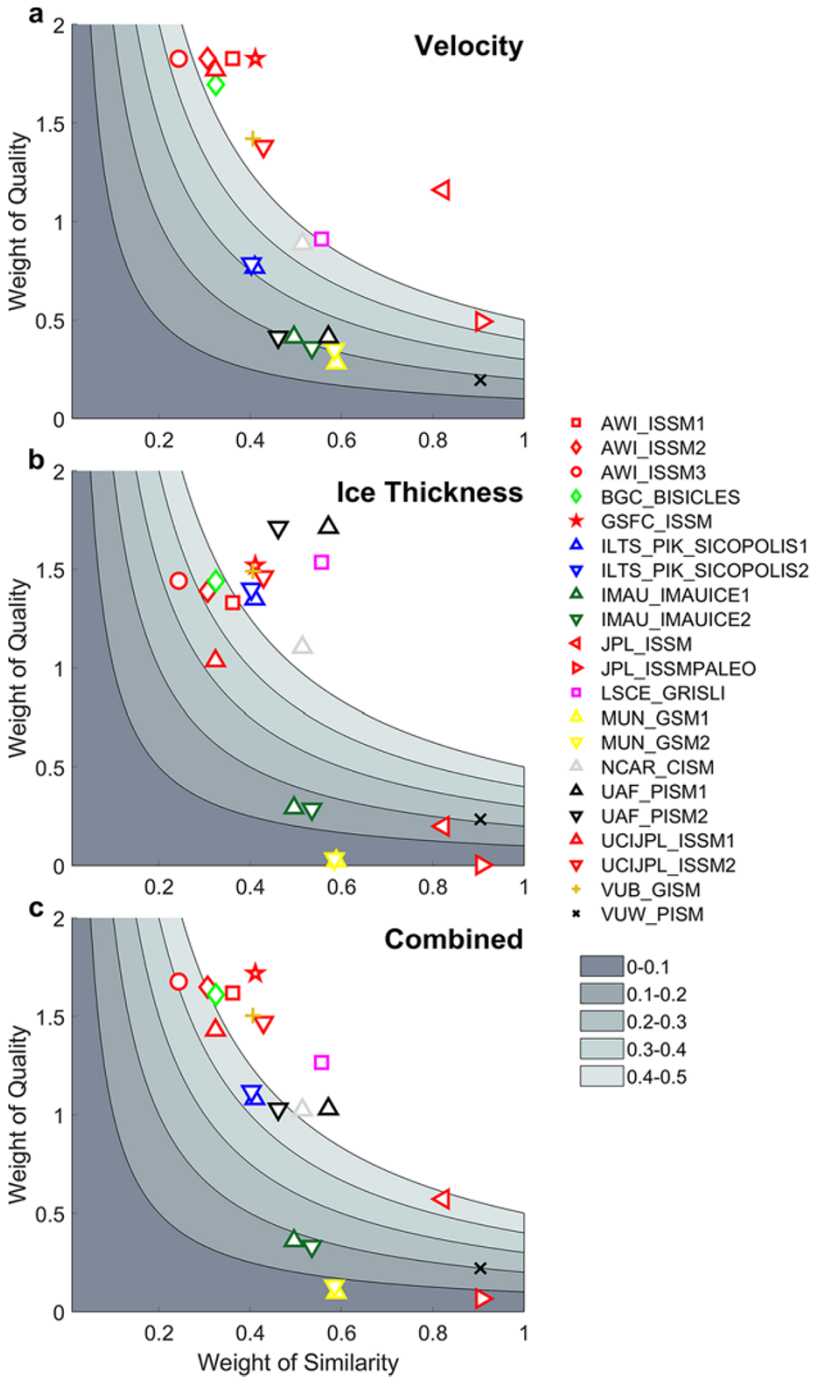

Figure 7. Results of ISMIP6-Greenland model weights of similarity (*x*-axis), weights of quality (*y*-axis) and total weights (indicated with the shaded area) where the input for quality weighting are (a) ice velocity only, (b) ice thickness only and (c) both ice velocity and thickness. The legends show the markers for each model. The same color is used for the same model variants; for example, red color for all ISSMs.

effects on the final sea level rise projections. In Table 2, we define the weighted ensemble mean as $\mu_{SLR} = \sum w_i \times SLR_i / \sum w_i$, and the weighted ensemble std dev. is defined as $\sigma_{SLR} = \sqrt{\sum_{i=1}^{N} w_i \times (SLR_i - \mu_{SLR})^2 / \sum_{i=1}^{N} w_i}$. For exploration purposes, we show the results with varying choices of weights in Table 2 including the total combined weights $w_i = w_q \times w_u$, quality weights only w_q, similarity weights only w_u, velocity weights only $w_i = w_{q(vel)} \times w_u$ and thickness weights only $w_i = w_{q(thickness)} \times w_u$.

We find that the multi-ensemble mean of sea level rise projections does not deviate much from the equal weighting case (mostly within $\pm$1cm), indicating that the weighting has minimal effects on the ensemble mean. However, we notice that the model weights decrease the std dev. values (the 'std dev. change' rows in Table 2). This decreasing effect varies among different experiments, ranging from minor effects (around −1%) to moderate decreases (around −30%). Using similarity weights only, the std dev. values are mostly increased, which can be anticipated because the similarity weighting is supposed to highlight the unique models, whose sea level projections may deviate more from the majority of the ensemble than others. For all other types of weights, the weighting effects have similar influences on the std dev. decrease.

We also note that the weighted ensemble std dev. is a simple statistical metric that does not take the distributions of original sea level projections into account. Therefore, it does not directly

Table 2. Weighted multi-model ensemble statistics of the chosen ISMIP6 experiments using various types of model weights

Weight		Experiments								
		exp05	exp06	exp08	expa01	expa02	expa03	exp09	exp10	exp07
Type	Statistics	MIROC5 Med RCP8.5	NorESM-M Med RCP8.5	HadGEM2-ES Med RCP8.5	IPSL-CM5A-MR Med RCP8.5	CSIRO-Mk3.6 Med RCP8.5	ACCESS1–3 Med RCP8.5	MIROC5 High RCP8.5	MIROC5 Low RCP8.5	MIROC5 High RCP2.6
Equal weight	Mean (cm)	10.14	6.92	8.28	7.70	4.43	5.59	9.96	8.35	3.17
	Std dev. (cm)	1.93	1.84	1.78	5.03	3.05	3.73	4.62	3.71	0.83
	Std dev. change	0	0	0	0	0	0	0	0	0
Total weight	Mean (cm)	10.27	7.07	8.30	9.43	5.41	6.82	10.07	8.35	3.28
	Std dev. (cm)	1.58	1.56	1.45	3.51	2.32	2.69	4.53	3.67	0.64
	Std dev. change	−18.3%	−15.2%	−18.8%	−30.3%	−24.0%	−27.8%	−2.0%	−1.1%	−22.8%
Quality only	Mean (cm)	10.39	7.17	8.39	9.46	5.47	6.87	10.38	8.59	3.34
	Std dev. (cm)	1.56	1.53	1.46	3.59	2.38	2.75	4.35	3.50	0.62
	Std dev. change	−19.6%	−16.7%	−18.3%	−28.6%	−22.2%	−26.3%	−5.8%	−5.5%	−25.3%
Similarity only	Mean (cm)	9.82	6.66	8.02	7.38	4.23	5.34	9.20	7.75	3.02
	Std dev. (cm)	1.96	1.84	1.74	5.07	3.06	3.75	4.95	4.03	0.91
	Std dev. change	1.1%	0	−2.3%	0.8%	0.1%	0.5%	7.1%	8.7%	9.2%
Velocity only	Mean (cm)	10.26	7.05	8.28	9.05	5.25	6.57	10.41	8.73	3.24
	Std dev. (cm)	1.64	1.61	1.50	4.05	2.56	3.04	3.99	3.22	0.73
	Std dev. change	−15.1%	−12.7%	−15.8%	−19.5%	−16.0%	−18.4%	−13.7%	−13.2%	−11.8%
Thickness only	Mean (cm)	10.18	6.98	8.23	9.50	5.40	6.85	9.65	7.95	3.27
	Std dev. (cm)	1.58	1.58	1.46	3.26	2.22	2.54	4.91	3.97	0.62
	Std dev. change	−18.2%	−14.0%	−18.3%	−35.1%	−27.2%	−31.8%	6.2%	7.1%	−25.1%

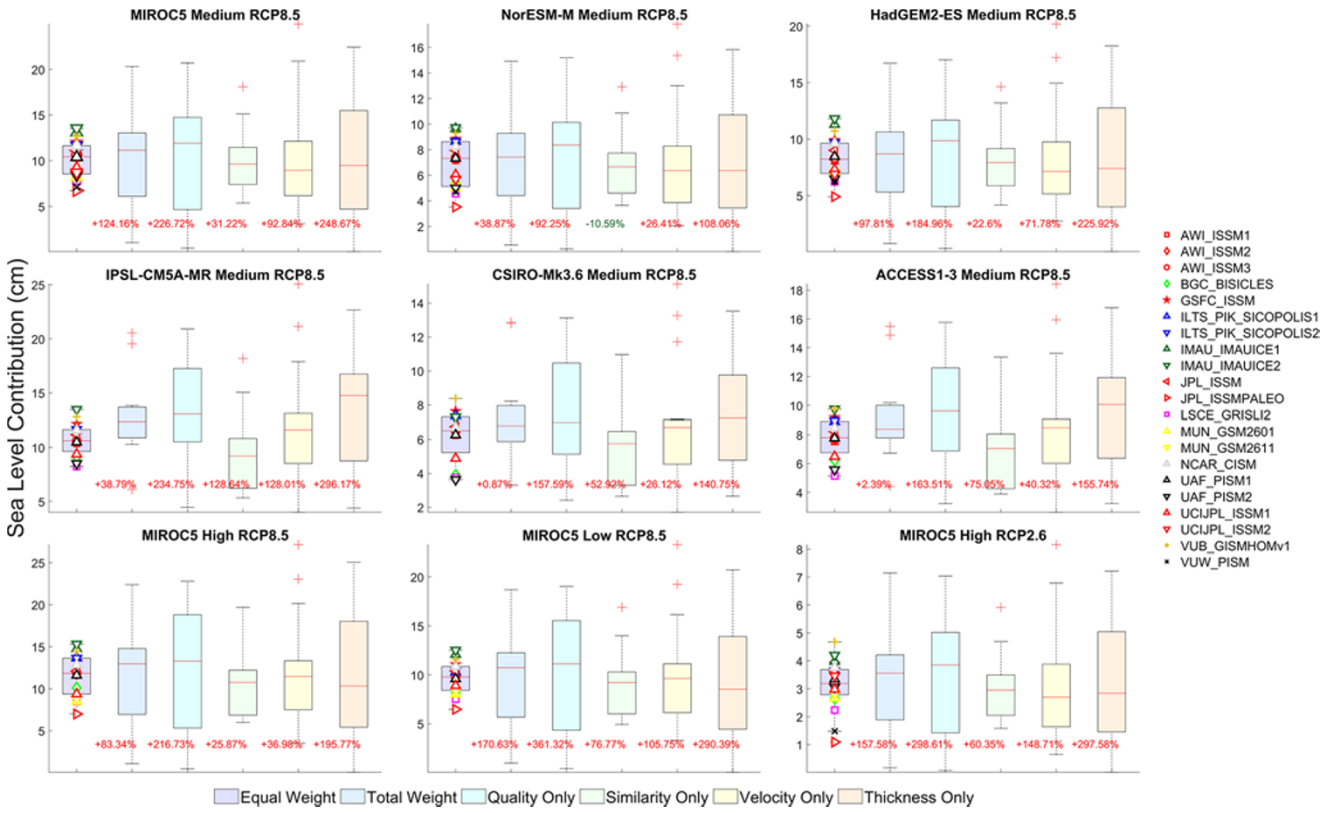

Figure 8. Boxplots of the updated ISMIP6-Greenland projections in 2100 using varying weighting types including equal weights (original simulations), total weights, quality weights alone, similarity weights alone, velocity weights alone and thickness weights alone. The original ISMIP6 projections are marked only on the first boxplot of each experiment. The text below each boxplot shows the reduction/increase of the model spread. The types of weighting schemes are indicated by the legends at the bottom.

translate to model spread (the interquartile range). In order to explore how much the weights modify model spreads and shift data distributions, we use four different ways of applying the model weights on the ISMIP6-Greenland projections: direct approach (Fig. 8), Monte Carlo sampling (Fig. 9), bootstrap resampling (Fig. 10) and kernel density estimation (KDE) using Gaussian kernel (Fig. 11). For the direct approach, the sea level projections are multiplied by the model weights in a straightforward manner, i.e. $SLE_{weighted} = SLE_{original} \times w_i$, where each model weight w_i is scaled up such that $\sum w_i$ equals to the number of models ($N = 21$). For the Monte Carlo sampling, we collect 2000 random samples with each sample randomly drawn from the original sea level rise projection and set the probability of each model equal to its weight. The statistics of both the direct approach and the Monte Carlo approach are then summarized by the boxplots in Figs 8 and 9. Bootstrap resampling collects 2000 samples as well, but each of them is randomly drawn with replacements that are equal to the size of the original samples ($N = 21$), in contrast to the Monte Carlo approach which draws only one value in each sample. We estimate the mean of sea level rise projections from the bootstrap samples and plot the probability density functions (PDFs) in Fig. 10. Finally, KDE builds directly upon original projections calibrated with the model weights, because it adjusts the peak and std dev. values of the Gaussian kernels locally around each 'data point' (in this case, each ISMIP6 projection). For instance, if a model has a larger weight, then larger confidence is implied for this value; therefore, the Gaussian kernel of this point becomes taller and slimmer. We show the PDFs of KDE results in Fig. 11, and the original ISMIP6 projections are marked on the x-axis as well.

In Figs 8 and 9, we simply define the model spread as the interquartile range (i.e. the middle 50 percentiles), which is the length of the box. The outliers are the models that deviate more than 1.5 times the interquartile range from the ensemble mean, and these are marked as red crosses. The changes in model spreads are shown by the text near each box, measured as a percentage of the original model spread. Note that the y-axes have different scales so that they align with the total range of their original ISMIP6 simulations in Fig. 9. For the direct approach, although the multi-model means have minor shifts, it gives much larger model spreads after applying the model weights as the original sea level rise projections are scaled either up or down by multiplying the weights. This increase is around two times larger for all weighting types except for the similarity weighting only. In contrast, the Monte Carlo approach gives mostly a reduction of the model spread. We observe that, by using total weights, most experiments have a decrease in the model spread, with the magnitude of reduction ranging from zero (e.g. CSIRO-Mk3.6 Medium RCP8.5) to moderate (−19.19% for MIROC5 Low RCP8.5) values. The quality weights mostly have similar effects as the total weights. Under similar weighting only, the model spreads are moderately increased for some experiments, which can also be observed in Table 2 due to the same reasoning above. We explore the impacts of univariate weighting as well. We find that the velocity weighting reduces model spreads for all experiments, while the effects of ice thickness on weighted ensemble are rather diverse among experiments. Ice thickness weighting

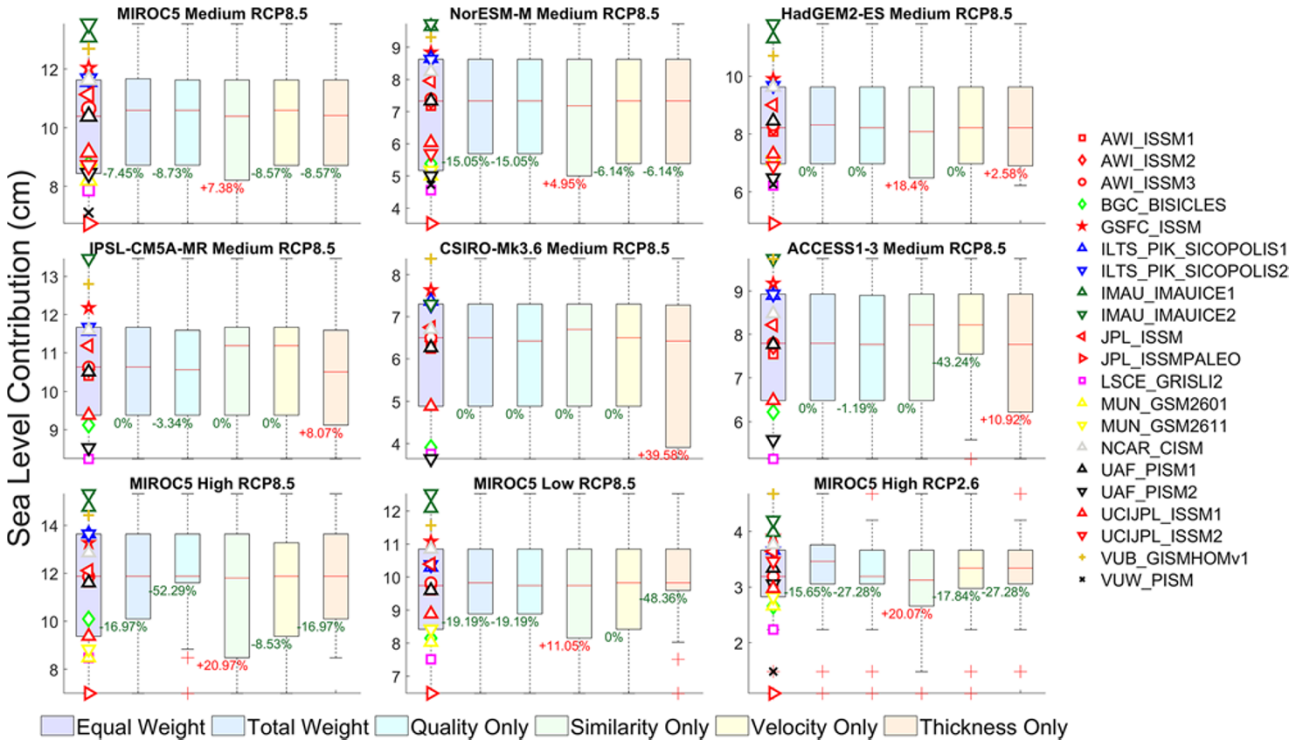

Figure 9. Boxplots of the updated ISMIP6-Greenland projections in 2100 using varying weighting types including equal weights (original simulations), total weights, quality weights alone, similarity weights alone, velocity weights alone and thickness weights alone. Figure 9 is similar to Fig. 8 but uses the Monte Carlo sampling approach. The original ISMIP6 projections are marked only on the first boxplot of each experiment. The text below each boxplot shows the reduction/increase of the model spread. The types of weighting schemes are indicated by the legends at the bottom.

alone does not always reduce the model spreads for all experiments, and a large increase in spread (39.58%) is observed for the CSIRO-Mk3.6 Medium RCP8.5 experiment. This increase is due to three models that received high weights (BGC_BICICLES, LSCE_GRISLI2 and UAF_PISM2), each of which generated lower-end sea level rise projections. This indicates the choice of using ice velocity alone to assign model weights based on present-day observation might be more optimal compared to univariate thickness weighting. Bootstrap resampling is used to estimate the distribution of sample means and the results are shown in Fig. 10. Note that this is not the distribution of sea level rise projections, but the estimation of the multi-model means. All weighting types except the similarity weighting scheme can reduce the variance of the distribution, which is similar to the results presented in Fig. 9. Finally, we use KDE to construct the distribution of sea level projections with weights (Fig. 11). The black thick curve shows the distribution of sea level rise projections, and all other curves show distributions with model weights. We observe that the distribution spreads are reduced for all Tier-1 core experiments including exp05, exp06, exp08 (the three experiments in the first row), exp09, exp10, and exp07 (the three experiments in the last row). In contrast, the Tier-2 experiments are less influenced by the choices of model weights.

In general, we conclude that the application of our model weights reduces the model spreads from minor to moderate levels depending on experiments, although it does not have major impacts on the multi-model mean. Finally, we note that the model spreads shown in Figs 8–11 are not directly comparable to the std dev. values in Table 2, as they are different metrics.

4. Conclusions and discussion

In this study, we have used the ClimWIP model weighting strategy to assign weights to ISMIP6-Greenland ice-sheet models to explore the influence on the sea level projections under model weights in contrast to the 'one model one vote' strategy, which was previously practiced in ISMIP6 literature. This model weighting strategy considers both model performance compared to observation and model inter-dependence among the ensemble model participants. We chose ice velocity and thickness of the initial ice-sheet state, which are the same diagnostics as in Goelzer and others (2020), to measure the model performance against observation. In contrast, we use as many ice-sheet model variables as we can to assign the independence weights. Furthermore, we consider both the initial states and future projections to measure the independence weights. The motivation is that the models having similar initial states and multiple submissions (such as UAF_PISM) may respond differently to climate forcing based on their implementation of ice-ocean interaction and other modeling options.

We also demonstrate the challenges of finding appropriate parameters involved in the model weighting scheme and how the choices of radius of quality (D_q) and similarity (D_u) can best facilitate the weighting. We select D_u as 0.5 times the mean of inter-model distance to effectively distinguish model differences, and we choose D_q as equal to the mean of inter-model distances so that the weighted ensemble shows decreased bias for both the present-day and future projections. For quality weighting, we found that the same model can have different performances when different diagnostics are used, which confirms the reasoning for using

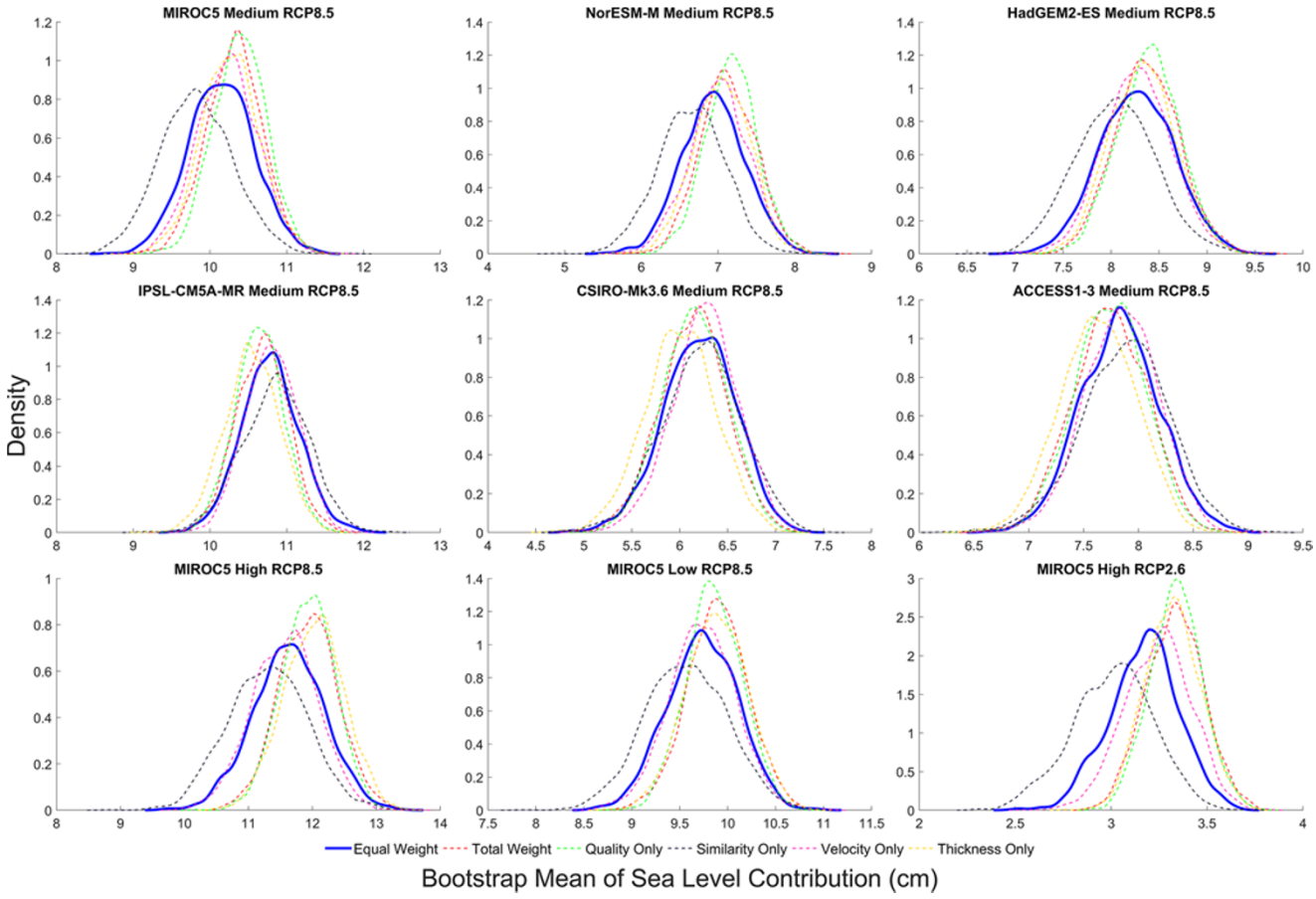

Figure 10. Distributions of bootstrap mean of ISMIP6-Greenland projections in 2100 using varying weighting types including equal weights (original simulations), total weights, quality weights alone, similarity weights alone, velocity weights alone and thickness weights alone. The types of weighting schemes are indicated by the legends at the bottom.

more than one single diagnostic. For example, UAF_PISM1 and PISM2 perform differently when only one variable is used (Fig. 7a and 7b), and the combined weights provide more balanced scores than the single variable scores. In the final model weighting results (Fig. 7c), most models receive weights ranging from 0.3 to 0.5.

The model weights are then utilized to update the projections of the six Tier-1 experiments plus three Tier-2 experiments. We do not observe large shifts of multi-model ensemble mean (mostly within $\pm$1cm), but the model spreads are indeed reduced to varying extents depending on the experiments. The simple weighted std dev. values indicate around 10–30% of model spread reduction. We also explore four different ways to apply the model weights on the projections, including the direct approach, Monte Carlo approach, Bootstrap mean approach and KDE approach, to find the impacts on the distribution of projections. With the exception of the direct approach, which increases all model spreads, the other approaches generally reduce model spreads, yet the magnitude of reduction varies considerably among experiments and types of model weights applied.

One limitation of this study is that we do not perform the tests regarding the choice of observation and correlation of diagnostics of ice-sheet models with sea level rise projections and explore the influences of the types of diagnostics. As an exploratory study of assigning model weights on ISMIP6-Greeland models, we limit our focus merely to the same metrics used in Goelzer and others (2020). The Bayesian calibrations by using velocity change, dynamic ice thickness change and mass change observations show very different posterior sea level rise distributions in Felikson and others (2023). Also, this study focuses on ISMIP6-Greenland model weighting, and the same practice may be used on ISMIP6-Antarctica (Seroussi and others, 2020) and ISMIP6-Antarctica-2300 in future research.

Another limitation arises from not exploring other model weighting schemes. For similarity weighting, the ClimWIP scheme stands on a data-driven point of view, that is, whether a pair of models are considered similar is judged by their initial states and simulation results. Other methods may be used to sort the models, such as developing family model genealogy as recently practiced in the climate modeling community for CMIP6 models (Kuma and others, 2023).

We also note that the refinement of the updated sea level distribution is not as considerable as the ones shown in the Bayesian calibration studies. However, they are not comparable due to the size of the ensemble. The size of the ensemble considered in Bayesian calibration tends to be much bigger (from several hundred to thousands) in contrast to ISMIP6-Greenland models (21 models). This is because the ISMIP6-Greenland models generally submitted one realization (at most three realizations) for the same model. In contrast, the Bayesian calibration studies were designed to explore the uncertainties involved in the whole parameter space, resulting in a large number of perturbed ensemble members branching from

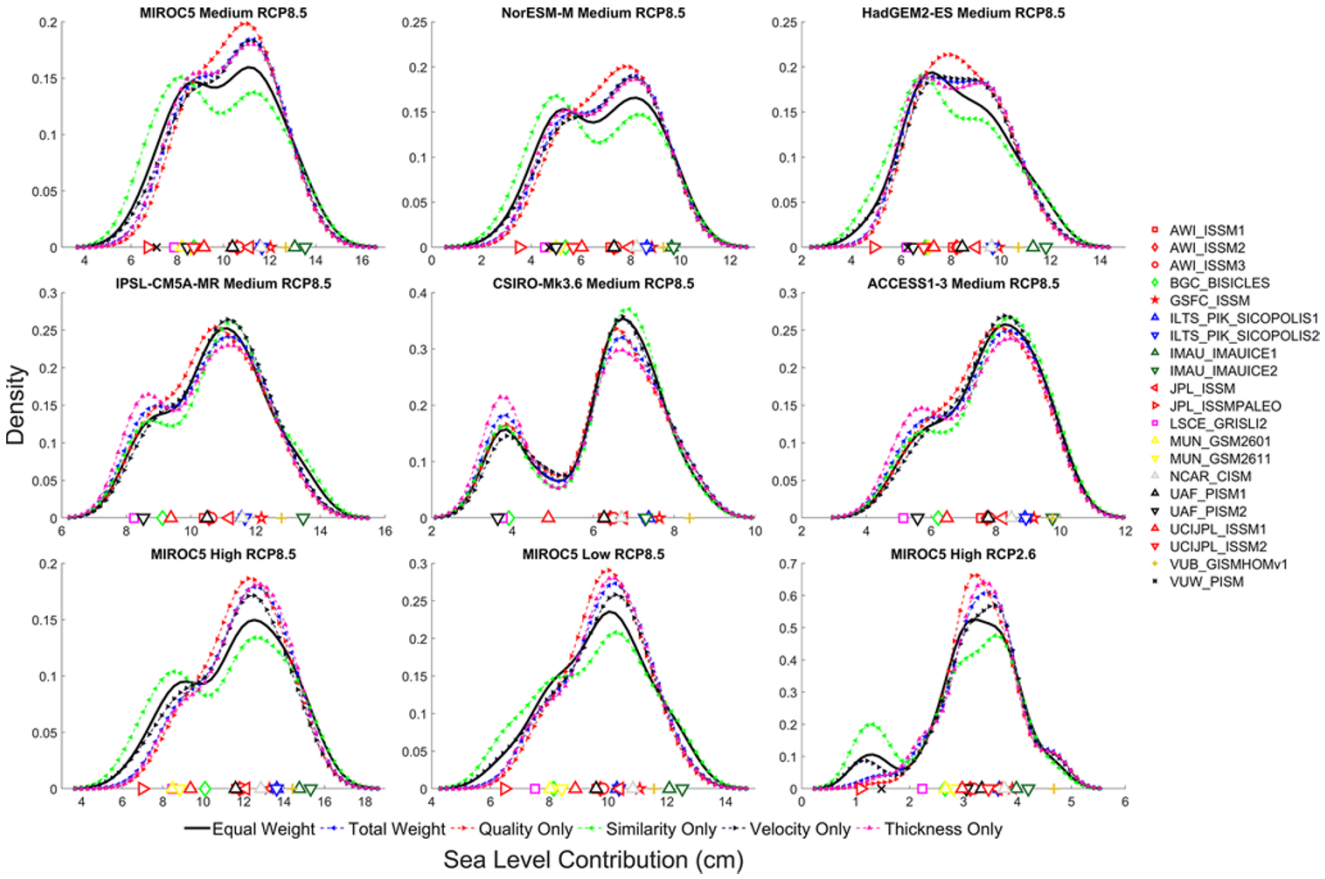

Figure 11. Distributions of ISMIP6-Greenland projections in 2100 KDE using varying weighting types including equal weights (original simulations), total weights, quality weights alone, similarity weights alone, velocity weights alone and thickness weights alone. The types of weighting schemes are indicated by the legends at the bottom. The original ISMIP6 projections are marked on the *x*-axis as well.

the same model. This makes their prior distributions of probabilistic sea level rise (before Bayesian updating) relatively flat and the posterior much sharper.

In conclusion, we show in this study that the ClimWIP scheme is skillful in producing model weights that effectively and reasonably quantify the model performance and inter-dependency. The resulting projections show mild to medium levels of decreased model spreads compared to the unweighted ensemble, although the multi-model means do not show considerable shifts. This highlights the potential of applying model weights to reduce ensemble spreads for ice-sheet intercomparison projects, given that the next phase, i.e. ISMIP7, may include a bigger size of model submissions and experiments that can lead to larger ensemble uncertainties.

Data availability statement. The ISMIP6 output data is accessible at https://theghub.org/. The BedMachine data are freely available at https://nsidc.org/data/nsidc-0756/versions/3. The MEaSUREs Greenland Ice Sheet Velocity Map from InSAR Data, Version 2, is freely available at https://nsidc.org/data/nsidc-0478/versions/2. The scripts used to generate the figures in this paper are available at https://doi.org/10.5281/zenodo.15265911.

Acknowledgements. We acknowledge the color schemes from Crameri and others (2020), which are used in multiple figures in this paper. We thank the Climate and Cryosphere (CliC) effort, which provided support for ISMIP6 through sponsoring workshops, hosting the ISMIP6 website and wiki and promoting ISMIP6. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modeling, coordinated and promoted CMIP5 and CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the CMIP data and providing access, the University at Buffalo for ISMIP6 data distribution and upload and the multiple funding agencies who support CMIP5 and CMIP6 and ESGF. We thank the ISMIP6 steering committee, the ISMIP6 model selection group and ISMIP6 data set preparation group for their continuous engagement in defining ISMIP6. This is ISMIP6 contribution No. 35.

Funding statement. This research was supported by grants from the NASA Sea Level Change Team (80NSSC24K1545) and Modeling, Analysis and Predictions Program (80NSSC25K7094).

Competing interests. The authors declare that they have no conflict of interest.

References

Aschwanden A and Brinkerhoff DJ (2022) Calibrated mass loss predictions for the Greenland ice sheet. *Geophysical Research Letters* **49**(19), e2022GL099058. doi: 10.1029/2022GL099058.

Aschwanden A, Fahnestock MA and Truffer M (2016) Complex Greenland outlet glacier flow captured. *Nature Communications* **7**(1), 10524. doi: 10.1038/ncomms10524.

Bamber JL, Layberry RL and Gogineni SP (2001) A new ice thickness and bed data set for the Greenland ice sheet: 1. Measurement, data reduction, and errors. *Journal of Geophysical Research: Atmospheres* **106**(D24), 33773–33780. doi: 10.1029/2001JD900054.

Bindschadler RA and 27 others (2013) Ice-sheet model sensitivities to environmental forcing and their use in projecting future sea level (the

SeaRISE project). *Journal of Glaciology* **59**(214), 195–224. doi: 10.3189/2013JoG12J125.

Brinkerhoff D, Aschwanden A and Fahnestock M (2021) Constraining subglacial processes from surface velocity observations using surrogate-based Bayesian inference. *Journal of Glaciology* **67**(263), 385–403. doi: 10.1017/jog.2020.112.

Brunner L, Lorenz R, Zumwald M and Knutti R (2019) Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environmental Research Letters* **14**(12), 124010. doi: 10.1088/1748-9326/ab492f.

Brunner L, Pendergrass AG, Lehner F, Merrifield AL, Lorenz R and Knutti R (2020) Reduced global warming from CMIP6 projections when weighting models by performance and Independence. *Earth System Dynamics* **11**(4), 995–1012. doi: 10.5194/esd-11-995-2020.

Crameri F, Shephard GE and Heron PJ (2020) The misuse of colour in science communication. *Nature Communications* **11**(1), 5444. doi: 10.1038/s41467-020-19160-7.

DeConto RM and Pollard D (2016) Contribution of Antarctica to past and future sea-level rise. *Nature* **531**(7596), 591–597. doi: 10.1038/nature17145.

Felikson D and 6 others (2023) Choice of observation type affects Bayesian calibration of Greenland Ice Sheet model simulations. *The Cryosphere* **17**(11), 4661–4673. doi: 10.5194/tc-17-4661-2023.

Fox-Kemper B and 17 others (2021) *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*. Cambridge and New York, NY: Cambridge University Press. 1211–1362.

Gladstone RM and 9 others (2012) Calibrated prediction of Pine Island Glacier retreat during the 21st and 22nd centuries with a coupled flowline model. *Earth and Planetary Science Letters* **333–334**, 191–199. doi: 10.1016/j.epsl.2012.04.022.

Goelzer H and 30 others (2018) Design and results of the ice sheet model initialisation experiments initMIP-Greenland: An ISMIP6 intercomparison. *The Cryosphere* **12**(4), 1433–1460. doi: 10.5194/tc-12-1433-2018.

Goelzer H and 41 others (2020) The future sea-level contribution of the Greenland ice sheet: A multi-model ensemble study of ISMIP6. *The Cryosphere* **14**(9), 3071–3096. doi: 10.5194/tc-14-3071-2020.

Hofer S and 6 others (2020) Greater Greenland Ice Sheet contribution to global sea level rise in CMIP6. *Nature Communications* **11**(1), 6289. doi: 10.1038/s41467-020-20011-8.

Jager E, Gillet-Chaulet F, Champollion N, Millan R, Goelzer H and Mouginot J (2024) The future of Upernavik Isstrøm through ISMIP6 framework: Sensitivity analysis and Bayesian calibration of ensemble prediction. *The Cryosphere* **18**(11), 5519–5550. doi: 10.5194/tc-18-5519-2024.

Joughin I, Smith H and Scambos T (2015) *MEaSUREs Greenland Ice Sheet Velocity Map from InSAR Data, Version 2. Data Set*. Boulder, CO: NASA National Snow and Ice Data Center Distributed Active Archive Center. doi: 10.5067/OC7B04ZM9G6Q.

Knutti R (2010) The end of model democracy? *Climatic Change* **102**(3), 395–404. doi: 10.1007/s10584-010-9800-2.

Knutti R, Furrer R, Tebaldi C, Cermak J and Meehl GA (2010) Challenges in combining projections from multiple climate models. *Journal of Climate* **23**(10), 2739–2758. doi: 10.1175/2009JCLI3361.1.

Knutti R, Sedláček J, Sanderson BM, Lorenz R, Fischer EM and Eyring V (2017) A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters* **44**(4), 1909–1918. doi: 10.1002/2016GL072012.

Kuma P, Bender FA-M and Jönsson AR (2023) Climate model code genealogy and its relation to climate feedbacks and sensitivity. *Journal of Advances in Modeling Earth Systems* **15**(7), e2022MS003588. doi: 10.1029/2022MS003588.

Larour E, Seroussi H, Morlighem M and Rignot E (2012) Continental scale, high order, high spatial resolution, ice sheet modeling using the Ice Sheet System Model (ISSM). *Journal of Geophysical Research: Earth Surface*, **117**(F1). doi: 10.1029/2011JF002140.

Lorenz R, Herger N, Sedláček J, Eyring V, Fischer EM and Knutti R (2018) Prospects and caveats of weighting climate models for summer maximum temperature projections over North America. *Journal of Geophysical Research: Atmospheres* **123**(9), 4509–4526. doi: 10.1029/2017JD027992.

Masson D and Knutti R (2011) Climate model genealogy. *Geophysical Research Letters* **38**(8). doi: 10.1029/2011GL046864.

Merrifield AL, Brunner L, Lorenz R, Medhaug I and Knutti R (2020) An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles. *Earth System Dynamics* **11**(3), 807–834. doi: 10.5194/esd-11-807-2020.

Morlighem M and 31 others (2017) BedMachine v3: Complete bed topography and ocean bathymetry mapping of Greenland from multibeam echo sounding combined with mass conservation. *Geophysical Research Letters* **44**(21), 11051–11061. doi: 10.1002/2017GL074954.

Morlighem M and 36 others (2020) Deep glacial troughs and stabilizing ridges unveiled beneath the margins of the Antarctic ice sheet. *Nature Geoscience* **13**(2), 132–137. doi: 10.1038/s41561-019-0510-8.

Nias IJ, Cornford SL, Edwards TL, Gourmelen N and Payne AJ (2019) Assessing uncertainty in the dynamical ice response to ocean warming in the Amundsen Sea Embayment, West Antarctica. *Geophysical Research Letters* **46**(20), 11253–11260. doi: 10.1029/2019GL084941.

Nias IJ, Nowicki S, Felikson D and Loomis B (2023) Modeling the Greenland ice sheet's committed contribution to sea level during the 21st century. *Journal of Geophysical Research: Earth Surface* **128**(2), e2022JF006914. doi: 10.1029/2022JF006914.

Nowicki S and 30 others (2013) Insights into spatial sensitivities of ice mass response to environmental change from the SeaRISE ice sheet modeling project II: Greenland. *Journal of Geophysical Research: Earth Surface* **118**(2), 1025–1044. doi: 10.1002/jgrf.20076.

Nowicki S and 8 others (2016) Ice Sheet Model Intercomparison Project (ISMIP6) contribution to CMIP6. *Geosci. Model Dev.*, **9**(12), 4521–4545. doi: 10.5194/gmd-9-4521-2016.

Nowicki S and 29 others (2020) Experimental protocol for sea level projections from ISMIP6 stand-alone ice sheet models. *The Cryosphere* **14**(7), 2331–2368. doi: 10.5194/tc-14-2331-2020.

O'Loughlin R (2024) Why we need lower-performance climate models. *Climatic Change* **177**(2), 21. doi: 10.1007/s10584-023-03661-7.

Payne AJ and 63 others (2021) Future sea level change under coupled model intercomparison project Phase 5 and Phase 6 scenarios from the Greenland and Antarctic ice sheets. *Geophysical Research Letters* **48**(16), e2020GL091741. doi: 10.1029/2020GL091741.

Pennell C and Reichler T (2011) On the effective number of climate models. *Journal of Climate* **24**(9), 2358–2367. doi: 10.1175/2010JCLI3814.1.

Rignot E and Mouginot J (2012) Ice flow in Greenland for the International Polar Year 2008–2009. *Geophysical Research Letters* **39**(11), doi: 10.1029/2012GL051634.

Ritz C, Edwards TL, Durand G, Payne AJ, Peyaud V and Hindmarsh RCA (2015) Potential sea-level rise from Antarctic ice-sheet instability constrained by observations. *Nature* **528**(7580), 115–118. doi: 10.1038/nature16147.

Rückamp M, Goelzer H and Humbert A (2020) Sensitivity of Greenland ice sheet projections to spatial resolution in higher-order simulations: The Alfred Wegener Institute (AWI) contribution to ISMIP6 Greenland using the ice-sheet and sea-level system model (ISSM). *The Cryosphere* **14**(10), 3309–3327. doi: 10.5194/tc-14-3309-2020.

Sanderson BM, Knutti R and Caldwell P (2015) A representative democracy to reduce interdependency in a multimodel ensemble. *Journal of Climate* **28**(13), 5171–5194. doi: 10.1175/JCLI-D-14-00362.1.

Sanderson BM, Wehner M and Knutti R (2017) Skill and independence weighting for multi-model assessments. *Geoscientific Model Development* **10**(6), 2379–2395. doi: 10.5194/gmd-10-2379-2017.

Seroussi H and 46 others (2020) ISMIP6 Antarctica: A multi-model ensemble of the Antarctic ice sheet evolution over the 21st century. *The Cryosphere* **14**(9), 3033–3070. doi: 10.5194/tc-14-3033-2020.

Shepherd A and 89 others (2020) Mass balance of the Greenland Ice Sheet from 1992 to 2018. *Nature* **579**(7798), 233–239. doi: 10.1038/s41586-019-1855-2.

Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres* **106**(D7), 7183–7192. doi: 10.1029/2000JD900719.