ISEFlow: A Flow-Based Neural Network Emulator for Improved Sea Level Projections and Uncertainty Quantification

Peter Van Katwyk[1,2,3], Baylor Fox-Kemper[1,3], Sophie Nowicki[4], Hélène Seroussi[5], and Karianne J. Bergen[1,2]

[1]Department of Earth, Environmental, and Planetary Sciences, Brown University, Providence, RI, USA
[2]Data Science Institute, Brown University, Providence, RI, USA
[3]Institute at Brown for Environment and Society, Brown University, Providence, RI, USA
[4]Department of Geology, University at Buffalo, Buffalo, NY, USA
[5]Thayer School of Engineering, Dartmouth College, Hanover, NH, USA

Correspondence: Peter Van Katwyk (peter_van_katwyk@brown.edu)

Abstract. Ice sheets are the primary contributors to global sea level rise, yet projecting their future contributions remains challenging due to the complex, nonlinear processes governing their dynamics and uncertainties in future climate scenarios. This study introduces ISEFlow, a neural network-based emulator of the ISMIP6 ice sheet model ensemble designed to accurately and efficiently ~~predict~~ emulate sea level contributions from both ice sheets while quantifying the sources of projection uncertainty. By integrating a normalizing flow architecture to capture data coverage uncertainty and a deep ensemble of ~~LSTM~~ recurrent neural network models to assess emulator uncertainty, ISEFlow separates uncertainties arising from training data from those inherent to the emulator. Compared to existing emulators such as ~~Emulandice~~ a Gaussian Process-based emulator modeled after emulandice and LARMIP, ISEFlow achieves substantially lower mean squared error and improved distribution approximation while maintaining faster ~~inference~~ prediction times. This study investigates the drivers of increased emulation accuracy and emission scenario distinction and finds that the inclusion of all available climate forcings, ice sheet model characteristics, and higher spatial resolution significantly enhances ~~predictive~~ emulation accuracy and the ability to capture the effects of varying emissions scenarios compared to other emulators. We include a detailed analysis of importance of input variables using Shapley Additive Explanations, and highlight ~~both the~~ which climate forcings and model characteristics ~~that~~ have the largest impact on sea level projections. ISEFlow offers ~~a computationally efficient tool for generating accurate~~ an additional tool for efficient, accurate exploration of future sea level projections~~, supporting climate risk assessments and informing policy decisions~~.

1 Introduction

Sea level rise is a critical worldwide concern, with ice sheets being the dominant contributor to global sea level rise (Shepherd et al., 2018; Bamber et al., 2019; Rignot et al., 2019). Despite extensive research over the past decades (Pattyn et al., 2017; Goelzer et al., 2017), uncertainty persists regarding the future contributions of the Antarctic Ice Sheet (AIS) and Greenland Ice Sheet (GrIS) (Fox-Kemper et al., 2021; Oppenheimer et al., 2019; Kopp et al., 2023). This uncertainty largely stems from

the complex, nonlinear processes governing ice sheet dynamics ~~, as well as~~ and uncertainty in future emission scenarios, as well as the deep uncertainty due to Marine Ice Cliff Instability (MICI) (DeConto and Pollard, 2016; Golledge et al., 2015). To address these challenges, ice sheet models (ISMs), such as those that took part in the Ice Sheet Model Intercomparison Project (ISMIP), are essential for understanding future sea level rise (Nowicki et al., 2016, 2020; Seroussi et al., 2020; Goelzer et al., 2020).

Given the computational demands and inherent complexities of running full-scale ~~climate~~ ice sheet simulations, emulators have emerged as a valuable tool for producing ~~climate~~ ice sheet projections efficiently and accurately (Van Katwyk et al., 2023; Edwards et al., 2021). Emulators (also commonly referred to as "surrogate models") are ~~small-scale~~ computationally efficient models that approximate ~~large~~computationally expensive, more complex physical models (or simulation models). Emulators differ from data-driven parameterizations that are designed to operate within a ~~climate~~ physical model to represent a key process (e.g., Sane et al., 2023): they are typically designed to be run independently and represent either the whole ~~climate~~ physical system (e.g., Nicklas et al., 2025; Smith et al., 2018) or a major component of it (e.g., Lam et al., 2023; Bi et al., 2023; Kochkov et al., 2024), often to facilitate data assimilation and uncertainty quantification applications (e.g., Nicklas et al., 2025). Emulators use either data-driven methods or simplifications of the physical models to learn the mapping from simulation inputs to simulation outputs. They are typically designed to be more computationally efficient than the simulation model they approximate; they allow researchers to explore a broader range of model inputs and parameters, and enable more efficient experimentation and exploration than is typically feasible with the larger simulation models (Sacks et al., 1989; Reichstein et al., 2019; Beusch et al., 2020). In recent years, emulators have become increasingly popular in Earth and climate sciences due to their ability to rapidly generate large numbers of simulations, providing valuable insights with much lower computational costs compared to full-scale Earth system models (Bochenek and Ustrnul, 2022; Kashinath et al., 2021). This computational efficiency has made emulators an essential tool for exploring a wide range of scenarios and understanding impacts across different climate systems (de Burgh-Day and Leeuwenburg, 2023).

~~Development~~ The development of machine learning (ML) architectures has ~~significantly advanced emulator development by providing frameworks for accurately and efficiently capturing complex dynamics. Recent ML emulators like~~ further advanced emulator capabilities, with recent examples such as FourCastNet (Pathak et al., 2022) and GraphCast (Lam et al., 2023) ~~are able to produce high-resolution weather predictions with low computational demands. These models illustrate MLΓÇÖ~~demonstrating ML's ability to handle intricate spatial and temporal interactions in weather and climate data ~~, providing both speed and accuracy in weather forecasting. As ML-driven emulators continue to improve, they expand the possibilities for more accurate, responsive climate modeling and uncertainty quantification across scenarios~~ at low computational cost (Weyn et al., 2021).

Emulators played a pivotal role in the ice sheet assessments of the Intergovernmental Panel on Climate Change (IPCC) ~~6th~~ Sixth Assessment Report (AR6), where ISMIP6, an ensemble of ice sheet models that produce ~~land ice~~ ice sheet evolution and sea level projections ~~(Nowicki et al., 2020; Seroussi et al., 2020; Goelzer et al., 2020; Payne et al., 2021)~~(Nowicki et al., 2020; Seroussi et , and the ~~Emulandice~~ emulandice emulator (Edwards et al., 2021) were used alongside the Linear Antarctic Response Model Intercomparison Project (LARMIP) (Levermann et al., 2020) ~~to produce a comprehensive view~~ and structured expert judgment to produce an assessment of future ice sheet contribution to sea level based on future socioeconomic scenarios (including

emissions) (Fox-Kemper et al., 2021). These tools were instrumental in expanding projections to include a range of Shared Socioeconomic Pathways (SSPs), providing a more comprehensive sampling of climate, ice sheet, and glacier data. However, both the ~~Emulandice~~ emulandice emulator and the LARMIP linear response models ~~have inherent limitations due to their simplified representations~~ focus on different aspects of ice sheet ~~dynamics~~emulation. Emulandice, a Gaussian Process (GP)-based emulator, ~~relies on~~ utilizes global mean surface air temperature ~~and~~ (GSAT), which enables efficient coupling to simple climate models (e.g., FaIR; Smith et al., 2018) and supports projections for a wider range of emissions scenarios than those directly simulated in ISMIP6, but does so with less spatial detail than emulators that use regional climate forcings. Emulandice uses fixed parameters for glacier retreat~~and~~ , sub-shelf basal melt~~(Edwards et al., 2021). LARMIP~~ , and ice shelf collapse that result from the ISMIP6 experimental design (Edwards et al., 2021). Given the limited time available to deliver projections for AR6 and the continued evolution of the ISMIP6 datasets and documentation at the time, emulandice was designed around a smaller set of ISMIP-wide input variables that were more consistently defined across models and more tractable to validate in a new multi-model emulation framework. LARMIP, on the other hand, employs linear response functions to estimate how different rates of ice-shelf basal melting on the Antarctic Ice Sheet (AIS) translate into ice mass loss and sea level rise. This approach assumes a linear relationship between basal melt and ice mass loss and does not take into account feedback mechanisms or changes in surface mass balance (SMB), ~~which limits its ability to capture the complex, nonlinear dynamics of ice sheets (Levermann et al., 2020). While these~~ in order to isolate the effects of interactions between the ocean and ice (Levermann et al., 2020).

These tools have contributed significantly to our understanding of future sea level rise ~~, particularly when attempting to understand processes relevant to ice sheet evolution, the simplifications required for computational efficiency reduce projection accuracy (Van Katwyk et al., 2023)~~and are valuable for scenario exploration and large-scale assessment (Van Katwyk et al., 2023) . At the same time, the design choices that make such emulators practical for integration into broader climate workflows can limit how well they capture the full complexity of regional ice sheet responses when directly emulating the ISMIP6 ensemble. For example, ~~Emulandice was unable to distinguish between different emission scenarios for the AIS in the IPCC~~ within the ISMIP6-based AR6 ~~, effectively producing the same~~ application, emulandice showed limited separation between Antarctic emission scenario projections, effectively producing very similar sea level outcomes ~~regardless of scenario. This limitation is critical, as accurately modeling~~ across scenarios. Distinguishing differences between emission scenarios is ~~essential for providing reliable information to government bodies and policymakers (Fox-Kemper et al., 2021) and is an important part~~ a key component of climate uncertainty quantification (Hawkins and Sutton, 2009; Lehner et al., 2020) and of providing decision-relevant sea level information (Fox-Kemper et al., 2021).

This study proposes a novel ML-based ice sheet emulator for both the AIS and GrIS that accurately and efficiently ~~projects future sea level~~ emulates ISMIP6 sea level projections, quantifies projection uncertainties, and captures ~~sea level projection sensitivity to emission scenarios~~differences between emissions scenario ensembles. This model, which we call ISEFlow (**Flow**-based **I**ce **S**heet **E**mulator), is a neural network (NN)-based emulator that leverages the computational advantages of NNs and incorporates a flow-based architecture (Nalisnick et al., 2019) to separate sources of projection uncertainty. This work directly builds upon Van Katwyk et al. (2023), which demonstrated that NN architectures ~~are able to~~ effectively approximate ice sheet

dynamics ~~effectively due to~~ through incorporation of more input variables ~~enabled by its computational efficiency. This work extends those previous results by proposing~~ and explicit modeling of temporal data. ISEFlow (v1.0) ~~,~~ extends those results as a specific model with available pretrained weights ~~to be used as ice sheet emulators~~ for Earth's two extant ice sheets—ISEFlow-AIS and ~~ISEFlow-GrIS—as well as~~ ISEFlow-GrIS—with an improved framework for separating sources of uncertainty~~in ice sheet emulators~~.

~~Along with demonstrating advances in projection accuracy and uncertainty quantification in the proposed ISEFlow model, this work~~ This work also carefully examines the underlying processes that drive ISEFlow~~ΓÇÖ~~'s performance in order to ~~verify~~ assess whether the emulator is learning ~~correct physical principles and to build confidence in its projections~~relationships that are consistent with established ice sheet dynamics. We perform a series of experiments, including sensitivity testing and an analysis ~~on~~ of input variable feature importances, ~~to verify that the emulator is learning in a way that reflects the real-world behavior of ice sheet models~~with SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) serving as the primary method for evaluating whether ISEFlow's learned relationships are physically consistent with prior understanding from the ice sheet literature. By understanding which variables and forcings most influence emulated sea level projections, we not only ~~validate ISEFlowΓÇÖ~~evaluate ISEFlow's accuracy and adherence to ~~ice flow physics~~the ice sheet dynamics represented in ISMIP6, but also validate the emulator's ability to differentiate between emission scenarios (using a Kolmogorov-Smirnov D Statistic) and identify the inputs necessary for capturing ~~sensitivity to different~~ differences between emission scenarios. This ~~ensures that ISEFlow provides reliable insights for future sea level projections and~~ work establishes a foundation for creating more effective ice sheet emulators and for using ISEFlow to efficiently explore the behavior of the ISMIP6 ensemble. Adoption of ISEFlow will allow climate and ice sheet scientists to curate better datasets, enhance the capabilities of ISMs, and improve our understanding of future sea level.

2 Methods

2.1 Data

The data used to train the ISEFlow emulator comes from ISMIP6, which is an ensemble of ice sheet models that produced ~~land ice~~ ice sheet and corresponding sea level projections driven by climate forcings from the ~~Climate~~ Coupled Model Intercomparison Project - phase 6 (CMIP6). The CMIP forcings include yearly-averaged atmospheric and oceanic forcing anomalies from 2015 to 2100 at 8 km and 5 km resolution for the AIS and GrIS respectively (Nowicki et al., 2020; Slater et al., 2020; Jourdain et al., 2020). The key emulator prediction target is the Sea Level Equivalent (SLE) contribution produced by each ice sheet model within ISMIP6 for each experiment that was run, with geographic localization of the ice mass losses for regional sea level rise projections, the consequences of key process treatments or ice sheet model characteristics, and other fingerprinting activities as additional goals. ~~Both input forcings and projected~~ Projected SLE in ISMIP6 and this study are reported as anomalies, or deviations from a "control" experiment (simulation with constant climate conditions), to reduce the effects of model drift throughout the projections (Nowicki et al., 2020).

We aggregate input climate forcing data by calculating the mean forcing value for each ISMIP6 region, as in Seroussi et al. (2020) and Goelzer et al. (2020). Previous emulators partitioned the ice sheets into large regions~~to increase computational efficiency~~, aggregating data into 3 regions for the AIS (East Antarctica, West Antarctica, Antarctic Peninsula) and 1 region for the GrIS. This coarser aggregation was chosen primarily to ensure independence of projections across regions (avoiding the need to model spatial correlations of model errors between basins) and to reduce the complexity of what was, at the time, a novel multi-model emulation framework. We maintain the input data in a resolution closer to the original resolution, dividing the AIS into 18 sectors based on drainage basins and the GrIS into six sectors ~~, to preserve the maximum amount of spatial information possible~~(see Appendix A1 for sector maps). This finer partitioning allows for more detailed projections, particularly at the sector level, which provides a detailed understanding of regional ice sheet dynamics and their contributions to global sea level rise. The ability to produce projections for each sector is a significant advantage over previous emulators that limited projections to only broad regions, offering enhanced information for regional sea level projections. As a direct improvement from Van Katwyk et al. (2023), we also include ISM characteristics, or choices made by modelers that reflect how ice sheet projections are calculated by the ISM rather than the encoded name of the ISM being used to make a particular projection. These characteristics are encoded as binary variables and given to ISEFlow as inputs along with the climate forcings for each year. The main ISM characteristics include numerical solving method, initialization method, and basal melt parameterization schemes (for the full list and details of the characteristics, see Appendix A4, which is based on Table 3 of Seroussi et al. (2020) and Table A1 of Goelzer et al. (2020)). Including ISM characteristics ~~offers enhanced generalization beyond modeling specific ISM configurations present~~ also improves generalization across the range of model configurations represented in ISMIP6, as ISEFlow is not restricted to ~~only emulating the original set of ISMIP models used for training , so it could anticipate the results of model updates to characteristics settings in ISMIP7 configurations, for example~~reproducing only the exact combinations of ISM settings seen during training and can therefore approximate projections for new combinations of existing model characteristics and forcings within the sampled ISMIP6 design space. We also note that we do not use any additional SSP simulations, such as those presented in **?**, and focus only on RCP-based ISMIP6 simulations.

With the processed CMIP forcings and ISM characteristics as inputs, and the SLE anomalies produced by ISMIP as outputs, we group the 86-year time series into training, validation, and testing sets. The dataset for each ice sheet contains projections that represent a combination of ice sheet model, CMIP input model, and experimental configuration, resulting in a dataset of 3,180 projections for the AIS and 907 for the GrIS. For the AIS, the training set consists of ~~635 projections of~~ 2,226 projections (70% of all projections) of 86 years each, totaling ~~54,610 training observations from each ice sheet sector~~191,436 training observations across all ice sheet sectors, and the validation and testing ~~set contains 136 full projections~~ sets each contain 477 86-year projections (15% each). For the GrIS, the training set consists of 635 projections of 86 years, totaling 54,610 training observations, and the validation and testing ~~set contains~~ sets each contain 136 ~~full~~ 86-year projections. For more detailed information on all data acquisition, preprocessing, and preparation steps, see Van Katwyk et al. (2023). For the open-source Python package for handling ISMIP6 data and creating the proposed emulators, see Section 5: Code and data availability.

2.2 ISEFlow

ISEFlow is a deep learning-based emulator designed to improve sea level ~~projection~~ emulation accuracy and quantify uncertainty. Deep learning, a broad class of machine learning methods that use multi-layer NN architectures (LeCun et al., 2015), has been extensively applied in Earth and Climate Sciences due to its ability to efficiently and accurately approximate complex Earth systems (Karpatne et al., 2019; Bergen et al., 2019). ISEFlow combines two NN architectures: a Normalizing Flow (Rezende and Mohamed, 2015) for probability density estimation and quantification of data coverage uncertainty, and a Deep Ensemble (Lakshminarayanan et al., 2017) of ~~LSTM~~ Long Short-Term Memory (LSTM) models (Hochreiter and Schmidhuber, 1997) to generate accurate projections and quantify emulator uncertainty (Figure 1).

In this study, we categorize uncertainty into two primary types: data coverage uncertainty and emulator uncertainty. Data coverage uncertainty is typically considered to be an aleatoric sampling uncertainty that arises when the NN model makes predictions in areas where the training data exhibits variability or disagreement, meaning that similar input conditions can lead to different outcomes in the training data. This reflects how well the training data represents the range of possible system behaviors, with greater uncertainty in regions where ~~the~~ there is greater disagreement within ISMIP6~~data distribution is more variable or less consistently constrained~~. Emulator uncertainty, on the other hand, refers to epistemic uncertainty, which stems from the emulator's limited knowledge due to incomplete data, uncertainties in the model structure, or limitations in the training process (Hüllermeier and Waegeman, 2021). ~~Note that the~~ The limited accuracy of ice sheet dynamics models in the ISMIP6 ensemble and their CMIP6 forcing data also bring in other types of uncertainty (e.g., model bias, scenario uncertainty, intrinsic chaotic variability, etc.), which are not the focus here. Quantifying data coverage uncertainty and emulator uncertainty allows us to separate uncertainty that arises from the emulator itself from uncertainty due to variability in the training data.

The uncertainty quantification technique employed by ISEFlow is typically used to quantify and separate aleatoric and epistemic uncertainty (Hüllermeier and Waegeman, 2021). However, as emulators approximate another ~~modelΓÇöin~~ model—in this case, ice sheet ~~modelsΓÇöthe~~ models—the concept of "aleatoric uncertainty" does not apply in its traditional sense~~.~~ ~~Aleatoric uncertainty represents the inherent variability in observed data, but here,~~ , because the "data" ~~is~~ are generated by a model rather than from direct observational processes. ~~Consequently, the uncertainty being quantified pertains not to natural variability but to the variability present in the ISM-generated training data, including disagreement in sea level projections associated with~~ We therefore use the term "data coverage uncertainty" to reflect that the variability being captured arises from disagreement across ISMs under similar input conditions~~across different ISMs. Therefore, we use the term "data coverage uncertainty" to reflect the uncertainty arising from the distribution of behaviors represented in the training data rather than from observational noise. This allows for a clear separation between the uncertainty intrinsic to the emulator and the uncertainty arising from~~ , not from observational noise. For most downstream applications, the relevant quantity is the total predictive uncertainty, which reflects the combined effect of these two components. In other words, the total uncertainty represents both uncertainty associated with variability and disagreement in the ISMIP6 training ensemble and uncertainty introduced by the emulator approximation itself. However, decomposing total uncertainty into data coverage uncertainty and emulator uncertainty remains scientifically useful because it helps distinguish whether uncertainty in a given prediction is primarily originating from

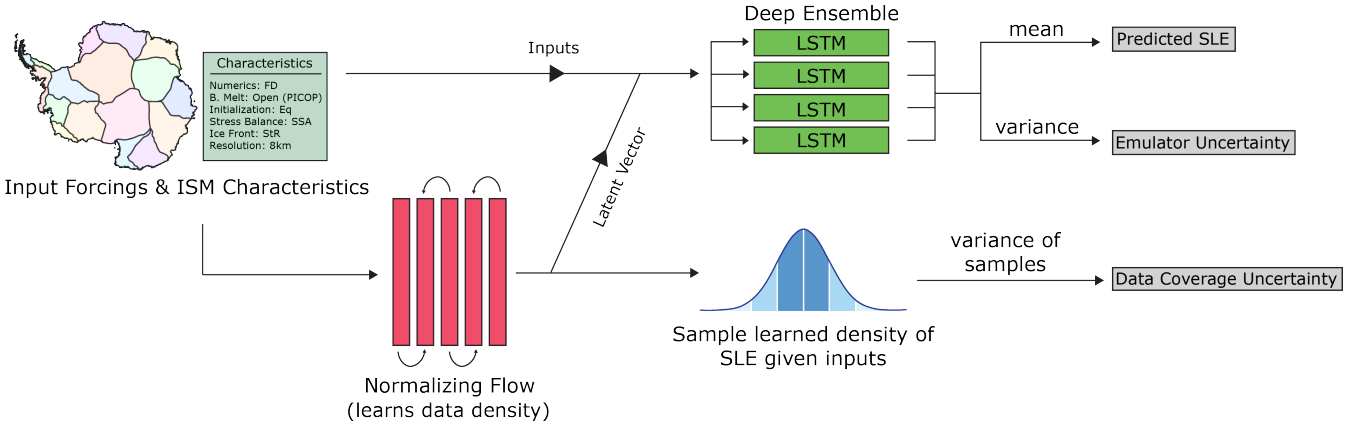

Figure 1. Architecture diagram for ISEFlow-AIS and ISEFlow-GrIS. The input data consists of climate forcings and ice sheet model characteristics. The input data ~~is~~ are first passed through a Normalizing Flow to approximate the data density. The resulting latent transformation is concatenated with the original input data and passed through the deep ensemble. The mean and variance of the resulting ensemble distribution represents the prediction and emulator uncertainty respectively. The learned data distribution given the inputs is then sampled and the variance of the learned distribution represents the data coverage uncertainty.

disagreement within the training data ~~.~~or from limitations in the emulator. This separation provides diagnostic insight that would be obscured if only a single aggregate uncertainty estimate were reported. While we predominantly use total predictive uncertainty in the present analysis, this decomposition could be useful in future work for identifying poorly constrained regions of the input space, guiding the design of future ISMIP experiments, and highlighting where additional simulations or more targeted model intercomparisons may be most informative.

For each projection, input forcings and ISM characteristics are first passed into a Masked Autoregressive Normalizing Flow (MAF) (Papamakarios et al., 2017), which consists of a series of learnable, reversible transformations that map the data distribution to a simple, latent distribution (e.g., a Gaussian distribution is used in ISEFlow). By learning these transformations, the data distribution can be efficiently sampled and the exact likelihood of the data can be computed, allowing us to estimate the conditional density of the data distribution given the input data (Rezende and Mohamed, 2015). The Normalizing Flow captures data coverage uncertainty by estimating the distribution of projected outcomes associated with a given set of input conditions, meaning that when similar inputs are associated with a wider range of outputs in the training data, the model expresses higher uncertainty. To calculate the data coverage uncertainty, we sample the learned distribution given the current input values, and take the variance of the generated samples.

The latent distribution generated from the input data by the Normalizing Flow, along with the input itself, is then passed through a Deep Ensemble of LSTM ~~modelsFÇöa~~ models—a type of NN particularly effective for time series data (Hochreiter and Schmidhuber, 1997; Van Katwyk et al., 2023). Using both the original data and latent space as inputs allows the LSTM models to leverage both the transformed, more structured latent representation and the original input features, ensuring that potentially useful information from the raw input is not lost during the transformation process. The mean of the ensemble pre-

dictions from the deep ensemble represents the overall projection, while the variance among the ensemble members quantifies emulator uncertainty. This hybrid uncertainty quantification approach ensures that both data limitations from the ISM models (via data coverage uncertainty) and emulator limitations (via emulator uncertainty) are captured. Separate emulators are trained for the Antarctic Ice Sheet (AIS) and the Greenland Ice Sheet (GrIS) to account for the distinct dynamics and processes of these systems and models.

We tested multiple configurations of ISEFlow architectures, including the number of transforms in the Normalizing Flow, the number of LSTM ensemble members, the training loss, and the architecture of each ensemble member within the Deep Ensemble. Given the large range of possible architectures and hyperparameters, we use Optuna (Akiba et al., 2019), an open-source framework for Bayesian parameter optimization, to efficiently explore and identify a range of optimal parameter configurations. We finalize the architecture using a grid search with the optimal parameter ranges, choosing the model with the lowest Mean Squared Error (MSE) on the validation dataset. The final ISEFlow architecture consists of a MAF with 5 flow layers, and 10 LSTMs of varying number of layers and layer sizes in the Deep Ensemble. Architecture variability within the ensemble members is necessary to accurately capture the emulator uncertainty (Lakshminarayanan et al., 2017). We use Python libraries PyTorch (Paszke et al., 2019) and nflows (Durkan et al., 2020) to implement ISEFlow. For more information on the ISEFlow architecture for both the AIS and GrIS, see Appendix B1. ISEFlow training was performed using a NVIDIA QuadroRTX GPU with 256 GB RAM. Training and ~~inference~~ prediction times were determined by logging wall times before and after model runs.

2.3 Comparison with previous emulators

We compare ISEFlow with the two ice sheet emulation tools used heavily in the IPCC AR6: LARMIP and ~~Emulandice.~~ a GP-based emulator following emulandice (Edwards et al., 2021). LARMIP is a linear response model that estimates ocean-driven dynamic ice mass change by convolving basal melt forcing time series with pre-computed linear response functions derived from ice sheet model simulations (Levermann et al., 2020). It is not an emulator in the general sense, as it does not represent the full range of ice sheet processes; it isolates the dynamic response to ocean forcing and does not account for surface mass balance (SMB) changes or nonlinear feedbacks. We run the LARMIP linear response functions using ISMIP6 AIS global surface air temperature (GSAT) as ~~inputs with the generated~~ the climate forcing input to generate the ensemble of 20,000 basal melt forcing time series as specified in Levermann et al. (2020). ~~LARMIP is a method used to understand ocean-driven basal melting only and is not specifically an emulator, so it requires additional data from ISMIP6. However,~~ Importantly, we do not add any SMB component to the LARMIP output, in contrast to the AR6 LARMIP assessment (Fox-Kemper et al., 2021) which combined the dynamic LARMIP projections with an SMB contribution. As a result, our LARMIP results represent only the dynamic component and likely show stronger scenario dependence than the AR6 assessment, because the opposing dynamic and SMB responses to warming are not combined here. The IPCC AR6 compares LARMIP alongside ~~the Emulandice emulator~~ emulandice (Figure 9.18, IPCC AR6 (Fox-Kemper et al., 2021)), so we follow that precedent here, referring to LARMIP as a "linear response model" throughout to distinguish it from data-driven emulators.

We implement a Gaussian ~~Process-based emulator that closely follows the Emulandice architecture presented in~~ Process (GP) emulator following a similar design to Edwards et al. (2021). We use a Gaussian Process with a power exponential ~~kernel with an optimal alpha (power) parameter of~~ kernel—which measures similarity between inputs as a smoothly decaying function of their distance, controlled by a power parameter (here set to 0.1) that governs how quickly correlation falls off—combined with an unbounded nugget kernel ~~to model random variability induced by variables not present in the dataset~~that adds a small independent noise term to each observation, accounting for random variability in sea level projections due to factors not explicitly included as inputs. We fit the data to a linear trend, and then train the zero-mean GP on the residuals of the linear function. ~~Following Edwards et al. (2021), Emulandice~~ The GP emulator is trained on ~~global~~ surface air temperature, ~~as well as a parameterization of glacier retreat for the GrIS and a~~ a sub-shelf basal melt ~~and an~~ parameterization (Jourdain et al., 2020) and ice-collapse boolean flag for the AIS~~. We also emphasize that the Emulandice model included in this study refers to the Emulandice architecture, rather than the exact Emulandice implementation presented in Edwards et al. (2021), as our implementation uses the ISMIP6-processed input forcings to train the emulator rather than the FaIR simple climate model (Smith et al., 2018).~~

, and a parameterization of retreat sensitivity (Slater et al., 2020) for the GrIS. Outputs are aggregated into three regions for the AIS (East Antarctica, West Antarctica, Antarctic Peninsula as defined by Shepherd et al. (2018)) and one region for the GrIS as was done in Edwards et al. (2021), rather than 18 regions for the AIS and six regions for the GrIS as is done in ISMIP6 and ISEFlow.

It is important to note that our GP emulator is an approximation to the emulandice design and differs from the original implementation in several key respects, which are documented fully in Appendix C. Most consequentially, we train on ISMIP6-processed forcings using individual GCM realizations rather than the ensemble-mean GSAT derived from CMIP5 and CMIP6 simulations used in emulandice. The original emulandice also incorporated additional simulations from **?** intended to broaden coverage of the input space, which we do not include. On the implementation side, emulandice relied on the RobustGaSP package (Gu et al., 2018) for marginal posterior mode estimation of GP hyperparameters, while we use the scikit-learn GP with maximum likelihood estimation. Emulandice was also trained on CMIP5 and CMIP6 GSAT and used FaIR (Smith et al., 2018) to generate future projections, whereas our implementation trains directly on the ISMIP6-processed CMIP forcings for the six available GCMs, without FaIR. Therefore, performance comparisons reflect our specific implementation rather than the original published emulandice, and should not be read as a direct evaluation of that model. We refer to this emulator as the "GP emulator" throughout the paper to make this distinction clear. All training and testing procedures were carried out on a 256 GB compute node with training and ~~inference~~ prediction times determined by logging wall times.

Accuracy of individual emulated projections is quantified using Mean Squared Error (MSE) and Mean Absolute Error (MAE). For each projection in the test set, the error at each time step is computed as the difference between the emulated SLE anomaly and the corresponding ISMIP6 SLE anomaly; these time-step errors are then averaged across all 86 years (2015–2100) to yield a per-projection error. MSE and MAE are the mean of the squared and absolute per-projection errors, respectively, averaged over all projections in the test set. In addition to emulation accuracy, we assess the emulator's ability to capture the distribution of projections through Kullback-Leibler (KL) (Kullback and Leibler, 1951) and Jensen-Shannon (JS) divergences

(Lin, 1991), which measure the similarity between the distribution of emulator projections at the year 2100 and the distribution of the true ISMIP6 projections. The KL divergence is asymmetric and sensitive to differences in the tails of distributions, while the JS divergence provides a symmetric and smoother measure of distribution similarity. Using both ensures a more comprehensive evaluation of how well the emulator captures the full range of the projection distribution at 2100.

2.4 Scientific Insights from Emulators

A key contribution of this work is to provide a framework for extracting scientific insights from emulators. We apply SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) to quantify the approximate contribution of various input features and ISM characteristics to ISEFlow's performance. SHAP is an approach that has been used extensively in the Earth Sciences (Yang et al., 2024) and other scientific domains (Gholami et al., 2024). SHAP calculates the contribution of each input feature to the ML model prediction by systematically considering predictions under all possible combinations of features and assigning an average marginal contribution for each feature. When applied to ISEFlow, it quantifies the impact of each climate forcing and ISM characteristic on the emulator~~FÇÖ~~'s output by estimating the difference in the predicted SLE anomaly when a feature is included or excluded. Here, we use SHAP to capture an overall approximation of feature importance across all predictions rather than for any individual projection, helping us interpret the contribution of different climate forcings and ISM characteristics within ISEFlow.

~~It is also important to note that while~~ While SHAP provides a detailed measure of each feature's contribution within the context of all inputs, it does not indicate how well the emulator would perform if certain inputs were used alone or in isolated combinations. To capture the predictive capabilities of surface air temperature and surface mass balance in isolation, which are commonly used forcings ~~used in isolation~~ for building climate emulators ~~(Edwards et al., 2021; Fettweis et al., 2013; Aschwanden et al., 2019)~~ (Fettweis et al., 2013; Aschwanden et al., 2019), we run the sensitivity tests reported in Section 3.2. We run SHAP on ISEFlow-AIS and ISEFlow-GrIS to determine which input forcings and ISM characteristics are most influential on emulator accuracy. We quantify the contribution of all variables, including aggregate forcings, or forcings that are calculated from other forcings (e.g., surface mass balance), and the underlying forcings (e.g., precipitation, runoff, evaporation) to provide a comprehensive view of feature importance. This captures the potentially complex, non-linear influences that individual forcings exert on model output which may be lost if only the aggregate forcings are used (Coulon et al., 2024). This dual approach allows us to distinguish between the predictive contributions of combined metrics and the complex interactions of individual variables, which ~~are essential for accurately assessing emulatorbehavior and performance~~help assess the trade-offs and design choices involved in selecting input variables for the emulator.

To better understand how ISEFlow-AIS and ISEFlow-GrIS are able to capture sensitivity to emission scenarios under Representative Concentration Pathway (RCP) 2.6 and RCP 8.5, we train separate models with the same architectural design as the ISEFlow models that classify the inputs into emission scenarios (i.e., the same inputs from ISEFlow are used but the target of prediction becomes the emission scenario rather than sea level). Then, by using SHAP on the classifier model, we are able to determine which input variables are most influential in distinguishing between scenarios (for information on the training procedure and performance, see Appendix B2). These analyses are useful for understanding which inputs drive the emulator's

Table 1. Performance metrics for ISEFlow (v1.0), ~~Emulandice (Edwards et al., 2021)~~the GP emulator (following a similar design to Edwards et al. (2021); see Appendix C for differences), and LARMIP (Levermann et al., 2020) for the Antarctic Ice Sheet (AIS) and Greenland Ice Sheet (GrIS). ~~Note that Emulandice refers~~ MSE is in mm^2 SLE, MAE is in mm SLE, and KLD, JSD, and KS D Statistic are dimensionless. Training time is tracked using minutes of wall-clock time. Prediction time is in seconds and represents the time it takes to produce projections for a new 86-year projections scenario. In the ~~Emulandice architecture trained with the same inputs as ISEFlow~~case of LARMIP, ~~rather than~~ the ~~original published model~~training time reflects the total time to generate all predictions, ~~which used a different~~ as emulator training ~~dataset and climate forcings~~is not required.

Ice Sheet	Emulator	MSE	MAE	KLD	JSD	~~Training Time (min)~~ KS D	Training Time	~~Inference Time (sec)~~ Prediction Time
AIS	ISEFlow-AIS	**1.20**	**0.53**	**0.05**	**0.009**	0.16	81.9	**0.58**
	~~Emulandice~~ GP emulator	3.03	1.10	10.83	0.289	0.12	23.6	3.17
	LARMIP	4.23	1.17	1.21	0.116	**0.31**	**20.3**	7.22
GrIS	ISEFlow-GrIS	**1.02**	**0.56**	**0.01**	**0.001**	0.15	63.1	**0.39**
	~~Emulandice~~ GP emulator	10.14	1.97	0.17	0.047	**0.26**	**11.4**	2.25

* KLD and JSD are evaluated at projection year 2100. KS D Stat is computed between the emulator-predicted RCP 2.6 and RCP 8.5 distributions at year 2100.

ability to accurately predict sea level, but also provides insights for capturing the uncertainty associated with future emissions. The results are intended to provide guidance for improving the selection of input features and refining the architecture for future emulators and future dynamical model simulations of ice sheet models. ~~We note that this~~ This work validates the emulator based on test data that was ~~help~~ held out of the original ISMIP6 ensemble~~:~~; we did not run additional dynamical model simulations to verify ISEFlow predictions.

3 Results

3.1 Emulation Performance

Emulator results are evaluated based on ~~individual~~ accuracy of individual emulated projections (where each projection consists of a single combination of ice sheet model, CMIP forcing, emission scenario, and model configuration, spanning 86 years), pro-

jection ~~accuracy, projection~~ distribution similarity, ability to distinguish between emission scenarios, and training and ~~inference~~ prediction time. As shown in Table 1 and consistent with previous results (Van Katwyk et al., 2023), ISEFlow demonstrates ~~superior~~ improved performance compared to ~~Emulandice~~ the GP emulator and LARMIP. For both the AIS and GrIS, ISEFlow achieves significantly lower Mean Squared Error (MSE) and Mean Absolute Error (MAE) values than other emulators, indicating more accurate ~~sea level projections~~emulation of ISMIP6. ISEFlow-AIS achieves an MSE of 1.20, compared to 3.03 for ~~Emulandice~~ the GP emulator and 4.23 for LARMIP. Similar improvements are observed for the GrIS, where the MSE for ISEFlow-GrIS is 1.02, significantly lower than ~~EmulandiceΓÇÖs 10.14.~~

~~In addition to projection accuracy, we assess the~~ the GP emulator's ~~ability to capture the distribution of projections through Kullback-Leibler (KL)(Kullback and Leibler, 1951) and Jensen-Shannon (JS) divergences (Lin, 1991), which measure the similarity between the emulatorΓÇÖs projection distribution at the year 2100 and the distribution of the true ISMIP6 projections. The KL divergence is asymmetric and sensitive to differences in the tails of distributions, while the JS divergence provides a symmetric and smoother measure of distribution similarity. Using both ensures a more comprehensive evaluation of how well the emulator captures the full range of the projection distribution.~~ 10.14 (see also Appendix D). As seen in Table 1, ISEFlow consistently achieves lower KL and JS divergence scores, demonstrating that ISEFlow is ~~not only~~ more accurate for ~~individual projections but also~~ emulating individual projections and better at capturing the full distribution of sea level rise ~~projections—a critical~~ projections—an important aspect of distinguishing sensitivity to emissions scenario.

~~Emulandice~~ The GP emulator trains faster than ISEFlow, ~~primarily because it emulates fewer regions , which results in less training data~~attributable to both the smaller number of regions and the smaller number of input features in the GP design. In contrast, ISEFlow maintains a finer spatial resolution by dividing the data into smaller regions to preserve more spatial information, which increases the time required for training. However, ~~inference~~ prediction time for ISEFlow is faster than for ~~Emulandice~~the GP emulator, making ISEFlow~~ΓÇÖ~~'s improved accuracy and spatial granularity more advantageous without sacrificing prediction efficiency. Reuse of v1.0 (this version) of ISEFlow-AIS and ISEFlow-GrIS with the pretrained weights provided here will enjoy these benefits without the additional training costs.

3.2 ISEFlow Sensitivity Testing

To verify that ISEFlow is learning the correct underlying physical relationships between input forcings, ice sheet model characteristics, and sea level, we conduct several additional experiments focusing on identifying variables beyond those that were included in ~~Emulandice~~ emulandice and LARMIP that enable the large accuracy improvements seen in Section 3.1 and the ability to distinguish between emission scenarios. First, we evaluate the impact of including ISM characteristics as inputs, comparing the performance of NN emulators trained with and without these characteristics (Table 2). The results show a substantial improvement in both the ~~projection~~ emulation accuracy and the ensemble approximation when ISM characteristics are included, which is expected given that the same forcing values can yield different SLE projections depending on the ISM used. Without ISM characteristics, the emulator tends to predict the average behavior of the ISMIP6 ensemble under similar forcings rather than capturing the unique responses of individual model configurations. For the AIS, the inclusion of ISM characteristics resulted in a 43.9% reduction in MSE, improving from 2.14 to 1.20, and a larger reduction was observed for the GrIS, where

Table 2. Results of adding ISM characteristics as inputs to ISEFlow. For both the AIS and GrIS, performance drastically increases in both ~~projection~~ emulation accuracy and ensemble approximation.

Ice Sheet	Emulator	MSE	MAE	KLD	JSD (2100)	Time to train
AIS	NN w/ Characteristics*	**1.20**	**0.53**	**0.05**	**0.009**	81.9 mins
	NN w/o Characteristics	2.14	0.74	0.266	0.026	**75.1 mins**
GrIS	NN w/ Characteristics*	**1.02**	**0.56**	**0.01**	**0.001**	63.1 mins
	NN w/o Characteristics	4.25	1.23	0.07	0.015	**58.4 mins**

*ISEFlow

the MSE decreased from 4.25 to 1.02. Both KL and JS divergences show substantial reductions when ISM characteristics are included, confirming that including characteristics not only improves ~~projection~~ emulation accuracy but also ~~better~~ improves the ability to emulate the distribution of possible sea levels. Using a larger number of input variables does, however, raise the question of potential overfitting. ISEFlow's performance is evaluated on a held-out test set that was withheld throughout training and hyperparameter optimization, and the PICP and Winkler Score metrics (Section 3.4) confirm that the emulator's uncertainty intervals are well-calibrated on this unseen data, providing evidence against overfitting.

We use SHAP to quantify individual ISM characteristics' contribution to emulator performance for both the AIS and GrIS, as shown in Figure 2, to determine which characteristics specifically are driving the increase in predictive accuracy. For the AIS, the most impactful ISM characteristic is which basal melt parameterization is used, followed by initialization method and model resolution, ~~consistently to~~ consistent with what was suggested in previous ISMIP6 analysis by Seroussi et al. (2023). Together, these three characteristics contribute significantly to the accuracy of the AIS emulator, suggesting that basal melt processes and initial conditions are important factors in determining future sea level projections. For the GrIS, the ice front retreat sensitivity to ocean forcing (based on Goelzer et al. (2020); Slater et al. (2020)) is the most important characteristic, followed by initialization method and the choice of a standard or open ocean forcing framework. Across both ice sheets, both the initialization method and parameters related to ocean forcing rank highly, emphasizing the role that both initial conditions and sensitivity to oceanic conditions play in projecting future sea level. Note that these rankings are *specific to ISEFlow's approximation of the ISMIP6 simulation suite*, in other applications (e.g., paleoclimatic change or if higher-resolution ocean or atmospheric models are used to formulate the forcing) the emulator would need to be re-trained and the rankings are likely to differ.

~~Next, we assess the role that additional forcings play in improving projection accuracy and the emulator's ability to distinguish between emission scenarios. Specifically, we compare the performance of emulators trained on all available forcings against emulators trained only on surface air temperature (Temperature) along with parameters for glacier retreat, sub-shelf basal melt, and ice shelf fracture, as described in Edwards et al. (2021). We also compare against emulators trained with SMB instead of Temperature to determine whether SMB (which is a function of precipitation, evaporation, sublimation, melt, and runoff, and is correlated with surface air temperature) offers more information for modeling SLE. Table 3 shows that across both ice sheets,~~

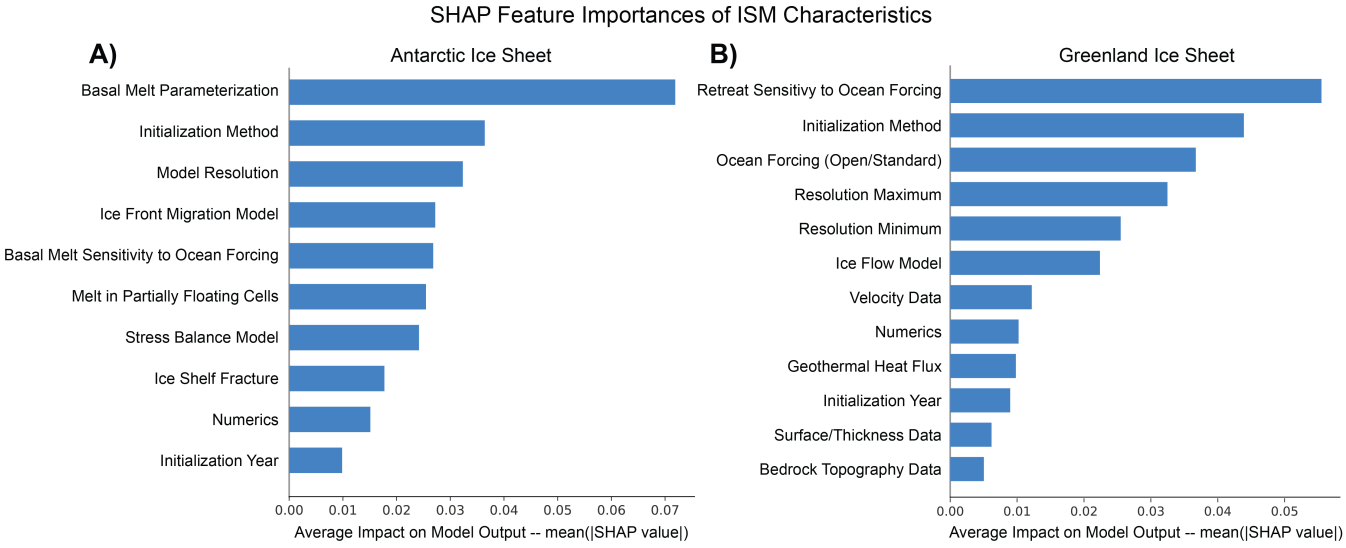

Figure 2. SHAP values representing the relative importance of different ISM characteristics for the Antarctic Ice Sheet (AIS, a) and Greenland Ice Sheet (GrIS, b) on ISEFlow projections.

~~emulators that used all available forcings consistently produce more accurate projections and demonstrate a greater capacity to distinguish between emission scenarios. It is conceivable that additional information (e.g., monthly or daily temperature forcing instead of annual mean) would continue this trend, but it is outside of the scope of the ISMIP6 design and thus inaccessible to the emulator, which is much more limited than a dynamical model in extrapolating to situations unlike its training data.~~

~~The Kolmogorov-Smirnov (KS) D statistic (Massey Jr, 1951), which measures the maximum distance between two distributions, is used to evaluate how well each emulator distinguishes between the sea level equivalent (SLE) distributions from different emission scenarios. Unlike MSE, which assesses overall projection accuracy, a higher D statistic indicates a stronger ability to differentiate between emission scenarios RCP2.6 and 8.5, reflecting the emulator's capacity to capture distinct distributions for each scenario.~~ Across both the AIS and GrIS, the NN emulator consistently outperforms the GP emulator in ~~projection~~ emulation accuracy. For the AIS, the NN trained on all forcings achieves the best accuracy ~~, with an MSE of~~ (MSE 2.14~~and a D statistic of 0.158. Even with only SMB, the NN maintains a similar ability to differentiate between scenarios (D statistic of 0.151) but with a slight decrease in accuracy compared to the all-variable emulator~~ (), followed closely by SMB (MSE 2.18) ~~. In contrast, training the NN on temperature alone results in weaker performance, with an MSE of~~ and then temperature alone (MSE 2.38~~and a D statistic of 0.132, showing that the inclusion of more input forcings is necessary to effectively capture scenario differences and create accurate projections~~). The GP emulators are generally less accurate than ISEFlow~~in predictive accuracy, but the GP trained on all variables demonstrates a similar ability to distinguish emission scenarios (D statistic of 0.153), although with lower projection accuracy (MSE~~ , with the all-forcing GP achieving an MSE of 2.60 ~~).~~

~~For the GrIS, the trend in predictive accuracy is similar to~~ for the AIS. The relatively small gap between the all-forcing and SMB-only configurations for the AIS could be because SMB already integrates the effects of precipitation, melt, and

Table 3. ~~Comparison of~~ Forcing sensitivity ablation comparing NN (ISEFlow without ISM characteristics) and GP emulators ~~trained on~~ across three input forcing configurations: all available forcings, local surface air temperature only~~(no ISM characteristics)~~, and SMB only, as well as categorical representations of glacier retreat, sub-shelf basal melt, and ice shelf fracture~~as specified in Edwards et al. (2021)~~. Results show that the NN architecture ~~consistently~~ outperforms the ~~Gaussian Process~~ GP in ~~both projection~~ emulation accuracy ~~and scenario sensitivity~~ for both the AIS and GrIS~~, and adding more forcings improves~~ . The GP shows strong ability to differentiate between scenarios for the GrIS. These results showcase the increased performance ~~further~~gained by adding additional input variables. LARMIP is not included in this sensitivity analysis because it uses a fundamentally different methodology (linear response functions) and accepts only a single input (GSAT), which does not lend itself to the same input variable sensitivity analysis performed here.

Ice Sheet	Architecture	Forcing Inputs	MSE	KS D Stat
AIS	NN	All*	**2.14**	**0.158**
		Temperature	2.38	0.132
		SMB	2.18	0.151
	GP	All	2.60	0.153
		Temperature	3.03	0.148
		SMB	2.88	0.126
GrIS	NN	All*	**4.25**	0.186
		Temperature	4.51	0.165
		SMB	4.41	0.171
	GP	All	7.02	0.201
		Temperature	10.14	**0.267**
		SMB	7.56	0.251

*No model characteristics included in inputs

runoff and is thus a compact summary of the climate state most directly relevant to surface mass changes, but the precise reason for this remains unclear and warrants further investigation. For the GrIS, the ~~AIS case, with the~~ NN trained on all forcings ~~achieving~~ again achieves the best results ~~, posting an MSE of~~ (MSE 4.25~~and a D statistic of 0.186. The GP emulator, while less accurate across all forcings, excels at distinguishing between emission scenarios for the GrIS. The GP trained on temperature forcings (equivalent to Emulandice) achieves a significantly higher D statistic of 0.267, demonstrating the greatest ability to separate the emission scenarios, despite having lower accuracy ()~~, while the GP emulator is substantially less accurate when using temperature alone (MSE 10.14). ~~These results reflect the findings in the IPCC AR6, where Emulandice effectively distinguished between emission scenarios for the GrIS but not for the AIS (Fox-Kemper et al., 2021). These findings show that NN-based emulators provide enhanced predictive accuracy and are capable of distinguishing between emission scenarios for both ice sheets. Additionally, they highlight that training ice sheet emulators using temperature alone, while helpful for~~

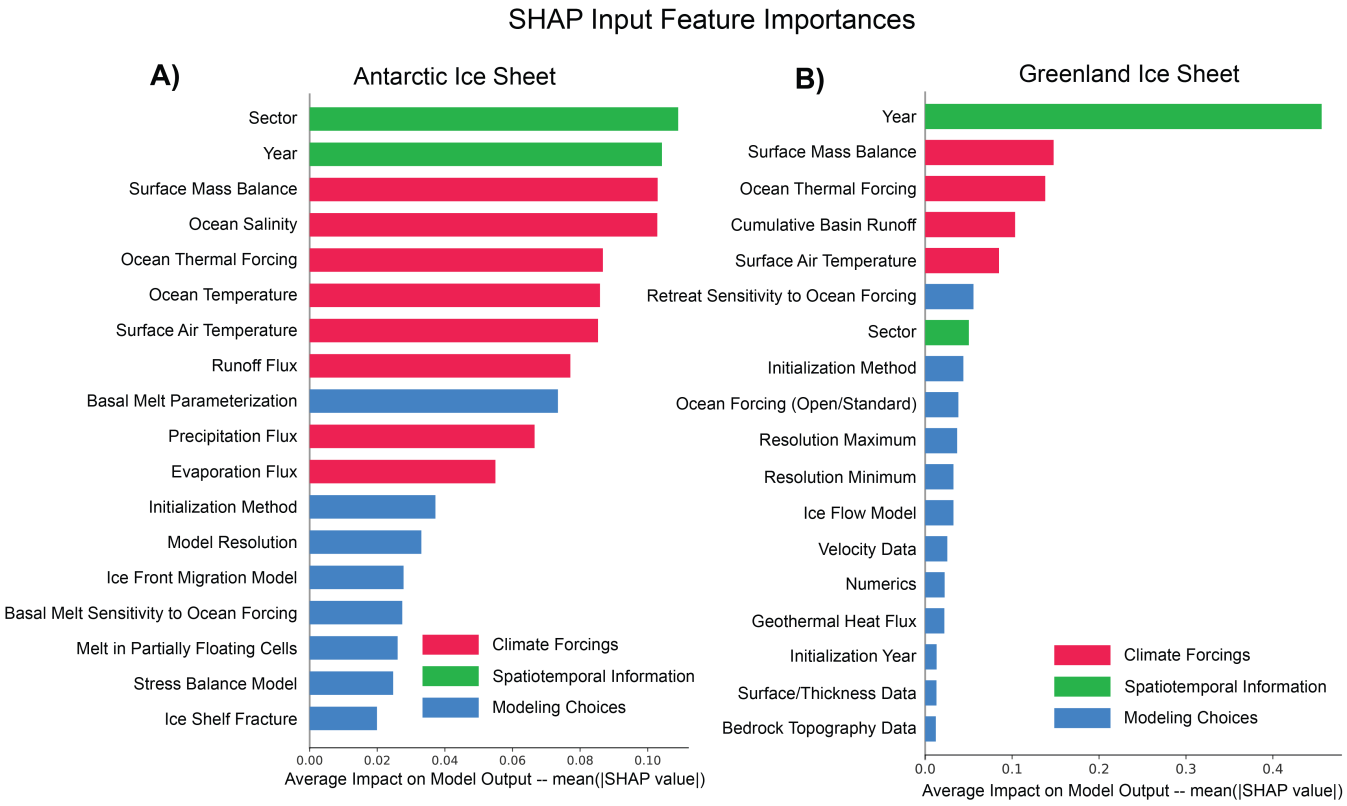

Figure 3. SHAP values for ISEFlow input variables for both the AIS (a) and GrIS (b). Bar color indicates the feature group, consisting of spatiotemporal features, climate forcings, and ISM characteristics. Results are computed using the full ISEFlow models with all input variables (including all characteristics shown in Appendix A4), but only the top 18 features by importance are shown for clarity.

~~distinguishing scenarios in some cases or when additional forcings are unavailable, is insufficient for producing accurate future sea level projections~~Scenario sensitivity results from these forcing configurations are discussed in Section 3.3.

Figure 3 shows the SHAP values for each of the input features in ISEFlow, separated into the following groups: spatiotemporal information, climate forcings, and ISM characteristics. For the AIS, spatiotemporal information (sector of the AIS and projection year) was the most impactful, followed by the climate forcings surface mass balance, ocean salinity, and ocean thermal forcing. Similarly, for the GrIS, year and surface mass balance were the top contributors. Across both ice sheets, spatiotemporal information was the feature group with the largest average contribution to model performance. Of the three feature groups, ISM characteristics were the least impactful on emulator projections, but still contribute enough information to increase emulator accuracy significantly (Table 2). For Greenland, this analysis is broadly consistent with Rohmer et al. (2022) in identifying retreat sensitivity (commonly denoted κ) and initialization as important controls on projected sea-level contribution, with other characteristics such as numerical scheme having little impact on projection values.

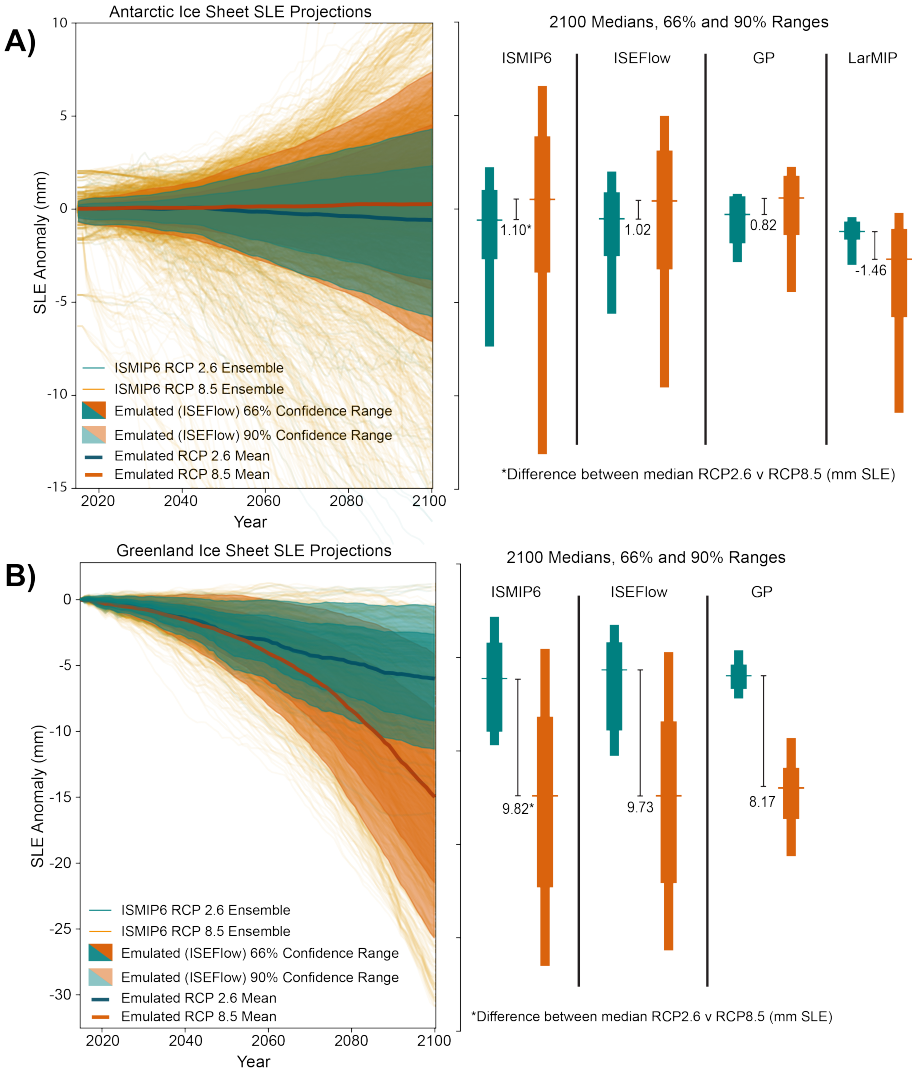

Figure 4. Plot based on IPCC Figures 9.17 and 9.18 (Fox-Kemper et al., 2021) showing the projections of the AIS (a) and GrIS (b) along with ISMIP6, ISEFlow, ~~Emulandice~~the GP emulator, and LARMIP median predictions at the year 2100 grouped by emissions scenario. The 5th/95th percentiles and medians shown represent the emulated projection distributions only, without additional sampling of emulator uncertainty, consistent with Figure 9.18 of IPCC AR6. ISEFlow accurately approximates the spread of each scenario distribution and identifies the difference in median SLE projections between emission scenarios. ~~Emulandice is unable to distinguish~~ The GP emulator shows smaller differences between ~~emission scenarios~~ RCP scenario medians for the AIS ~~and fails~~ compared to ~~approximate the ensemble spread~~ISEFlow. LARMIP, due to the use of linear response functions, skews to negative SLE anomalies and does not account for any plausibility of projections with positive SLE anomalies, which leads to an inaccurate approximation of both the median emission scenario differences and the spread of emission scenario projections. We note that some ISMIP6 projections do not begin at zero anomaly in 2015, which reflects differences in model initialization and reference states within the ISMIP6 ensemble, rather than post-processing in this study. Following ISMIP6 convention, SLE anomalies are reported as deviations from a control experiment: negative values indicate ice mass loss relative to the control, which corresponds to a positive contribution to sea level rise; this convention may differ from IPCC figures which report SLE as a positive sea level contribution.

3.3 Scenario Sensitivity

~~A significant advancement of the ISEFlow emulator is its ability to effectively distinguish between different emission scenarios for both the AIS and GrIS~~We evaluate scenario sensitivity using the Kolmogorov-Smirnov (KS) D statistic (Massey Jr, 1951), which measures the maximum distance between two cumulative distributions; higher values indicate greater separation between the emulator-predicted RCP 2.6 and RCP 8.5 distributions, with 0 indicating identical distributions. Figure 4 shows that ~~, unlike Emulandice, which had difficulty differentiating between scenarios for the AIS in the IPCC AR6,~~ ISEFlow captures the variations in sea level projections across emission scenarios more accurately ~~, providing a clearer understanding of potential future outcomes~~than the GP emulator. The KS test results (Table ~~??~~1) further confirm this: for the AIS, the full ISEFlow configuration achieves a D statistic of 0.161 (95% CI: 0.154–0.175) compared to 0.124 (95% CI: 0.119–0.142) for the GP emulator, a statistically distinguishable difference under bootstrap resampling ($\Delta D = 0.033$, ~~demonstrating that ISEFlow has a strong ability to distinguish between RCP 2.6 and RCP 8.5 scenarios, particularly for the AIS, where the NN emulator outperforms the GP and Emulandice in both scenario separation and projection accuracy~~95% CI: 0.020–0.049) (Tibshirani and Efron, 1993) . However, for the GrIS, the GP emulator ~~trained on temperature forcings excels in distinguishing between emission scenarios, achieving the highest KS D statistic, although at the cost of lower predictive accuracy. This is likely caused by Emulandice's inability to emulate the ensemble spread (Table 1 KLD & JSD) , which results in a substantial divide between the distributions of RCP 2.6 projections and RCP 8.5 (Figure 4). Note that a reason for Emulandice's strong distinction between scenarios is~~ achieves a higher D statistic of 0.264 (95% CI: 0.248–0.281) compared to 0.153 (95% CI: 0.144–0.165) for ISEFlow ($\Delta D = -0.111$, 95% CI: -0.131–-0.091).

These two tables reflect complementary aspects of emulator design. Table 3 holds the input set constant across architectures—both the NN and GP are trained on the same forcing configurations (All, Temperature, or SMB only), with no ISM characteristics—so that differences in KS D reflect architectural behavior rather than input choice. Under these conditions, the GP achieves comparable or higher KS D values for the GrIS (D = 0.201–0.267 vs. 0.165–0.186 for the NN), while the NN leads for the AIS. Table 1, by contrast, compares the full ISEFlow design (all forcings plus ISM characteristics) against the GP emulator as deployed following the emulandice design (GSAT only); here ISEFlow achieves higher KS D for the AIS. Together, these results indicate that the NN architecture's advantage in scenario sensitivity for the AIS grows with access to a richer input set, and that the GP's higher GrIS D statistics in Table 3 partly reflect its underestimation of the ~~spread of the~~ ISMIP6 ensemble ~~which separates the distributions of GrIS outcomes excessively.~~ spread (Table 1, KLD & JSD), which inflates the apparent separation between RCP distributions.

These findings are broadly consistent with the IPCC AR6 results, where ~~Emulandice was able to distinguish~~ emulandice distinguished GrIS scenarios more effectively than AIS scenarios ~~.~~(Fox-Kemper et al., 2021). The KS D test compares the two RCP sub-ensembles as they exist in ISMIP6, which have different sizes and sample different portions of the input uncertainty space; the test therefore does not isolate genuine scenario dependence from differences due to the uneven ISMIP6 experimental design. Studies with more balanced ensemble designs have found limited scenario differences in ISMIP6 projections before 2100 (e.g., Lowry et al., 2021; Coulon et al., 2025). Importantly, the scenario separation observed in the emulators is bounded

by what exists in the underlying ISMIP6 ensemble itself. Applying the KS test directly to the ISMIP6 simulations yields D = 0.162 (95% CI: 0.151–0.176) for the AIS and D = 0.172 (95% CI: 0.164–0.181) for the GrIS—statistically significant but modest in magnitude, indicating that even the true ensemble does not cleanly separate RCP 2.6 from RCP 8.5 at year 2100. ISEFlow's emulated D statistics (0.16 for AIS, 0.15 for GrIS; Table 1) are therefore close to this empirical ceiling, suggesting the emulator captures nearly all of the scenario signal present in the training data.

LARMIP, as a linear response model~~with a constant SMB~~, distinguishes between emission scenarios by applying response functions that link basal melt rates to ice mass loss. However, these linear assumptions introduce limitations by ~~neglecting~~ omitting key nonlinear self-dampening and amplifying processes (Levermann et al., 2020), which can lead to substantial inaccuracies. While LARMIP ~~achieves some differentiation across scenarios ,~~ can differentiate across scenarios well, particularly because of the absence of added SMB, its inability to capture nonlinear ice sheet responses results in notable projection errors~~, reducing its reliability as an emulator. However, LARMIPΓÇÖ~~. LARMIP's computational efficiency and ability to isolate ~~ocean dynamics~~ the ocean-driven dynamic component, however, make it valuable for quickly assessing a wide range of potential outcomes, particularly in preliminary assessments where the primary goal is to estimate the uncertainty range associated with ocean-driven ~~basal melting~~dynamic change.

~~Comparison of the ability of ISEFlow, Emulandice, and LARMIP to distinguish between emission scenarios for the Antarctic Ice Sheet (AIS) and Greenland Ice Sheet (GrIS) using the KS D statistic. A Kolmogorov-Smirnov (KS) test was conducted between predicted distributions of RCP 2.6 and RCP 8.5 emission scenario projections. The MSE (Ensemble) and MSE (mean) represent the error of approximating the entire ISMIP ensemble, and the mean of the ISMIP ensemble respectively. Ice Sheet Emulator MSE (Ensemble) MSE (mean) KS D Stat ISEFlow **1.20** 0.001 0.16 Emulandice 3.03 0.026 0.12 LARMIP-2 4.23 0.683 **0.31** ISEFlow **1.23** 0.007 0.15 Emulandice 10.14 0.060 **0.26**~~

~~To identify the drivers of ISEFlow's ability to distinguish between emission scenarios, we run SHAP on the classification model mentioned~~ To identify which climate forcings carry the most scenario-distinguishing information, we apply SHAP to the emission scenario classification model described in Section 2.4 (Figure 5). This classification model is a diagnostic tool—a separate model trained specifically to predict which emission scenario a given set of climate forcings belongs to—and is distinct from ISEFlow itself. Its purpose is not to directly test ISEFlow's scenario sensitivity, but rather to identify which variables in the climate forcing data most strongly encode scenario information, thereby informing emulator design choices. Surface air temperature is the primary driver in distinguishing scenarios for both the AIS and GrIS. However, the classifier also highlighted that the inclusion of additional variables significantly enhances the model's ability to differentiate between scenarios. The values shown in Figure 5 represent the average absolute SHAP values for each feature, which reflect the average impact of each feature on the model's output in distinguishing emission scenarios. These SHAP values do not sum to 1 or any specific target, as they are not proportions or probabilities but rather independent measures of feature importance and should be interpreted as relative magnitudes compared to the other features. For the AIS, the average SHAP value of all other features combined is 0.434, while the average contribution from surface air temperature is 0.214. Likewise, for the GrIS, the combined impact of other features (0.153) was nearly equal to that of surface air temperature (0.182). This finding ~~illustrates that integrating a broader~~ shows that while temperature is the single strongest scenario indicator, the combined

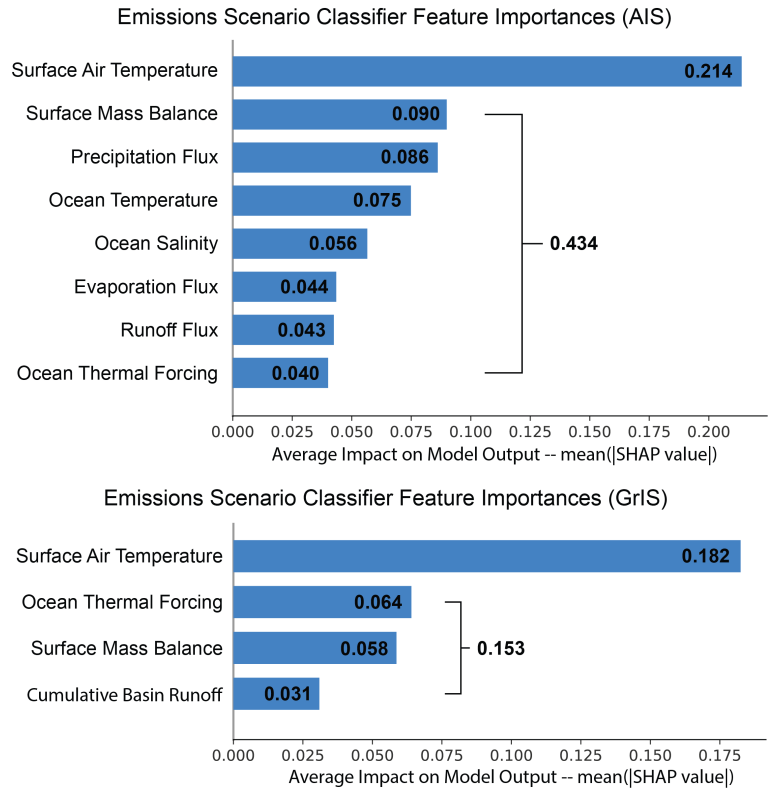

Figure 5. SHAP values showing the importance of various features in distinguishing between emission scenarios for the AIS (top) and GrIS (bottom). Surface air temperature is the most influential forcing for both ice sheets. However, for the AIS, the combined importance of the remaining forcings is more than twice that of temperature alone, while for the GrIS, the remaining forcings contribute an approximately equal amount to model prediction as temperature ~~along~~alone.

contribution of all other forcings is comparable to or exceeds that of temperature alone, supporting the design choice to train ISEFlow on a broad set of climate forcings~~is essential to improve the model's sensitivity to different emission scenarios~~. These results are not contradicted by the sensitivity analysis in Table 3, where SMB-only configurations often outperform temperature-only configurations in emulation accuracy (MSE). The classification model asks which variables best *distinguish emission scenarios*—a question where temperature excels because it is the most direct expression of the scenario forcing. The emulation model, by contrast, asks which variables best *predict ice sheet response*—a question where SMB performs strongly because it integrates precipitation, melt, and runoff into a compact summary of the climate state most directly relevant to surface mass changes.

3.4 Uncertainty Quantification Comparison of Emulators

To directly compare the quantified uncertainty between ISEFlow and ~~Emulandice~~the GP emulator, we trained a separate model (Regional ISEFlow) with the same architecture as ISEFlow using the same dataset and input variables as ~~Emulandice~~the GP emulator, including aggregated regional data (as opposed to higher resolution sector data) and limiting forcings to local surface air temperature, glacier retreat, sub-shelf basal melt, and ice shelf fracture (Table 4). This approach ensures a fair architectural comparison between the models in terms of their ability to quantify uncertainty. To evaluate the uncertainty estimates of each emulator, we use several metrics that assess both accuracy and sharpness, ensuring that the uncertainty predictions are reliable and precise. We primarily compare the ~~regional~~ Regional ISEFlow emulator with the ~~Emulandice to compare~~ GP emulator to assess the emulator architecture's ability to produce accurate uncertainty, but add metrics for the proposed ISEFlow models to show the usefulness of using more forcings that are higher resolution.

Prediction Interval Coverage Probability (PICP) (Khosravi et al., 2011) measures how often the true values fall within the predicted intervals, offering insight into the model's reliability. A higher PICP indicates better coverage, but it must be considered alongside Mean Prediction Interval Width (MPIW) (Pearce et al., 2018), which quantifies the sharpness of the intervals. While a ~~higher PICP~~ PICP close to the target (95%) is desirable, excessively wide intervals (higher MPIW) may reflect underconfidence, suggesting the model is being too conservative with its uncertainty estimates. Similarly, excessively low MPIW may reflect overconfidence. MPIW helps balance coverage (PICP) with precision by ensuring intervals are precise. A Continuous Ranked Probability Score (CRPS) (Gneiting and Raftery, 2007) is used to assess the accuracy of probabilistic predictions. It measures the difference between predicted cumulative distributions and actual outcomes, with lower values indicating better probabilistic performance. Finally, the Winkler Score (Winkler, 1972) combines coverage and sharpness into a single metric, penalizing both overly wide intervals and intervals that do not encompass the true values, providing a comprehensive view of the quality of quantified uncertainty. Region-level projections were combined together and metrics were calculated on the combined dataset.

When comparing ~~the~~ Regional ISEFlow and ~~Emulandice~~ the GP emulator (Table 4), Regional ISEFlow ~~performs better on PICP , indicating more reliable coverage of the true outcomes. However, Emulandice~~ achieves higher PICP for the GrIS but lower PICP for the AIS relative to the 95% target. The GP emulator achieves narrower intervals (lower MPIW) ~~,~~ but at the expense of ~~significantly worse PICP. This suggests that Emulandice tends~~ substantially worse coverage for the GrIS (PICP of 54.1%), indicating a tendency to underestimate uncertainty ~~, leading to intervals that fail to capture the true values consistently~~for that ice sheet. The coverage (PICP) is much lower than the emulandice implementation presented in Edwards et al. (2021) , which could be due to differing training data or specific implementation differences, as outlined in Appendix C. The CRPS values are ~~very close between the two models, but Regional ISEFlowstill slightly outperforms Emulandice~~close between Regional ISEFlow and the GP emulator. We also note that the larger uncertainty intervals produced by our GP emulator and Regional ISEFlow, while penalized by accuracy-sharpness metrics relative to the target 95% PICP, may be viewed as an advantage in the context of sea level projections, where the ISMIP6 ensemble is known to underestimate real-world uncertainty.

Table 4. Comparison of emulator uncertainty quantification performance. A separate ISEFlow emulator was trained on the same data and inputs as ~~Emulandice~~ our GP emulator (regional projections instead of sector projections, surface temperature forcings, etc.) to provide a fair architectural comparison~~to the Emulandice emulator~~. Metrics include the Continuous Ranked Probability Score (CRPS), Prediction Interval Coverage Probability (PICP; target: 95%), Mean Prediction Interval Width (MPIW), and the Winkler Score. Between the ~~regional~~ Regional ISEFlow emulator and ~~Emulandice~~the GP emulator, Regional ISEFlow ~~is more accurate (~~achieves higher PICP ~~),~~ for the GrIS but ~~has wider~~ lower PICP for the AIS, while the GP emulator achieves narrower intervals (~~higher~~ lower MPIW). The Winkler ~~score~~ Score indicates that the ~~regional~~ Regional ISEFlow ~~models~~ emulators do better at balancing ~~the~~ interval width and ~~accuracy of intervals~~coverage. The full ISEFlow emulators, ISEFlow-AIS and ISEFlow-GrIS~~are consistently most accurate (lowest PICP) and approximate~~ , show the ~~probability distributions best (highest CRPS)~~largest improvements across nearly all metrics, ~~showcasing~~ demonstrating the benefit of including additional higher-resolution forcings and ISM characteristics.

Ice Sheet	Emulator	PICP	MPIW	CRPS	Winkler
AIS	Regional ISEFlow	0.826	4.168	0.463	10.777
	~~Emulandice~~ GP emulator	0.918	**2.274**	0.371	21.561
	ISEFlow-AIS*	**0.986**	4.535	**0.235**	**4.768**
GrIS	Regional ISEFlow	0.858	6.647	0.725	16.273
	~~Emulandice~~ GP emulator	0.541	**2.820**	0.668	28.817
	ISEFlow-GrIS*	**0.969**	3.786	**0.263**	**4.339**

* Trained on all available forcings and ISM characteristics.

~~When comparing the~~ The full ISEFlow-AIS and ISEFlow-GrIS emulators, compared to both Regional ISEFlow and ~~Emulandice, ISEFlow-AIS and -GrIS~~ the GP emulator, show significant improvements across nearly all metrics. ~~It achieves~~ They achieve the lowest CRPS and Winkler scores, indicating more accurate and ~~sharper~~ well-calibrated uncertainty quantification. ~~It also attains higher coverage (PICP), balancing reliability and interval sharpness.~~ While the MPIW for the full ISEFlow emulators is ~~slightly higher than Emulandice~~higher than the GP emulator, the trade-off is well justified by the substantial gains in coverage and overall accuracy ~~,~~ as reflected by the Winkler Score. These results demonstrate the ~~superior uncertainty quantification abilities when~~ benefit of using a wider range of input forcings at ~~a~~ higher resolution and including ISM characteristics~~in making predictions~~.

4 Discussion

This study proposes ISEFlow as an accurate and efficient ice sheet emulator ~~, based on prediction accuracy and quality of quantified uncertainty. ISEFlow extends the~~ of the ISMIP6 ensemble, extending the work of Van Katwyk et al. (2023)~~and~~

~~significantly improves sea level projection accuracy and uncertainty quantification compared to the widely used emulators (Edwards et al., 2021). For both the Antarctic Ice Sheet (AIS) and Greenland Ice Sheet (GrIS), ISEFlow achieved lower Mean Squared Error (MSE), indicating greater predictive accuracy . Our results show that this increase in accuracy stems from both .~~
The accuracy improvements over our GP emulator and LARMIP baselines stem from two key design choices: the inclusion of ISM characteristics as input features, ~~as well as~~ and the incorporation of all available ISMIP6 climate forcings at finer spatial resolution.

A primary goal of this study is to ~~determine the ability of ISEFlow to capture the correct underlying processes, and whether the emulator~~assess whether ISEFlow's learned approximation of ice sheet dynamics is rooted in physical principles. ~~The SHAP analysis offers a window into ISEFlow's internal workings, showing which features have the greatest influence on sea level projections from 2015 to 2100 following the ISMIP6 protocol. This not only allows us to see how ISEFlow is learning, but also enables us to see which variables or which ISM characteristics have the greatest effect on sea level projections.~~

From the feature importance analysis (SHAP), it is evident that ISEFlow is learning relationships that align with domain knowledge of ice sheet dynamics. For both the AIS and GrIS, surface mass balance and ocean thermal forcing are among the top contributors, which is consistent with the understanding that these are primary drivers of ice sheet changes. We also see the importance of including forcings that are used to calculate SMB and thermal forcing, as doing so enables the emulator to learn how components of SMB and ocean thermal forcing change over time (Coulon et al., 2024). Interestingly, we also find that spatial information plays a more significant role for the AIS, reflecting the greater spatial variability of ice sheet processes there. In contrast, the GrIS exhibits more spatially homogeneous ice sheet behavior, which is captured by the comparatively lower importance of spatial information.

For AIS, basal melt parameterizations, particularly the sensitivity of basal melt to ocean forcing, emerge as important model characteristics, supporting the notion that processes at the ice-ocean interface dominate AIS dynamics. These findings are consistent with results presented in Seroussi et al. (2023), which focused on isolating uncertainty sources with constant SMB. Similarly, for the GrIS, SMB-related inputs like runoff and precipitation are dominant, in line with the importance of surface-driven changes in this region. As discussed in Section 3.1, the GrIS SHAP rankings are broadly consistent with those reported by Rohmer et al. (2022), who also identified retreat sensitivity (κ) and initialization as key controls on projected GrIS sea-level contribution using SHAP analysis.

ISEFlow has the potential to significantly enhance the ISMIP6 projection suite by supplementing experiments that could not be executed within the limited timeline available for the IPCC assessment cycles. For instance, ISMIP6 primarily provided projections for only two Representative Concentration Pathways (RCP2.6 and RCP8.5), ~~limiting the resolution of future socioeconomic~~ relying on emulandice to fill in intermediate scenarios. ISEFlow could be employed as a complementary tool to emulate sea level projections for intermediate Shared Socioeconomic Pathways (~~SSPs~~as was done with emulandice), which were not included in ISMIP6. This would provide a more detailed understanding of the ice sheet response across the full spectrum of potential emission scenarios. Additionally, ISEFlow could serve to explore other experiments that were lower priority (e.g., Tier 2 & 3 in Nowicki et al. (2020)), which could provide further insights into ice sheet dynamics. ~~By filling these gaps,~~

~~ISEFlow not only augments the comprehensiveness of sea level projections but also ensures that policymakers have access to a broader, more robust set of scenarios to inform climate action.~~

This analysis also highlights areas where improvements in ice sheet modeling could enhance emulator projection accuracy. For the AIS, the results underscore the critical role of basal melt processes and the need for high-fidelity parameterizations of ice-ocean interactions. The importance of model initialization methods suggests that more consistent or detailed initial conditions could reduce model variability and improve emulator training. This has implications for ISMIP7, where careful consideration of initialization practices could enhance the reliability of emulators like ISEFlow. More broadly, ISEFlow's ability to learn how model characteristics and forcing choices affect projected sea level makes it a useful tool for ISMIP7 experimental design. It can be used to predict how different model configurations affect projections, helping to prioritize which simulations to run in order to most effectively reduce projection uncertainty. We also highlight that as we have access to more projections with each model characteristic or a more even sampling of modeling choices across the experimental protocol, the ISEFlow emulator will better approximate the effect that each one has on sea level projections. Likewise, as modeling groups conduct additional simulations exploring a wider range of model configurations beyond prescribed projections, emulators like ISEFlow will benefit from a broader data space, further improving their ability to capture the diversity of ice sheet responses and enhancing the robustness of future projections.

These ~~enhanced~~ capabilities enable ISEFlow to serve ~~multiple~~ distinct purposes in the CMIP7/AR7 cycle ~~. As FaIR-Emulandice was used in AR6 to impute SSPs not simulated in ISMIP6, ISEFlow offers an alternative which can be used for a multi-emulator assessment. ISEFlow also has other capabilities of use. It has learned to predict behavior based on dynamical model characteristics, which makes it valuable as a tool in designing ISMIP7 configurations. It also should~~ beyond those already discussed. ISEFlow could prove useful in evaluating CMIP7 forcing products for inclusion into the ISMIP7 ensemble~~(during CMIP6 this process was based on expert opinion and CMIP6 input variable-to-observation comparison, not based on any ice sheet model output)~~. As ice sheet models are run after the CMIP ensemble is mostly complete, ~~it is always a rush~~ there is always time pressure to deliver meaningful ice sheet model data in time for the IPCC assessment report. ISEFlow makes it possible to emulate what the ISMIP6 configurations of ~~all of~~ the dynamical ice sheet models would have predicted when given any CMIP7 climate model forcing data, ~~and thus it can speed~~ thereby speeding the ice sheet model ensemble simulation process. A promising direction for future work is to develop a temperature-only version of ISEFlow that could be coupled with simple climate models such as FaIR (Smith et al., 2018) or FACTS (Kopp et al., 2023), which would allow projections for any emissions scenario without requiring full regional GCM forcing data. While the sensitivity tests in Section 3.2 demonstrate that using local temperature alone yields lower accuracy than using all available forcings, the temperature-only configuration remains informative and this coupling represents a practical path toward broader applicability.

~~The findings of this study underscore the importance~~ These results highlight the value of incorporating a ~~diverse set of input forcings and model characteristics~~ broader set of climate forcings into ice sheet emulators ~~, challenging the assumption that temperature alone is sufficient for accurately capturing ice sheet responses. The improved performance of ISEFlow suggests that future emulators should integrate a wider range of physical variables to enhance predictive accuracy and scenario sensitivity. This insight extends beyond ice sheets to other components of the Earth system where emulators are increasingly~~

used, such as glaciers (Jouvet and Cordonnier, 2023), weather (Li et al., 2024), and ocean circulation models (Guo et al., 2024). In these domains, relying on a single dominant forcing variable may overlook critical interactions and feedback mechanisms, leading to reduced reliability in long-term projections. Furthermore, the demonstrated success of combining normalizing flows with deep ensembles for uncertainty quantification provides a scalable framework for other climate emulation tasks, enabling better separation of model when that information is available. The classification SHAP analysis (Section 3.3) supports this: while surface air temperature is the single strongest scenario-distinguishing variable, the combined contribution of all other forcings is comparable in magnitude, indicating that regional forcings encode scenario-relevant information not captured by temperature alone. Different emulator designs, however, involve genuine trade-offs. Emulandice (Edwards et al., 2021) uses GSAT by design for several deliberate reasons: it naturally captures the correlation between the two ice sheets induced by global climate change, it integrates directly into the FACTS sea level framework (Kopp et al., 2023), it provides access to large probabilistic climate ensembles based on AR6-assessed climate sensitivity distributions, and data-driven uncertainties. As machine learning-based emulators continue to evolve, these results highlight the necessity of balancing computational efficiency with physical interpretability to ensure that emulators remain robust and scientifically credible tools for scientific modelingit allows projections to be rapidly updated for any emissions scenario without waiting for new GCM simulations. ISEFlow requires regional CMIP forcing inputs, which limits its applicability to scenarios with available GCM output and introduces dependencies on a specific set of climate model simulations, but enables higher emulation accuracy within the ISMIP6 ensemble. We present ISEFlow as a complementary tool, consistent with a broader trend toward diverse input variables in Earth system emulators (Jouvet and Cordonnier, 2023; Li et al., 2024; Guo et al., 2024).

Although ISEFlow demonstrates significant advances in emulator accuracy and uncertainty quantification, there are several limitations to this study that present opportunities for future improvement. First, the training data used to develop ISEFlow is based solely on ISMIP6 simulations, which, while extensive, represent a limited subset of possible climate forcings and ice sheet model configurations. This restricts the emulator's ability to generalize beyond the specific experimental conditions provided by ISMIP6, potentially limiting its applicability to future or more extreme climate scenarios. The current dataset also does not include enough simulations with varying ISM characteristics to fully disentangle their effects on model outputs. For example, bedrock topography and stress balance are known to be important factors in accurate dynamical ice sheet modeling, but these characteristics do not vary often or substantially within the ISMIP6 ensemble data used to train ISEFlow. It is also important to clarify that for ISEFlow to be used for policy-relevant sea level projections, additional steps are required beyond the scope of this work. Some of these changes include the ability to choose prior probability distributions over model parameters and scenarios, broader uncertainty sampling (encompassing deep uncertainties not represented in ISMIP6), accounting for correlations between ice sheet contributions, and combining with other sea level rise components.

Additionally, while we argue that the physics underlying the ice sheet behavior is embedded in the data, as ISMIP6 is composed of physics-based models, ISEFlow itself does not explicitly encode physical constraints. This data-driven approach, while effective for emulation, may fail to capture phenomena outside the range of training data or enforce physical laws not well-represented in the data. FurthermoreAdditionally, ML-based models like ISEFlow are inherently challenging to interpret, making it difficult to fully understand how predictions are generated. While we employ SHAP as a tool to investigate the

model's behavior and identify important features, SHAP provides an approximation of feature importance rather than causal relationships. Finally, although ISEFlow quantifies emulator and data coverage uncertainties, it does not address other significant sources of uncertainty, such as those arising from structural errors in the ice sheet models, the handling of MICI, external forcing uncertainties, or potential feedbacks between processes not captured by the dataset. ~~Addressing these limitations in future work~~Improving in these areas, such as by expanding the scope of simulations, incorporating additional sources of uncertainty, or integrating physics-informed constraints, will be crucial for improving the robustness and applicability of ISEFlow in sea level rise projections.

5 Conclusions

This study presents ISEFlow, a NN-based ice sheet emulator, and demonstrates its ability to ~~produce accurate~~ accurately emulate projections of future sea levels. We compare ISEFlow to state-of-the-art emulators and show that ISEFlow creates more ~~accurate~~ accurately emulated ISMIP6 ice sheet model projections, is faster to run, and is able to capture climate sensitivity to emission scenarios. The use of SHAP analysis provides a clear understanding of the key drivers of ISEFlow's predictions, revealing which input features and ISM characteristics most influence the accuracy of sea level projections. This not only strengthens confidence in ISEFlow~~ΓÇÖ~~'s predictions but also offers valuable insights into ice sheet model characteristics and their impact on sea level projections. Another significant ~~contributions~~ contribution of this study is demonstrating how the emulator can efficiently capture complex ice sheet dynamics while maintaining the ability to quantify and separate sources of uncertainty. ISEFlow~~ΓÇÖ~~'s ability to quantify both data coverage and emulator uncertainty separately ensures that areas of ISM disagreement are properly accounted for, reducing the risk of overconfidence in projections where training data ~~is~~ are variable. This hybrid uncertainty quantification provides a more detailed estimate of future sea level changes, making ISEFlow a powerful tool for scientific research.

There are important opportunities for future work that could further enhance the applicability and impact of ISEFlow. ~~In the IPCC AR6, Emulandice relied on temperature forcings from the Finite amplitude Impulse Response (FaIR) simple climate model (Smith et al., 2018) because it represents a wider distribution of possible forcing distributions based on more CMIP models than the six used in ISMIP6. In this study, however, we use the same set of forcings from the six CMIP models used in ISMIP6 to build each emulator, which enabled a direct comparison between ISEFlow , Emulandice, and LARMIP. Future work should extend ISEFlow by applying it~~ Extending ISEFlow to a larger selection of CMIP forcings beyond those included in ~~ISMIP, which could provide a more comprehensive range of projections and ensure that ISEFlow remains adaptable~~ ISMIP6 would broaden the range of achievable projections and improve adaptability to future climate scenarios. ~~By applying ISEFlow to a broader set of forcings, we may uncover new insights into the potential differences between emission scenariosthat were previously underrepresented~~A temperature-only version of ISEFlow suitable for coupling with simple climate models such as FaIR (Smith et al., 2018) or FACTS (Kopp et al., 2023) represents a promising path toward broader applicability across emissions scenarios. Future work could also ~~include the further analysis of the quantified uncertainty from the emulator. Detailed experimentation could be carried out to understand where the emulated uncertainty originates and what~~ examine the

quantified uncertainty in greater depth, investigating which forcings contribute most to ~~quantified uncertainty~~ each uncertainty component and how the decomposition evolves across the projection period.

ISEFlow is also able to contribute to future ice sheet projections, including those generated for ISMIP7, which will include different experiments, different models, and offer a new set of sea level projections to the year 2300. As shown in ISMIP6 Antarctica projections to 2300 (Seroussi et al., 2024), these longer time horizons offer an opportunity to explore how the ice sheet will respond to various climate scenarios over extended periods. Retraining the model on this updated dataset will enable us to evaluate how ice sheet dynamics and sensitivities evolve beyond 2100, as widespread retreat and collapse of some West Antarctic basins are simulated with 30-40% of the ISMIP6 ensemble by 2300. This future work will not only broaden the scope of ISEFlow~~ΓÇÖ~~'s applications but also reinforce its value as a key tool for understanding ice sheet dynamics.

Code and data availability. The processed datasets and code needed to reproduce the ISMIP6 data processing, model training, evaluation, and results for both ISEFlow-AIS and ISEFlow-GrIS are archived in a Zenodo repository found at https://doi.org/10.5281/zenodo.14908114 (Van Katwyk et al., 2025). The original ISMIP6 datasets, from which ISEFlow training data was processed, can be found at the following: https://doi.org/10.5281/zenodo.11176009 (Forcings), https://doi.org/10.5281/zenodo.11176023 (GrIS outputs), and https://doi.org/10.5281/zenodo.11176027 (AIS outputs).

Author contributions. PV conceptualized the study, designed the research methodology, and developed and evaluated the emulator. KJB and BFK contributed to the research methodology, model development, and evaluation. HS and SN curated the datasets and provided additional scientific background and insights. All co-authors contributed to the interpretation of results and discussion. PV drafted the original manuscript with substantial input from all co-authors. All authors reviewed, edited, and approved the final manuscript.

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

Acknowledgements. PV is supported by the National Science Foundation Graduate Research Fellowship Program under Grant 2040433. BFK was supported by the Brown University Equitable Climate Futures initiative and Schmidt Futures ΓÇô a philanthropic initiative that seeks to improve societal outcomes through the development of emerging science and technologies. SN was supported by the NASA Sea Level Change Team and a philanthropic gift from the Schmidt Family. HS was supported by grants from NASA Cryospheric Science (80NSSC22K0383), NASA Sea Level Change Team (80NSSC24K1532), and Novo Nordisk Foundation under the Challenge Programme 2023 (Grant number NNF23OC00807040). KJB was supported by the SciAI Center, and funded by the Office of Naval Research (ONR), under Grant Number N00014-23-1-2729. ~~AI assistance was~~ The authors used generative AI tools to assist with copy editing and formatting the submitted code; all scientific content, analysis, and interpretations were developed and verified by the authors. AI assistance was also used for initial citation formatting; all entries have been manually verified before submission. Computational resources and services required

for this study were provided by the Center for Computation and Visualization, Brown University. We also thank the Climate and Cryosphere (CliC) and their efforts to host ISMIP6, along with the World Climate Research Programme, which 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 all those involved with ISMIP6 for making this research possible. The authors thank ~~... and ... anonymous reviewers for their~~ Chris Smith, Denis Felikson, and Tamsin Edwards for their invaluable feedback on the manuscript. This is ISMIP6 contribution number 34.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623–2631, https://doi.org/10.1145/3292500.3330701, 2019.

Aschwanden, A., Fahnestock, M. A., Truffer, M., Brinkerhoff, D. J., Hock, R., Khroulev, C., Mottram, R., and Khan, S. A.: Contribution of the Greenland Ice Sheet to sea level over the next millennium, Science Advances, 5, eaav9396, https://doi.org/10.1126/sciadv.aav9396, 2019.

Bamber, J. L., Oppenheimer, M., Kopp, R. E., Aspinall, W. P., and Cooke, R. M.: Ice sheet contributions to future sea-level rise from structured expert judgment, Proceedings of the National Academy of Sciences, 116, 11 195–11 200, https://doi.org/10.1073/pnas.1817205116, 2019.

Bergen, K. J., Johnson, P. A., de Hoop, M. V., and Beroza, G. C.: Machine learning for data-driven discovery in solid Earth geoscience, Science, 363, eaau0323, https://doi.org/10.1126/science.aau0323, 2019.

Beusch, L., Gudmundsson, L., and Seneviratne, S. I.: Emulating Earth system model temperatures with MESMER: from global mean temperature trajectories to grid-point-level realizations on land, Earth System Dynamics, 11, 139–159, https://doi.org/10.5194/esd-11-139-2020, 2020.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, 619, 533–538, https://doi.org/10.1038/s41586-023-06185-3, 2023.

Bochenek, B. and Ustrnul, Z.: Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives, Atmosphere, 13, 180, https://doi.org/10.3390/atmos13020180, 2022.

Coulon, V., Klose, A. K., Kittel, C., Edwards, T., Turner, F., Winkelmann, R., and Pattyn, F.: Disentangling the drivers of future Antarctic ice loss with a historically calibrated ice-sheet model, The Cryosphere, 18, 653–681, https://doi.org/10.5194/tc-18-653-2024, 2024.

Coulon, V., Klose, A. K., Edwards, T., Turner, F., Pattyn, F., and Winkelmann, R.: From short-term uncertainties to long-term certainties in the future evolution of the Antarctic Ice Sheet, Nature Communications, 16, 10 385, https://doi.org/10.1038/s41467-025-66178-w, 2025.

de Burgh-Day, C. O. and Leeuwenburg, T.: Machine learning for numerical weather and climate modelling: a review, Geoscientific Model Development, 16, 6433–6477, https://doi.org/10.5194/gmd-16-6433-2023, 2023.

DeConto, R. M. and Pollard, D.: Contribution of Antarctica to past and future sea-level rise, Nature, 531, 591–597, https://doi.org/10.1038/nature17145, 2016.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G.: nflows: normalizing flows in PyTorch, https://doi.org/10.5281/zenodo.4296287, 2020.

Edwards, T. L., Nowicki, S., Marzeion, B., Hock, R., Goelzer, H., Seroussi, H., Jourdain, N. C., Slater, D. A., Turner, F. E., Smith, C. J., et al.: Projected land ice contributions to twenty-first-century sea level rise, Nature, 593, 74–82, https://doi.org/10.1038/s41586-021-03302-y, 2021.

Fettweis, X., Franco, B., Tedesco, M., van Angelen, J. H., Lenaerts, J. T. M., van den Broeke, M. R., and Gallée, H.: Estimating the Greenland ice sheet surface mass balance contribution to future sea level rise using the regional atmospheric climate model MAR, The Cryosphere, 7, 469–489, https://doi.org/10.5194/tc-7-469-2013, 2013.

Fox-Kemper, B., Hewitt, H., Xiao, C., Aðalgeirsdóttir, G., Drijfhout, S., Edwards, T., Golledge, N., Hemer, M., Kopp, R., Krinner, G., Mix, A., Notz, D., Nowicki, S., Nurhati, I., Ruiz, L., Sallée, J.-B., Slangen, A., and Yu, Y.: Ocean, Cryosphere

and Sea Level Change, p. 1211–1362, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896.011, 2021.

Gholami, H., Darvishi, E., Moradi, N., Mohammadifar, A., Song, Y., Li, Y., Niu, B., Kaskaoutis, D., and Pradhan, B.: An interpretable (explainable) model based on machine learning and SHAP interpretation technique for mapping wind erosion hazard, Environmental Science and Pollution Research, 31, 64 628–64 643, https://doi.org/10.1007/s11356-024-35521-x, 2024.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association, 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Goelzer, H., Robinson, A., Seroussi, H., and Van De Wal, R. S.: Recent Progress in Greenland Ice Sheet Modelling, Current Climate Change Reports, 3, 291–302, https://doi.org/10.1007/s40641-017-0073-y, 2017.

Goelzer, H., Nowicki, S., Payne, A., Larour, E., Seroussi, H., Lipscomb, W. H., Gregory, J., Abe-Ouchi, A., Shepherd, A., Simon, E., Agosta, C., Alexander, P., Aschwanden, A., Barthel, A., Calov, R., Chambers, C., Choi, Y., Cuzzone, J., Dumas, C., Edwards, T., Felikson, D., Fettweis, X., Golledge, N. R., Greve, R., Humbert, A., Huybrechts, P., Le clec'h, S., Lee, V., Leguy, G., Little, C., Lowry, D. P., Morlighem, M., Nias, I., Quiquet, A., Rückamp, M., Schlegel, N.-J., Slater, D. A., Smith, R. S., Straneo, F., Tarasov, L., van de Wal, R., and van den Broeke, M.: The future sea-level contribution of the Greenland ice sheet: a multi-model ensemble study of ISMIP6, The Cryosphere, 14, 3071–3096, https://doi.org/10.5194/tc-14-3071-2020, 2020.

Golledge, N. R., Kowalewski, D. E., Naish, T. R., Levy, R. H., Fogwill, C. J., and Gasson, E. G. W.: The multi-millennial Antarctic commitment to future sea-level rise, Nature, 526, 421–425, https://doi.org/10.1038/nature15706, 2015.

Gu, M., Wang, X., and Berger, J. O.: Robust Gaussian stochastic process emulation, The Annals of Statistics, 46, 3038 – 3066, https://doi.org/10.1214/17-AOS1648, 2018.

Guo, Z., Lyu, P., Ling, F., Luo, J.-J., Boers, N., Ouyang, W., and Bai, L.: ORCA: A Global Ocean Emulator for Multi-year to Decadal Predictions, arXiv e-prints, pp. arXiv–2405, https://doi.org/https://doi.org/10.48550/arXiv.2405.15412, 2024.

Hawkins, E. and Sutton, R.: The Potential to Narrow Uncertainty in Regional Climate Predictions, Bulletin of the American Meteorological Society, 90, 1095–1108, https://doi.org/10.1175/2009BAMS2607.1, 2009.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Hüllermeier, E. and Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, Machine Learning, 110, 457–506, https://doi.org/10.1007/s10994-021-05946-3, 2021.

Jourdain, N. C., Asay-Davis, X., Hattermann, T., Straneo, F., Seroussi, H., Little, C. M., and Nowicki, S.: A protocol for calculating basal melt rates in the ISMIP6 Antarctic ice sheet projections, The Cryosphere, 14, 3111–3134, https://doi.org/10.5194/tc-14-3111-2020, 2020.

Jouvet, G. and Cordonnier, G.: Ice-flow model emulator based on physics-informed deep learning, Journal of Glaciology, 69, 1941–1955, https://doi.org/10.1017/jog.2023.73, 2023.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., and Kumar, V.: Machine Learning for the Geosciences: Challenges and Opportunities, IEEE Transactions on Knowledge and Data Engineering, 31, 1544–1554, https://doi.org/10.1109/TKDE.2018.2861006, 2019.

Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., et al.: Physics-informed machine learning: case studies for weather and climate modelling, Philosophical Transactions of the Royal Society A, 379, 20200 093, https://doi.org/10.1098/rsta.2020.0093, 2021.

Khosravi, A., Nahavandi, S., Creighton, D., and Atiya, A. F.: Comprehensive Review of Neural Network-Based Prediction Intervals and New Advances, IEEE Transactions on Neural Networks, 22, 1341–1356, https://doi.org/10.1109/TNN.2011.2162110, 2011.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., et al.: Neural general circulation models for weather and climate, Nature, 632, 1060–1066, https://doi.org/10.1038/s41586-024-07744-y, 2024.

Kopp, R. E., Oppenheimer, M., O'Reilly, J. L., Drijfhout, S. S., Edwards, T. L., Fox-Kemper, B., Garner, G. G., Golledge, N. R., Hermans, T. H., Hewitt, H. T., et al.: Communicating future sea-level rise uncertainty and ambiguity to assessment users, Nature Climate Change, 13, 648–660, https://doi.org/10.1038/s41558-023-01691-8, 2023.

Kullback, S. and Leibler, R. A.: On Information and Sufficiency, The Annals of Mathematical Statistics, 22, 79–86, http://www.jstor.org/stable/2236703, 1951.

Lakshminarayanan, B., Pritzel, A., and Blundell, C.: Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles, in: Advances in Neural Information Processing Systems, pp. 6402–6413, 2017.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521, 436–444, 2015.

Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E. M., Brunner, L., Knutti, R., and Hawkins, E.: Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6, Earth System Dynamics, 11, 491–508, https://doi.org/10.5194/esd-11-491-2020, 2020.

Levermann, A., Winkelmann, R., Albrecht, T., Goelzer, H., Golledge, N. R., Greve, R., Huybrechts, P., Jordan, J., Leguy, G., Martin, D., Morlighem, M., Pattyn, F., Pollard, D., Quiquet, A., Rodehacke, C., Seroussi, H., Sutter, J., Zhang, T., Van Breedam, J., Calov, R., DeConto, R., Dumas, C., Garbe, J., Gudmundsson, G. H., Hoffman, M. J., Humbert, A., Kleiner, T., Lipscomb, W. H., Meinshausen, M., Ng, E., Nowicki, S. M. J., Perego, M., Price, S. F., Saito, F., Schlegel, N.-J., Sun, S., and van de Wal, R. S. W.: Projecting Antarctica's contribution to future sea level rise from basal ice shelf melt using linear response functions of 16 ice sheet models (LARMIP-2), Earth System Dynamics, 11, 35–76, https://doi.org/10.5194/esd-11-35-2020, 2020.

Li, L., Carver, R., Lopez-Gomez, I., Sha, F., and Anderson, J.: Generative emulation of weather forecast ensembles with diffusion models, Science Advances, 10, eadk4489, https://doi.org/10.1126/sciadv.adk4489, 2024.

Lin, J.: Divergence measures based on the Shannon entropy, IEEE Transactions on Information Theory, 37, 145–151, https://doi.org/10.1109/18.61115, 1991.

Lowry, D. P., Krapp, M., Golledge, N. R., and Alevropoulos-Borrill, A.: The influence of emissions scenarios on future Antarctic ice loss is unlikely to emerge this century, Communications Earth & Environment, 2, 221, https://doi.org/10.1038/s43247-021-00289-2, 2021.

Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, Curran Associates, Inc., 2017.

Massey Jr, F. J.: The Kolmogorov-Smirnov Test for Goodness of Fit, Journal of the American Statistical Association, 46, 68–78, https://doi.org/10.1080/01621459.1951.10500769, 1951.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B.: Hybrid models with deep and invertible features, in: International Conference on Machine Learning, pp. 4723–4732, PMLR, 2019.

Nicklas, J. M., Fox-Kemper, B., and Lawrence, C.: Efficient Estimation of Climate State and Its Uncertainty Using Kalman Filtering with Application to Policy Thresholds and Volcanism, Journal of Climate, 38, 1235–1270, https://doi.org/10.1175/JCLI-D-23-0580.1, 2025.

Nowicki, S., Goelzer, H., Seroussi, H., Payne, A. J., Lipscomb, W. H., Abe-Ouchi, A., Agosta, C., Alexander, P., Asay-Davis, X. S., Barthel, A., Bracegirdle, T. J., Cullather, R., Felikson, D., Fettweis, X., Gregory, J. M., Hattermann, T., Jourdain, N. C., Kuipers Munneke, P.,

Larour, E., Little, C. M., Morlighem, M., Nias, I., Shepherd, A., Simon, E., Slater, D., Smith, R. S., Straneo, F., Trusel, L. D., van den Broeke, M. R., and van de Wal, R.: Experimental protocol for sea level projections from ISMIP6 stand-alone ice sheet models, The Cryosphere, 14, 2331–2368, https://doi.org/10.5194/tc-14-2331-2020, 2020.

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, Geoscientific Model Development, 9, 4521–4545, https://doi.org/10.5194/gmd-9-4521-2016, 2016.

Oppenheimer, M., Glavovic, B. C., Hinkel, J., van de Wal, R., Magnan, A. K., Abd-Elgawad, A., Cai, R., et al.: Sea Level Rise and Implications for Low-Lying Islands, Coasts and Communities, IPCC Special Report on the Ocean and Cryosphere in a Changing Climate, 2019.

Papamakarios, G., Pavlakou, T., and Murray, I.: Masked autoregressive flow for density estimation, in: Advances in Neural Information Processing Systems, pp. 2338–2347, 2017.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, Advances in Neural Information Processing Systems, 32, 2019.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., and Anandkumar, A.: FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators, arXiv preprint arXiv:2202.11214, https://doi.org/10.48550/arXiv.2202.11214, 2022.

Pattyn, F., Favier, L., Sun, S., and Durand, G.: Progress in Numerical Modeling of Antarctic Ice-Sheet Dynamics, Current Climate Change Reports, 3, 174–184, https://doi.org/10.1007/s40641-017-0069-7, 2017.

Payne, A. J., Nowicki, S., Abe-Ouchi, A., Agosta, C., Alexander, P., Albrecht, T., Asay-Davis, X., Aschwanden, A., Barthel, A., Bracegirdle, T. J., Calov, R., Chambers, C., Choi, Y., Cullather, R., Cuzzone, J., Dumas, C., Edwards, T. L., Felikson, D., Fettweis, X., Galton-Fenzi, B. K., Goelzer, H., Gladstone, R., Golledge, N. R., Gregory, J. M., Greve, R., Hattermann, T., Hoffman, M. J., Humbert, A., Huybrechts, P., Jourdain, N. C., Kleiner, T., Munneke, P. K., Larour, E., Le clec'h, S., Lee, V., Leguy, G., Lipscomb, W. H., Little, C. M., Lowry, D. P., Morlighem, M., Nias, I., Pattyn, F., Pelle, T., Price, S. F., Quiquet, A., Reese, R., Rückamp, M., Schlegel, N.-J., Seroussi, H., Shepherd, A., Simon, E., Slater, D., Smith, R. S., Straneo, F., Sun, S., Tarasov, L., Trusel, L. D., Van Breedam, J., van de Wal, R., van den Broeke, M., Winkelmann, R., Zhao, C., Zhang, T., and Zwinger, T.: 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, e2020GL091 741, https://doi.org/https://doi.org/10.1029/2020GL091741, e2020GL091741 2020GL091741, 2021.

Pearce, T., Zaki, M., Brintrup, A., Anastassacos, N., and Neely, A.: Uncertainty in Neural Networks: Bayesian ensembling, stat, 1050, 12, https://doi.org/10.48550/arXiv.1810.05546, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine learning in Python, The Journal of Machine Learning Research, 12, 2825–2830, 2011.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat, F.: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Rezende, D. and Mohamed, S.: Variational inference with normalizing flows, in: International Conference on Machine Learning, pp. 1530–1538, PMLR, 2015.

Rignot, E., Mouginot, J., Scheuchl, B., Van Den Broeke, M., Van Wessem, M. J., and Morlighem, M.: Four decades of Antarctic Ice Sheet mass balance from 1979–2017, Proceedings of the National Academy of Sciences, 116, 1095–1103, https://doi.org/10.1073/pnas.1812883116, 2019.

Rohmer, J., Thieblemont, R., Le Cozannet, G., Goelzer, H., and Durand, G.: Improving interpretation of sea-level projections through a machine-learning-based local explanation approach, The Cryosphere, 16, 4637–4657, https://doi.org/10.5194/tc-16-4637-2022, 2022.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P.: Design and Analysis of Computer Experiments, Statistical Science, 4, 409–423, 1989.

Sane, A., Reichl, B. G., Adcroft, A., and Zanna, L.: Parameterizing Vertical Mixing Coefficients in the Ocean Surface Boundary Layer Using Neural Networks, Journal of Advances in Modeling Earth Systems, 15, e2023MS003 890, https://doi.org/10.1029/2023MS003890, e2023MS003890 2023MS003890, 2023.

Seroussi, H., Nowicki, S., Payne, A. J., Goelzer, H., Lipscomb, W. H., Abe-Ouchi, A., Agosta, C., Albrecht, T., Asay-Davis, X., Barthel, A., Calov, R., Cullather, R., Dumas, C., Galton-Fenzi, B. K., Gladstone, R., Golledge, N. R., Gregory, J. M., Greve, R., Hattermann, T., Hoffman, M. J., Humbert, A., Huybrechts, P., Jourdain, N. C., Kleiner, T., Larour, E., Leguy, G. R., Lowry, D. P., Little, C. M., Morlighem, M., Pattyn, F., Pelle, T., Price, S. F., Quiquet, A., Reese, R., Schlegel, N.-J., Shepherd, A., Simon, E., Smith, R. S., Straneo, F., Sun, S., Trusel, L. D., Van Breedam, J., van de Wal, R. S. W., Winkelmann, R., Zhao, C., Zhang, T., and Zwinger, T.: ISMIP6 Antarctica: a multi-model ensemble of the Antarctic ice sheet evolution over the 21st century, The Cryosphere, 14, 3033–3070, https://doi.org/10.5194/tc-14-3033-2020, 2020.

Seroussi, H., Verjans, V., Nowicki, S., Payne, A. J., Goelzer, H., Lipscomb, W. H., Abe-Ouchi, A., Agosta, C., Albrecht, T., Asay-Davis, X., Barthel, A., Calov, R., Cullather, R., Dumas, C., Galton-Fenzi, B. K., Gladstone, R., Golledge, N. R., Gregory, J. M., Greve, R., Hattermann, T., Hoffman, M. J., Humbert, A., Huybrechts, P., Jourdain, N. C., Kleiner, T., Larour, E., Leguy, G. R., Lowry, D. P., Little, C. M., Morlighem, M., Pattyn, F., Pelle, T., Price, S. F., Quiquet, A., Reese, R., Schlegel, N.-J., Shepherd, A., Simon, E., Smith, R. S., Straneo, F., Sun, S., Trusel, L. D., Van Breedam, J., Van Katwyk, P., van de Wal, R. S. W., Winkelmann, R., Zhao, C., Zhang, T., and Zwinger, T.: Insights into the vulnerability of Antarctic glaciers from the ISMIP6 ice sheet model ensemble and associated uncertainty, Cryosphere, 17, 5197–5217, https://doi.org/10.5194/tc-17-5197-2023, 2023.

Seroussi, H., Pelle, T., Lipscomb, W. H., Abe-Ouchi, A., Albrecht, T., Alvarez-Solas, J., Asay-Davis, X., Barre, J.-B., Berends, C. J., Bernales, J., Blasco, J., Caillet, J., Chandler, D. M., Coulon, V., Cullather, R., Dumas, C., Galton-Fenzi, B. K., Garbe, J., Gillet-Chaulet, F., Gladstone, R., Goelzer, H., Golledge, N., Greve, R., Gudmundsson, G. H., Han, H. K., Hillebrand, T. R., Hoffman, M. J., Huybrechts, P., Jourdain, N. C., Klose, A. K., Langebroek, P. M., Leguy, G. R., Lowry, D. P., Mathiot, P., Montoya, M., Morlighem, M., Nowicki, S., Pattyn, F., Payne, A. J., Quiquet, A., Reese, R., Robinson, A., Saraste, L., Simon, E. G., Sun, S., Twarog, J. P., Trusel, L. D., Urruty, B., Van Breedam, J., van de Wal, R. S. W., Wang, Y., Zhao, C., and Zwinger, T.: Evolution of the Antarctic Ice Sheet Over the Next Three Centuries From an ISMIP6 Model Ensemble, Earth's Future, 12, e2024EF004 561, https://doi.org/10.1029/2024EF004561, e2024EF004561 2024EF004561, 2024.

Shepherd, A., Ivins, E., Rignot, E., Smith, B., van den Broeke, M., Velicogna, I., Whitehouse, P., Briggs, K., Joughin, I., Krinner, G., Nowicki, S., Payne, T., Scambos, T., Schlegel, N., A, G., Agosta, C., Ahlstrøm, A., Babonis, G., Barletta, V., Blazquez, A., Bonin, J., Csatho, B., Cullather, R., Felikson, D., Fettweis, X., Forsberg, R., Gallee, H., Gardner, A., Gilbert, L., Groh, A., Gunter, B., Hanna, E., Harig, C., Helm, V., Horvath, A., Horwath, M., Khan, S., Kjeldsen, K. K., Konrad, H., Langen, P., Lecavalier, B., Loomis, B., Luthcke, S., McMillan, M., Melini, D., Mernild, S., Mohajerani, Y., Moore, P., Mouginot, J., Moyano, G., Muir, A., Nagler, T., Nield, G., Nilsson, J., Noel, B., Otosaka, I., Pattle, M. E., Peltier, W. R., Pie, N., Rietbroek, R., Rott, H., Sandberg-Sørensen, L., Sasgen, I., Save, H., Scheuchl,

B., Schrama, E., Schröder, L., Seo, K.-W., Simonsen, S., Slater, T., Spada, G., Sutterley, T., Talpe, M., Tarasov, L., van de Berg, W. J., van der Wal, W., van Wessem, M., Vishwakarma, B. D., Wiese, D., Wouters, B., and The IMBIE team: Mass balance of the Antarctic Ice Sheet from 1992 to 2017, Nature, 558, 219–222, https://doi.org/10.1038/s41586-018-0179-y, 2018.

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

Smith, C. J., Forster, P. M., Allen, M., Leach, N., Millar, R. J., Passerello, G. A., and Regayre, L. A.: FAIR v1.3: a simple emissions-based impulse response and carbon cycle model, Geoscientific Model Development, 11, 2273–2297, https://doi.org/10.5194/gmd-11-2273-2018, 2018.

Tibshirani, R. J. and Efron, B.: An introduction to the bootstrap, Monographs on statistics and applied probability, 57, 1–436, 1993.

Van Katwyk, P., Fox-Kemper, B., Seroussi, H., Nowicki, S., and Bergen, K. J.: A Variational LSTM Emulator of Sea Level Contribution From the Antarctic Ice Sheet, Journal of Advances in Modeling Earth Systems, 15, e2023MS003 899, https://doi.org/10.1029/2023MS003899, e2023MS003899 2023MS003899, 2023.

Van Katwyk, P., Fox-Kemper, B., Nowicki, S., Seroussi, H., and Bergen, K.: Code for "ISEFlow: A Flow-Based Neural Network Emulator for Improved Sea Level Projections and Uncertainty Quantification" (v1.0.0), https://doi.org/10.5281/zenodo.14908114, 2025.

Weyn, J. A., Durran, D. R., Caruana, R., and Cresswell-Clay, N.: Sub-Seasonal Forecasting With a Large Ensemble of Deep-Learning Weather Prediction Models, Journal of Advances in Modeling Earth Systems, 13, e2021MS002 502, https://doi.org/10.1029/2021MS002502, e2021MS002502 2021MS002502, 2021.

Winkler, R. L.: A Decision-Theoretic Approach to Interval Estimation, Journal of the American Statistical Association, 67, 187–191, https://doi.org/10.1080/01621459.1972.10481224, 1972.

Yang, R., Hu, J., Li, Z., Mu, J., Yu, T., Xia, J., Li, X., Dasgupta, A., and Xiong, H.: Interpretable machine learning for weather and climate prediction: A review, Atmospheric Environment, 338, 120 797, https://doi.org/10.1016/j.atmosenv.2024.120797, 2024.

Appendix A: Data

A1 Ice Sheet Sector Maps

Climate forcings and outputted SLE from ISMIP6 are aggregated into distinct sectors on each the Antarctic and Greenland ice sheets. These sectors are based on those used in ISMIP6 (Nowicki et al., 2016), and are seen in Figures A1 and A2.

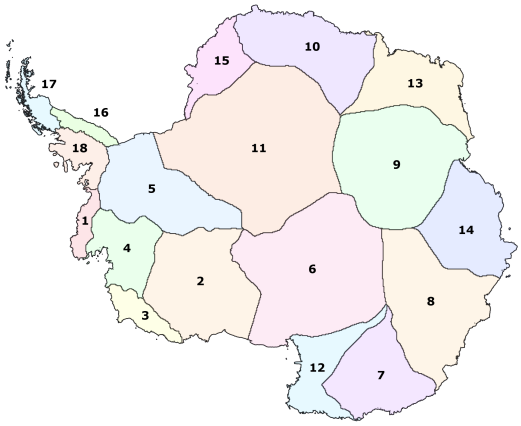

Figure A1. Map of Antarctic Ice Sheet sectors as used in ISMIP6 AIS (Seroussi et al., 2020).

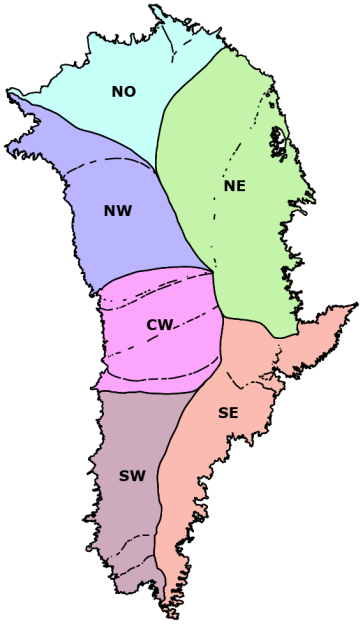

Figure A2. Map of Greenland Ice Sheet sectors as used in ISMIP6 GrIS (Goelzer et al., 2020).

A2 ISEFlow Input Variables

Table A1. Input variables to the ISEFlow-AIS neural network emulator. Climate forcings are annually averaged sector-level anomalies from the ISMIP6 Antarctic forcing dataset, conservatively regridded from the native GCM grid to the ice sheet model sectors. Ocean temperature, salinity, and thermal forcing fields are derived from CMIP outputs and extrapolated to unresolved regions, including continental shelf areas and sub-ice-shelf cavities, to provide complete boundary forcing (Jourdain et al., 2020). Each temporal forcing includes the current year and five lagged years (six total features per forcing). Categorical ice sheet model (ISM) characteristics are one-hot encoded. The final feature set contains 98 input dimensions. For all categorical variable values, refer to Van Katwyk et al. (2023).

Feature	Name	Units	Temporal	Type	Number
Projection Year	year	–	Temporal	Numeric	1
Sector	sector	–	Static	Numeric	1
Precipitation Anomaly	pr_anomaly	$\frac{kg}{m^2 \cdot s}$	Temporal	Numeric	6
Evaporation Anomaly	evspsbl_anomaly	$\frac{kg}{m^2 \cdot s}$	Temporal	Numeric	6
Runoff Anomaly	mrro_anomaly	$\frac{kg}{m^2 \cdot s}$	Temporal	Numeric	6
Surface Mass Balance	smb_anomaly	$\frac{kg}{m^2 \cdot s}$	Temporal	Numeric	6
Surface Temperature	ts_anomaly	K	Temporal	Numeric	6
Thermal Forcing	thermal_forcing	°C	Temporal	Numeric	6
Salinity	salinity	PSU	Temporal	Numeric	6
Ocean Temperature	temperature	°C	Temporal	Numeric	6
Numerics	numerics_*	–	Static	One-hot	3
Initialization Year	initial_year	year	Static	Numeric	1
Stress Balance	stress_balance_*	–	Static	One-hot	6
Resolution	resolution_*	km	Static	One-hot	6
Initialization Method	init_method_*	–	Static	One-hot	6
Melt Treatment	melt_*	–	Static	One-hot	3

Table A2. Input variables to the ISEFlow-GrIS neural network emulator. Climate forcings are annually averaged sector-level anomalies from the ISMIP6 Greenland forcing dataset, derived from regionally downscaled climate products. Ocean forcing is represented through prescribed tidewater glacier retreat parameterizations driven by runoff and ocean temperature changes, alongside open-framework simulations that apply direct ocean temperature and runoff forcing (Slater et al., 2020). Each temporal forcing includes the current year and five lagged years (six total features per forcing). Categorical ice sheet model (ISM) characteristics are one-hot encoded. The final feature set contains 90 input dimensions. For all categorical variable values, refer to Van Katwyk et al. (2023).

Feature	Name	Units	Temporal	Type	Number
Projection Year	year	–	Temporal	Numeric	1
Sector	sector	–	Static	Numeric	1
Surface Mass Balance Anomaly	aSMB	$\frac{kg}{m^2 \cdot s}$	Temporal	Numeric	6
Surface Temperature Anomaly	aST	K	Temporal	Numeric	6
Thermal Forcing	thermal_forcing	°C	Temporal	Numeric	6
Basin Runoff	basin_runoff	–	Temporal	Numeric	6
Initialization Year	initial_year	year	Static	Numeric	1
Numerics	numerics_*	–	Static	One-hot	4
Ice Flow Approximation	ice_flow_*	–	Static	One-hot	4
Initialization Method	initialization_*	–	Static	One-hot	9
Initial SMB Product	initial_smb_*	–	Static	One-hot	7
Velocity Product	velocity_*	–	Static	One-hot	2
Bed Topography Product	bed_*	–	Static	One-hot	2
Surface Thickness Product	surface_thickness_*	–	Static	One-hot	1
Geothermal Heat Flux Product	ghf_*	–	Static	One-hot	3
Minimum Resolution	res_min_*	km	Static	One-hot	13

A3 Sensitivity Test and Regional ISEFlow Input Variables

Table A3. Input variables for the ISEFlow-AIS and ISEFlow-GrIS emulators under the three forcing sensitivity configurations reported in Table 3: All available forcings (All), surface air temperature only (Temperature), and surface mass balance only (SMB), as well as the basal melt parameterization, retreat parameterization (κ), and ice shelf fracture. The "All" configuration corresponds to the same climate forcings used in the full ISEFlow emulators (Tables A1 and A2), but excludes ISM characteristics.

Feature	Units	Temporal	All	Temp.	SMB
AIS					
Projection Year	–	Temporal	✓	✓	✓
Sector	–	Static	✓	✓	✓
Basal Melt Param.	–	Static	✓	✓	✓
Ice shelf fracture	–	Static	✓	✓	✓
Precipitation Anomaly	$\frac{kg}{m^2 \cdot s}$	Temporal	✓		
Evaporation Anomaly	$\frac{kg}{m^2 \cdot s}$	Temporal	✓		
Runoff Anomaly	$\frac{kg}{m^2 \cdot s}$	Temporal	✓		
Surface Mass Balance	$\frac{kg}{m^2 \cdot s}$	Temporal	✓		✓
Surface Temperature	K	Temporal	✓	✓	
Thermal Forcing	°C	Temporal	✓		
Salinity	PSU	Temporal	✓		
Ocean Temperature	°C	Temporal	✓		
GrIS					
Projection Year	–	Temporal	✓	✓	✓
Sector	–	Static	✓	✓	✓
Retreat sensitivity	–	Static	✓	✓	✓
Surface Mass Balance Anomaly	$\frac{kg}{m^2 \cdot s}$	Temporal	✓		✓

Table A4. Input variables for Regional ISEFlow, trained on the same dataset and input variables as the GP emulator (following a similar design to Edwards et al. (2021)) to provide a fair architectural comparison in Table 4. For the AIS, projections are aggregated into three regions (East Antarctica, West Antarctica, and Antarctic Peninsula; Shepherd et al. 2018) rather than 18 sectors. For the GrIS, projections are aggregated into a single region rather than six sectors. Regional ISEFlow uses surface air temperature as the sole temporal climate forcing, plus categorical ISM parameters representing glacier retreat, sub-shelf basal melt, and ice shelf fracture. Categorical variables are one-hot encoded.

Feature	Units	Temporal	Type	Number
AIS (3 aggregated regions)				
Projection Year	–	Temporal	Numeric	1
Region	–	Static	Numeric	1
Surface Temperature	K	Temporal	Numeric	1
Basal Melt Param.	–	Static	One-hot	4
Ice Shelf Fracture	–	Static	One-hot	2
			Total Features	**9**
GrIS (1 aggregated region)				
Projection Year	–	Temporal	Numeric	1
Region	–	Static	Numeric	1
Surface Temperature	K	Temporal	Numeric	1
Retreat Sensitivity	–	Static	One-hot	3
			Total Features	**6**

A4 Ice Sheet Model Characteristics

Table A5. Ice Sheet Model Characteristics for AIS, per Table 3 of (Seroussi et al., 2020). The three ISMIP6-wide parameters shared across all models in the experimental protocol are: Ocean Sensitivity (basal melt parameterization), Ice Shelf Fracture, and Ocean Forcing (open vs. standard framework). The remaining characteristics are those of individual ISMs.

Characteristic	Description	Possible Values
Numerics	Numerical solving method	Finite difference (FD), finite elements (FE), finite volumes (FV)
Stress Balance	Stress balance method	Includes Hybrid, HO, L1L2, Stokes, SIA+SSA
Resolution	Model resolution ~~in kilometers~~	Ice sheet model resolution in kilometers
Initialization	Initialization method	Spin-up (SP), spin-up with ice thickness target values (SP+), data assimilation (DA), data assimilation with relaxation (DA+), data assimilation of ice geometry only (DA*), and equilibrium state (Eq)
Initial Year	Initialization starting year	Year when initialization starts, varies by model (e.g., 1990, 2000)
Melt in Partially Floating Cells	Melt model for partially floating cells	Melt either applied or not applied over the entire cell based on floating condition (floating condition) and melt applied based on a sub-grid scheme (sub-grid) with N/A referring to models without partially floating cells
Ice Front Migration Model	Ice front migration scheme	Strain rate (StR), retreat only (RO), fixed front (fix), minimum thickness height (MH) and divergence and accumulated damage (Div)
Open Melt Parameterization	Basal melt rate in an open framework	Linear function of thermal forcing (lin), quadratic local function of thermal forcing (quad), PICO parameterization (PICO), PICOP parameterization (PICOP), plume model (Plume), and nonlocal parameterization with slope dependence of the melt (nonlocal + Slope)
Standard Melt Parameterization	Basal melt in standard framework	Local or nonlocal quadratic function of thermal forcing, and local or nonlocal anomalies (Local, Nonlocal, Local anom, Nonlocal anom)
Ocean Forcing	Framework for oceanic forcing	Open framework (open) where modelers choose their own interpretation, or a standard framework (standard) with a fixed set of equations
Ocean Sensitivity	Sensitivity of basal melt to ocean forcing	~~Includes High, Medium, Low,~~ Low, medium and high ocean sensitivity corresponding to the 5th, 50th and ~~Pine Island Glacier-Larsen (PIGL)~~ 95th percentile values of the ΓÇ£MeantAntΓÇ¥ γ_0 distribution (Jourdain et al., 2020)
Ice Shelf Fracture	Inclusion of ice shelf fracturing	True or False

Table A6. Ice Sheet Model Characteristics for GrIS, per Table A1 of (Goelzer et al., 2020). The three ISMIP6-wide parameters shared across all models in the experimental protocol are: Ocean Sensitivity (retreat parameter κ), Ice Shelf Fracture, and Ocean Forcing (open vs. standard framework). The remaining characteristics are those of individual ISMs.

Characteristic	Description	Possible Values
Numerics	Numerical solving method	Finite difference (FD), finite elements (FE), finite volumes with adaptive mesh refinement (FV)
Ice Flow	Ice flow model	Includes shallow-ice approximation (SIA), shallow shelf approximation (SSA), higher order (HO), and a combination of SIA and SSA (HYB)
Initialization	Initialization method	Includes data assimilation of velocity (DAv), data assimilation of surface elevation (DAs), data assimilation of ice thickness (DAi), spin-up (SP), transient glacial cycles (CYC), nudging to ice mask (NDm), and nudging to surface elevation (NDs)
Initial Year	Initialization starting year	Year when initialization starts, varies by model (e.g., 1990, 2000)
Initial SMB	Initial surface mass balance model	RACMO2.1 (RA1), RACMO2.3 (RA3), HIRHAM5 (HIR), MAR (MAR), BOX reconstruction (BOX), ISMB (implied SMB)
Velocity	Velocity data used to initialize the model	Includes Rignot and Mouginot (RM), Joughin et al. (J)
Bed	Bed topography data used to ~~initalize~~ initialize the model	Includes Morlighem et al. (M), Bamber et al. (G)
Surface/Thickness	Surface elevation and thickness data	Includes Morlighem et al. (M), Bamber et al. (G)
Geothermal Heat Flux	Geothermal heat flux model	Shapiro and Ritzwoller (SR), Greve (GR), and MIX for mixed models
Res. Min	Model resolution minimum	Minimum resolution, e.g., 4 km, 8 km
Res. Max	Model resolution maximum	Maximum resolution, e.g., 16 km, 32 km
Ocean Forcing	Framework for oceanic forcing	Open framework (open) where modelers choose their own interpretation, or a standard framework (standard) with a prescribed set of retreat and melt equations (Slater et al., 2020)
Ocean Sensitivity	Sensitivity of tidewater glacier retreat to ocean forcing (retreat parameter κ)	Median, 75th percentile, and 25th percentile of the sensitivity parameter κ as defined in Slater et al. (2020); Goelzer et al. (2020)

Appendix B: ISEFlow

B1 ISEFlow Architecture

Tables B1 and B2 detail the architecture for the Deep Ensemble for ISEFlow-AIS and ISEFlow-GrIS.

Table B1. ISEFlow-AIS Deep Ensemble Model Architecture.

LSTM Layers	LSTM Hidden Units	Criterion
1	128	HuberLoss
1	512	HuberLoss
1	512	HuberLoss
2	128	HuberLoss
1	256	L1Loss
1	512	MSELoss
2	128	MSELoss
2	512	MSELoss
1	256	L1Loss
1	64	HuberLoss

Table B2. ISEFlow-GrIS Deep Ensemble Model Architecture.

LSTM Layers	LSTM Hidden Units	Criterion
2	128	HuberLoss
2	256	MSELoss
2	128	HuberLoss
2	128	MSELoss
2	256	HuberLoss
1	256	L1Loss
1	128	HuberLoss
2	64	MSELoss
2	256	HuberLoss
1	256	L1Loss

B2 Emission Scenario Predictors

The emission scenario predictors are diagnostic tools trained separately from ISEFlow and are not components of the ISEFlow emulator. Their sole purpose is to identify which input variables most strongly encode emission scenario information, as described in Section 2.4. The emission scenario predictors were designed to classify climate forcings and ice sheet model characteristics into their respective emission scenarios (RCP2.6 or RCP8.5). These models were trained using the same input data as ISEFlow, enabling a direct evaluation of the features contributing to scenario distinction. The training workflow was implemented in PyTorch, with the model architecture being similar to a predictor model within ISEFlow.

The model architecture consisted of an input layer with dimension equal to the number of input features, including climate forcings (e.g., surface air temperature, SMB components, ocean salinity, and thermal forcing) and ISM characteristics (e.g., basal melt parameterization and initialization method). The architecture followed the ISEFlow architecture, but with a change to the output layer, which was a single neuron with a sigmoid activation function for binary classification between RCP2.6 and RCP8.5.

The training strategy employed the Binary Cross-Entropy Loss (BCELoss) as the loss function and the Adam optimizer with a learning rate of 1×10^{-3}. A batch size of 32 was used, and the model was trained for 100 epochs with early stopping based on validation loss. A cosine annealing learning rate schedule was applied to improve convergence. The model with the lowest validation loss was saved as the final checkpoint. All training was performed on an NVIDIA QuadroRTX GPU.

Performance Metrics

The performance of the scenario classifiers for the AIS and GrIS is summarized in Table B3. Both models achieved high accuracy and effectively distinguished between RCP2.6 and RCP8.5. The results indicate strong classification performance, with slightly higher accuracy for the GrIS, reflecting the clearer distinction in forcings between RCP scenarios for the Greenland region.

Table B3. Emission scenario predictor performance metrics for the AIS and GrIS.

Ice Sheet	Accuracy	Precision	Recall	F1-Score	Validation Loss
AIS	77.80%	0.90	0.62	0.73	0.46
GrIS	87.92%	0.94	0.82	0.88	0.27

Appendix C: Differences Between Our GP Emulator and the Original Emulandice

Throughout this paper, we refer to our Gaussian Process baseline as the "GP emulator" rather than "emulandice" to make clear that it is an approximation to the emulandice architecture of Edwards et al. (2021), not a reproduction of the original code or an identical reimplementation of all design choices. The following differences are important to keep in mind when interpreting comparisons between our GP emulator and the published emulandice results.

1. **Training data and GCM forcing.** The original emulandice was trained on CMIP5 and CMIP6 **global** mean surface air temperature (GSAT) simulations, using the ensemble mean of all available GCM realizations where multiple realizations existed. Our GP emulator is trained directly on the ISMIP6-processed regional CMIP forcings for the six GCMs used in ISMIP6, using individual GCM realizations rather than ensemble means.

2. **Additional training simulations.** The original emulandice included additional simulations specifically designed to improve performance and coverage of the GSAT input space, including SSP-based projections (**?**). Our GP emulator does not include these additional simulations and is trained only on the core ISMIP6 experimental ensemble.

3. **GP implementation and hyperparameter estimation.** Emulandice used the RobustGaSP package, which estimates GP hyperparameters via marginal posterior mode estimation (Gu et al., 2018). Our GP emulator uses the scikit-learn GP implementation (Pedregosa et al., 2011), which uses maximum likelihood estimation (MLE). Differences in hyperparameter estimation can lead to differences in fitted kernel parameters and resulting uncertainty quantification.

4. **Use of FaIR for prediction.** Emulandice was designed to generate projections by coupling to the FaIR simple climate model (Smith et al., 2018), which provides GSAT time series for a wide range of emissions scenarios. FaIR projections were used for *prediction*, not for *training*; training was performed on CMIP5/CMIP6 GSAT. Our GP emulator is both trained and evaluated using ISMIP6-processed CMIP forcings, without the use of FaIR.

5. **Categorical ISM parameters.** In our implementation of the GP emulator, we treat the input parameters for glacier retreat and sub-shelf basal melt as categorical rather than continuous variables. To our understanding, the continuous values for these parameters take on only a select number of values, so we treat them as discrete.

These differences mean that performance comparisons between our GP emulator and ISEFlow reflect our specific GP implementation. In particular, the 54.1% PICP for our GrIS GP emulator (Table 4), compared with 94.1% reported by Edwards et al. (2021) , is a consequence of the differences in training data and implementation listed above, and should not be interpreted as a reflection of the original emulandice's uncertainty quantification performance.

Appendix D: Emulator Validation vs ISMIP6 Simulations

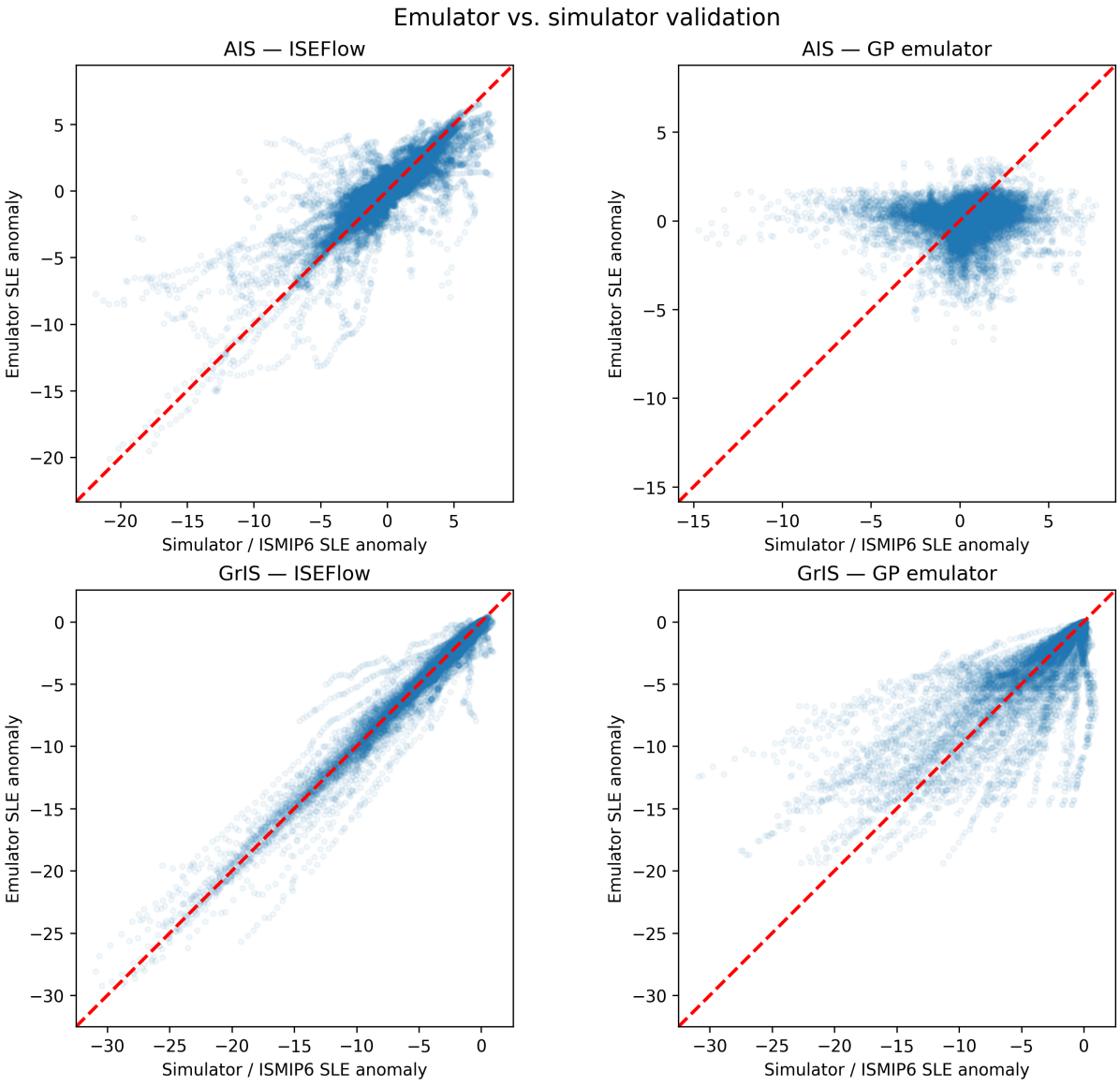

Figure D1. Emulator vs. simulator scatter plots for ISEFlow (left) and the GP emulator (right), for AIS (top) and GrIS (bottom). Each point is one emulator prediction plotted against the corresponding ISMIP6 simulator output (SLE anomaly, cm). The red dashed line is the 1:1 reference. ISEFlow predictions cluster tightly along the diagonal for both ice sheets, while the GP emulator shows substantially larger scatter, particularly for GrIS, consistent with the lower PICP and higher error metrics reported in Table 1. For the AIS, the GP tends to revert to the mean at near zero, resulting in a more horizontal trend.

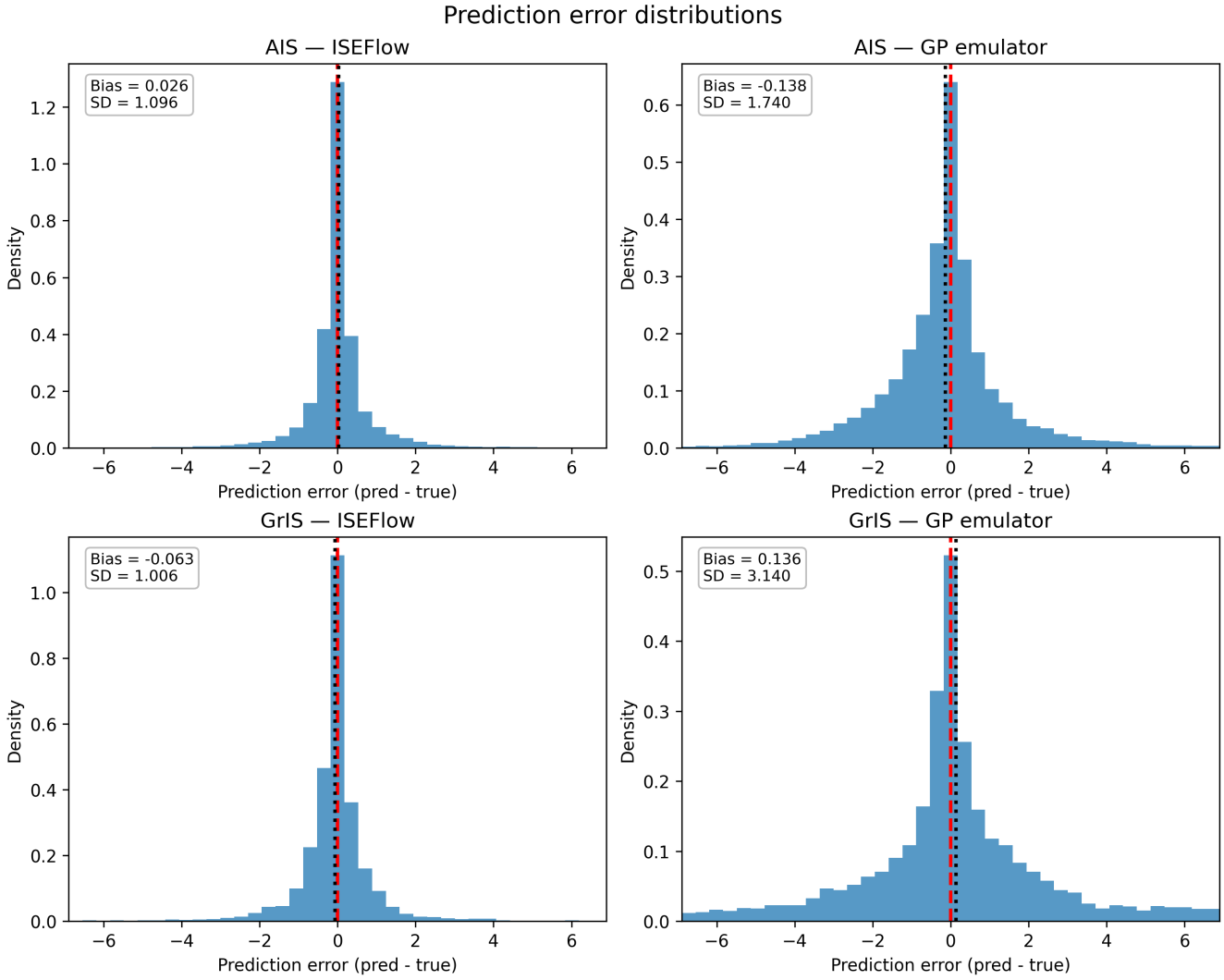

Figure D2. Prediction error distributions (emulated − simulated, cm SLE) for ISEFlow (left) and the GP emulator (right), for AIS (top) and GrIS (bottom). Bias and standard deviation of errors are shown in each panel. ISEFlow errors are tightly centred near zero (AIS: bias = 0.026, SD = 1.096; GrIS: bias = −0.063, SD = 1.006), while the GP emulator exhibits substantially heavier tails (AIS: SD = 1.740; GrIS: SD = 3.140). The larger GP emulator errors for GrIS, and the asymmetric tail visible in the AIS panel, are consistent with the implementation differences from the original emulandice described in Appendix C.