https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


# 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 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 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 and LARMIP, ISEFlow achieves substantially lower mean squared error and improved distribution approximation while maintaining faster inference times. This study investigates the drivers of increased 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 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 climate forcings and model characteristics that have the largest impact on sea level projections. ISEFlow offers a computationally efficient tool for generating accurate 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 uncertainty in future carbon emission scenarios (DeConto and Pollard, 2016; Golledge et al., 2015). To address these challenges, ice sheet models (ISMs), such as those that

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


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 simulations, emulators have emerged as a valuable tool for producing climate projections efficiently and accurately (Van Katwyk et al., 2023; Edwards et al., 2021). Emulators (also commonly referred to as "surrogate models") are small-scale models that approximate large, more complex physical models (or simulation models). Emulators differ from data-driven parameterizations that are designed to operate within a climate 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 system (e.g., Nicklas et al., 2025; Smith et al., 2018) or a major component of it (e.g., Lam et al., 2023b; 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 of machine learning (ML) architectures has significantly advanced emulator development by providing frameworks for accurately and efficiently capturing complex dynamics. Recent ML emulators like FourCastNet (Pathak et al., 2022) and GraphCast (Lam et al., 2023a) are able to produce high-resolution weather predictions with low computational demands. These models illustrate 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 (Weyn et al., 2021).

Emulators played a pivotal role in the ice sheet assessments of the Intergovernmental Panel on Climate Change (IPCC) 6th Assessment Report (AR6), where ISMIP6, an ensemble of ice sheet models that produce land ice and sea level projections (Nowicki et al., 2020; Seroussi et al., 2020; Goelzer et al., 2020; Payne et al., 2021), and the 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 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 emulator and the LARMIP linear response models have inherent limitations due to their simplified representations of ice sheet dynamics. Emulandice, a Gaussian Process (GP)-based emulator, relies on global mean surface air temperature and fixed parameters for glacier retreat and sub-shelf basal melt (Edwards et al., 2021). LARMIP employs linear response functions to estimate how different rates of ice-shelf basal melting on the Antarctic Ice Sheet (AIS) translate into

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


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 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). For example, Emulandice was unable to distinguish between different emission scenarios for the AIS in the IPCC AR6, effectively producing the same sea level outcomes regardless of scenario. This limitation is critical, as accurately modeling 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 of climate uncertainty quantification (Hawkins and Sutton, 2009; Lehner et al., 2020).

This study proposes a novel ML-based ice sheet emulator for both the AIS and GrIS that accurately and efficiently projects future sea level, quantifies projection uncertainties, and captures sea level projection sensitivity to carbon emission scenarios. 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 approximate ice sheet dynamics effectively due to incorporation of more input variables enabled by its computational efficiency. This work extends those previous results by proposing ISEFlow (v1.0), 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 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 carefully examines the underlying processes that drive ISEFlow's performance in order to verify the emulator is learning correct physical principles and to build confidence in its projections. We perform a series of experiments, including sensitivity testing and an analysis on input variable feature importances, to verify that the emulator is learning in a way that reflects the real-world behavior of ice sheet models. By understanding which variables and forcings most influence emulated sea level projections, we not only validate ISEFlow's accuracy and adherence to ice flow physics, but also identify the inputs necessary for capturing sensitivity to different emission scenarios. This ensures that ISEFlow provides reliable insights for future sea level projections and establishes a foundation for creating more effective ice sheet emulators. 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 and corresponding sea level projections driven by climate forcings from the Climate Model Intercomparison Project - phase 6 (CMIP6). The CMIP forcings include yearly-averaged atmospheric and oceanic forcing anomalies from 2015 to

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


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 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. 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. 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 A2, 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 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.

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. For the AIS, the training set consists of 635 projections of 86 years each, totaling 54,610 training observations, and the validation and testing set contains 136 full projections. 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 136 full 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 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

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


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 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 is sparse or unevenly distributed, meaning the NN model has less information about how the system behaves under those specific conditions. This reflects how well the training data represents the full range of possible input scenarios, with greater uncertainty in underrepresented or unexplored parts of the ISMIP6 data distribution. 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 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 sparsity 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 model—in this case, ice sheet models—the concept of "aleatoric uncertainty" does not apply in its traditional sense. Aleatoric uncertainty represents the inherent variability in observed data, but here, the "data" is generated by a model rather than from direct observational processes. Consequently, the uncertainty being quantified pertains not to natural variability but to the distribution and coverage of the ISM-generated data, as well as the variability of sea level projections coming from ISMs with the same configurations. Because of the limited number of simulation projections run, there are very few simulations with the same model configuration (same ISM characteristics), so the variability within projections with the same configuration will be much less prevalent than the variability due to differing configurations and the lack of simulations for all possible experiments. Therefore, we use the term "data coverage uncertainty" to reflect the uncertainty that arises due to an uneven sampling of the input space. This allows for a clear separation between the uncertainty intrinsic to the emulator and the uncertainty arising from sparsity or unevenness in the training data.

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 measuring the density of data sampling across different regions of the input space, meaning that

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


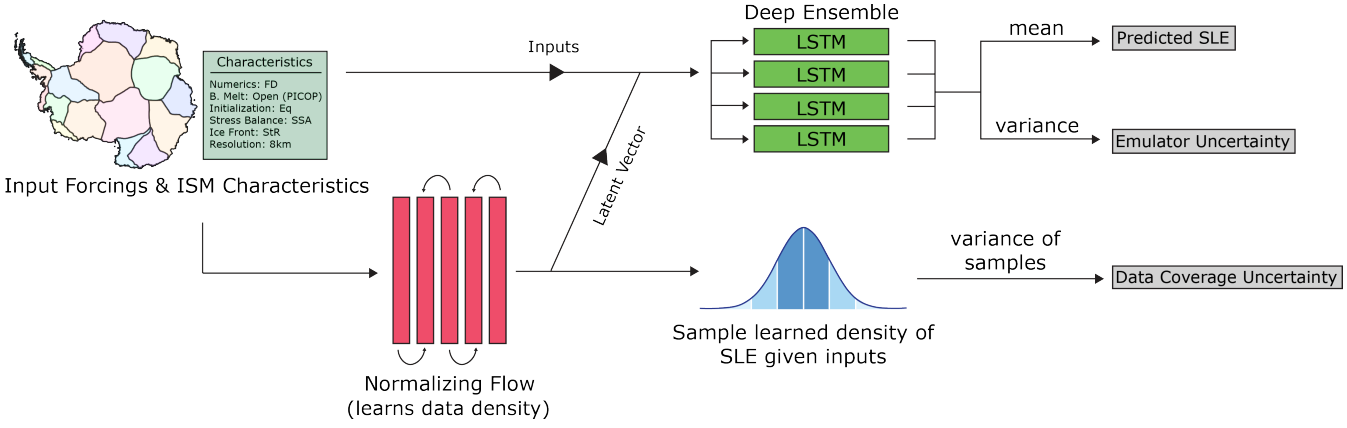


**Figure 1.** Architecture diagram for ISEFlow-AIS and ISEFlow-GrIS. The input data consists of climate forcings and ice sheet model characteristics. The input data is first passed through a Normalizing Flow to approximate the data density. The resulting latent transformation is concatenated with the 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.

in areas where the data is highly variable, 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 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 predictions 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.

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

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


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 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. We run the LARMIP linear response functions using ISMIP6 AIS global surface air temperature (GSAT) as inputs with the generated 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, the IPCC AR6 compares LARMIP alongside the Emulandice emulator (Figure 9.18, IPCC AR6 (Fox-Kemper et al., 2021)), so we follow that precedent here.

We implement a Gaussian Process-based emulator that closely follows the Emulandice architecture presented in Edwards et al. (2021). We use a Gaussian Process with a power exponential kernel with an optimal alpha (power) parameter of 0.1 with an unbounded nugget kernel to model random variability induced by variables not present in the dataset. 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 is trained on global surface air temperature, as well as a parameterization of glacier retreat for the GrIS and a sub-shelf basal melt and an 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).

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. All training and testing procedures were carried out on a 256 GB compute node with training and inference times determined by logging wall times.

### 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'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

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


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 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), 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 emulator behavior and performance.

To better understand how ISEFlow-AIS and ISEFlow-GrIS are able to capture sensitivity to carbon 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 ability to accurately predict sea level, but also provides insights for capturing the uncertainty associated with future carbon 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 work validates the emulator based on test data that was help 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 projection accuracy, projection distribution similarity, ability to distinguish between carbon emission scenarios, and training and inference time. As shown in Table 1 and consistent with previous results (Van Katwyk et al., 2023), ISEFlow demonstrates superior performance compared to Emulandice 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. ISEFlow-AIS achieves an MSE of 1.20, compared to 3.03 for

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


**Table 1.** Performance metrics for ISEFlow (v1.0), Emulandice (Edwards et al., 2021), and LARMIP (Levermann et al., 2020) for the Antarctic Ice Sheet (AIS) and Greenland Ice Sheet (GrIS). Note that Emulandice refers to the Emulandice architecture trained with the same inputs as ISEFlow, rather than the original published model, which used a different training dataset and climate forcings.

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

* KLD and JSD are evaluated at projection year 2100.

Emulandice and 4.23 for LARMIP. Similar improvements are observed for the GrIS, where the MSE for ISEFlow-GrIS is 1.02, significantly lower than Emulandice's 10.14.

In addition to projection 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 emulator's 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. 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 better at capturing the full distribution of sea level rise projections—a critical aspect of distinguishing sensitivity to emissions scenario.

Emulandice trains faster than ISEFlow, primarily because it emulates fewer regions, which results in less training data. 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 time for ISEFlow is faster than for Emulandice, making ISEFlow'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 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 im-

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


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

provement in both the projection 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 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 accuracy but also better improves the ability to emulate the distribution of possible sea levels.

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 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 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

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


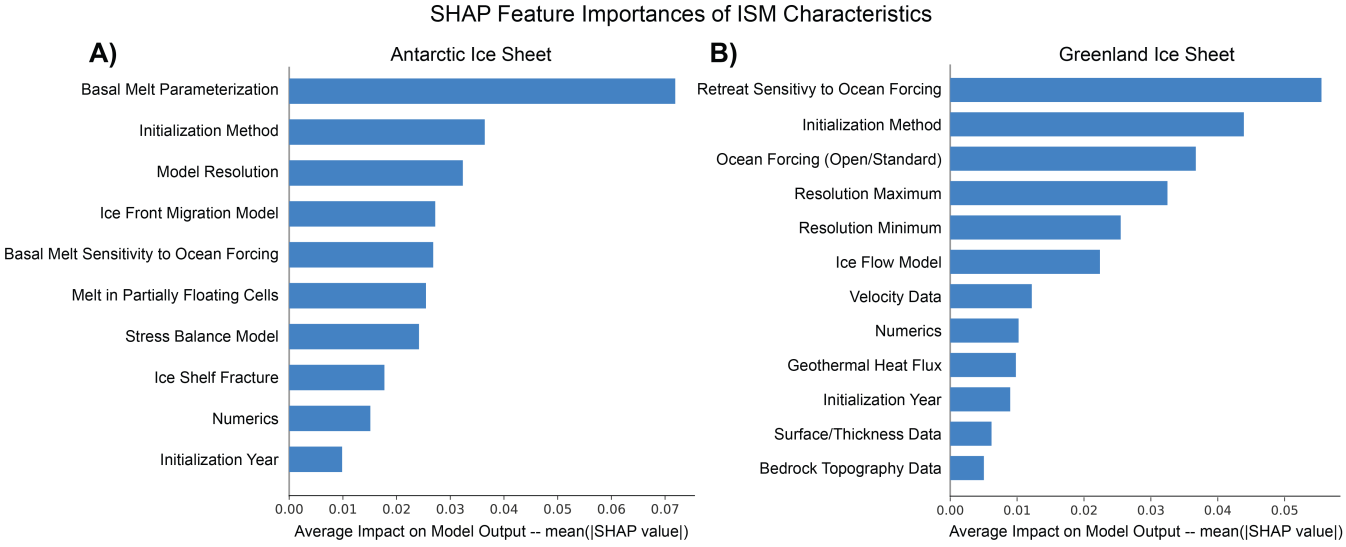


**Figure 2.** SHAP values representing the relative importance of different ISM characteristics for the Antarctic Ice Sheet (AIS, a) and Greenland Ice Sheet (GrIS, b) on ISEFlow projections.

runoff, and is correlated with surface air temperature) offers more information for modeling SLE. Table 3 shows that across both ice sheets, 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 accuracy. For the AIS, the NN trained on all forcings achieves the best accuracy, with an MSE of 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 (MSE 2.18). In contrast, training the NN on temperature alone results in weaker performance, with an MSE of 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 2.60).

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


**Table 3.** Comparison of NN and GP emulators trained on input forcings only (no ISM characteristics), 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 in both projection accuracy and scenario sensitivity for both the AIS and GrIS, and adding more forcings improves performance further.

| 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

For the GrIS, the trend in predictive accuracy is similar to the AIS case, with the NN trained on all forcings achieving the best results, posting an MSE of 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 (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 distinguishing scenarios in some cases or when additional forcings are unavailable, is insufficient for producing accurate future sea level projections.

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

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


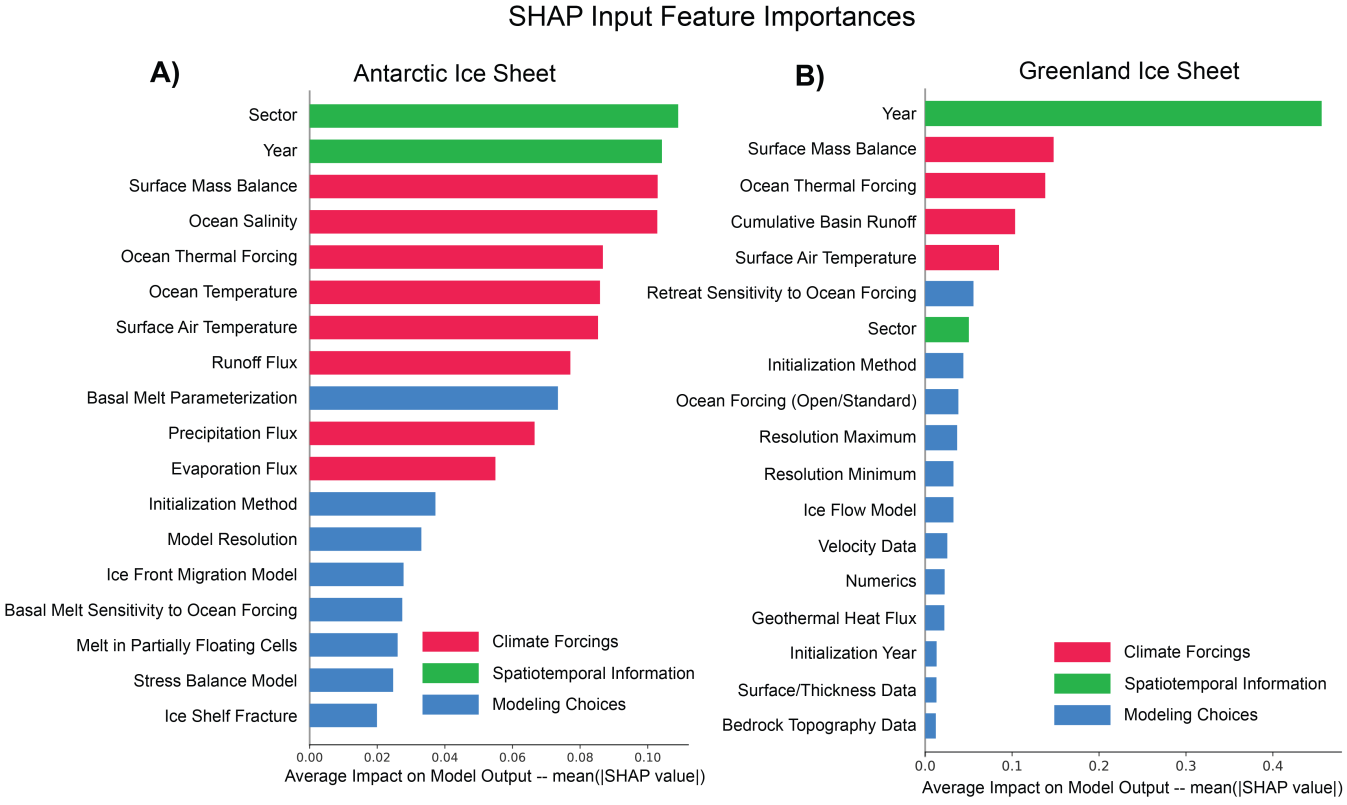


**Figure 3.** SHAP values for ISEFlow input variables for both the AIS (a) and GrIS (b). Bar color indicates the feature group, consisting of spatiotemporal features, climate forcings, and ISM characteristics.

feature groups, ISM characteristics were the least impactful on emulator projections, but still contribute enough information to increase emulator accuracy significantly (Table 2).

### 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. 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. The KS test results (Table 4) further confirm this, 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. 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

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


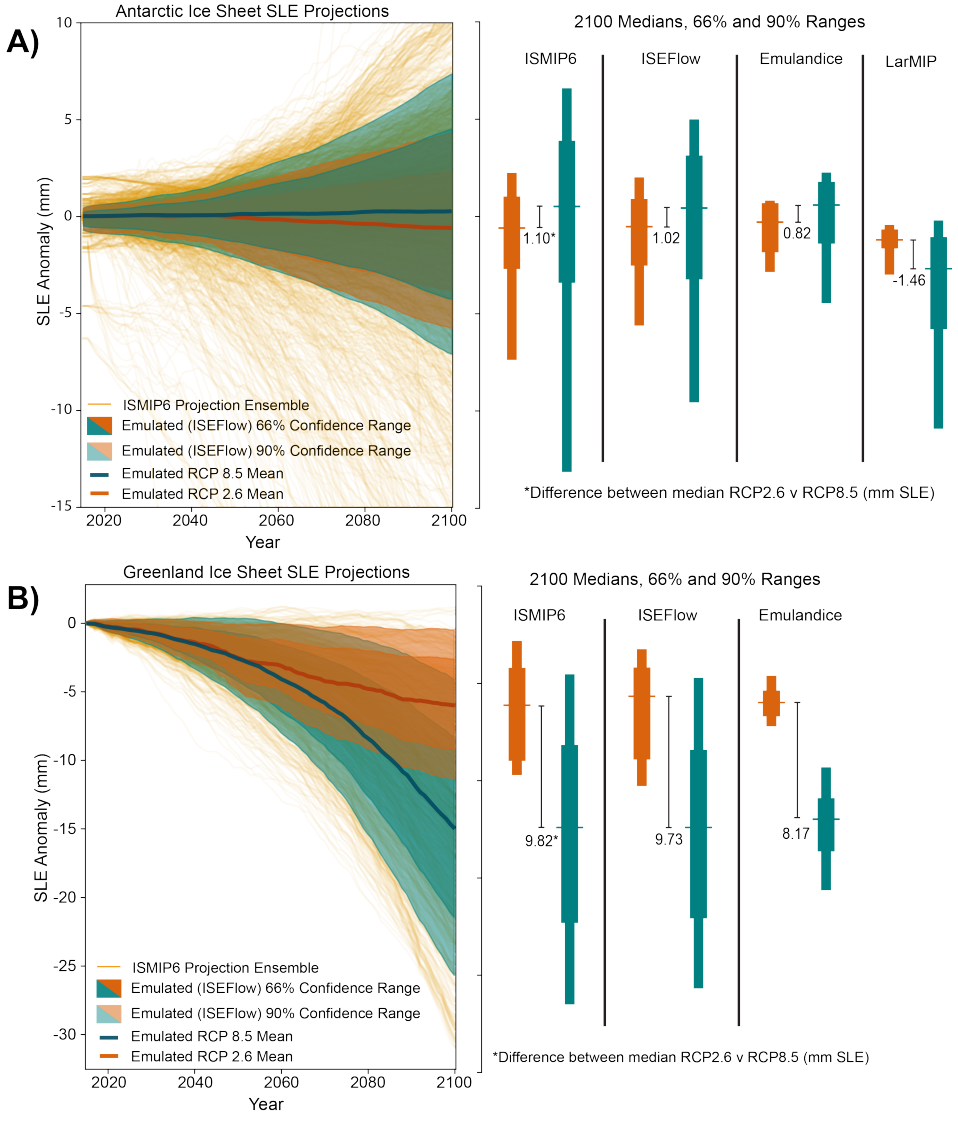


**Figure 4.** Plot based on IPCC Figures 9.17 and 9.18 (Fox-Kemper et al., 2021) showing the projections of the AIS (a) and GrIS (b) along with ISMIP6, ISEFlow, Emulandice, and LARMIP median predictions at the year 2100 grouped by emissions scenario. ISEFlow accurately approximates the spread of each scenario distribution and identifies the difference in median SLE projections between emission scenarios. Emulandice is unable to distinguish between emission scenarios for the AIS and fails to approximate the ensemble spread. 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.

and RCP 8.5 (Figure 4). Note that a reason for Emulandice's strong distinction between scenarios is its underestimation of the

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


**Table 4.** 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 |
|---|---|---|---|---|
| AIS | ISEFlow | **1.20** | 0.001 | 0.16 |
| | Emulandice | 3.03 | 0.026 | 0.12 |
| | LARMIP-2 | 4.23 | 0.683 | **0.31** |
| GrIS | ISEFlow | **1.23** | 0.007 | 0.15 |
| | Emulandice | 10.14 | 0.060 | **0.26** |

spread of the ISMIP6 ensemble which separates the distributions of GrIS outcomes excessively. These findings are consistent with the IPCC AR6 results, where Emulandice was able to distinguish GrIS scenarios more effectively than AIS scenarios. 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 key nonlinear self-dampening and amplifying processes (Levermann et al., 2020), which can lead to substantial inaccuracies. While LARMIP achieves some differentiation across scenarios, its inability to capture nonlinear ice sheet responses results in notable projection errors, reducing its reliability as an emulator. However, LARMIP's computational efficiency and ability to isolate ocean dynamics 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.

To identify the drivers of ISEFlow's ability to distinguish between emission scenarios, we run SHAP on the classification model mentioned in Section 2.4 (Figure 5). 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 set of climate forcings is essential to improve the model's sensitivity to different emission scenarios.

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


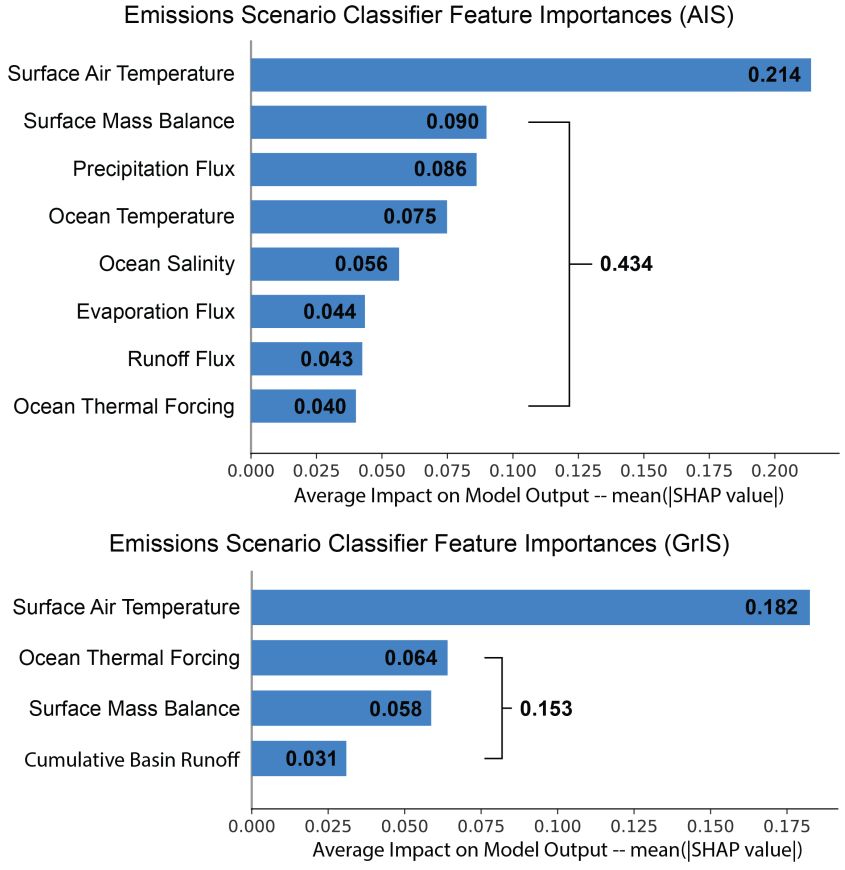


**Figure 5.** SHAP values showing the importance of various features in distinguishing between carbon emission scenarios for the AIS (top) and GrIS (bottom). Surface air temperature is the most influential forcing for both ice sheets. However, for the AIS, the combined importance of the remaining forcings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

### 3.4 Uncertainty Quantification Comparison of Emulators

To directly compare the quantified uncertainty between ISEFlow and Emulandice, we trained a separate model (Regional ISEFlow) with the same architecture as ISEFlow using the same dataset and input variables as Emulandice, including aggregated regional data and limiting forcings to surface air temperature, glacier retreat, sub-shelf basal melt, and ice shelf fracture (Table 5). This approach ensures a fair 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 ISEFlow emulator with the Emulandice to compare 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.

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


**Table 5.** Comparison of emulator uncertainty quantification performance. A separate ISEFlow emulator was trained on the same data as Emulandice (regional projections instead of sector projections, surface temperature forcings, etc.) to provide a fair comparison to the Emulandice emulator. Metrics include the Continuous Ranked Probability Score (CRPS), Prediction Interval Coverage Probability (PICP), Mean Prediction Interval Width (MPIW), and the Winkler Score. Between the regional ISEFlow emulator and Emulandice, ISEFlow is more accurate (higher PICP), but has wider intervals (higher MPIW). The Winkler score indicates that the regional ISEFlow models do better at balancing the width and accuracy of intervals. The full ISEFlow emulators, ISEFlow-AIS and ISEFlow-GrIS are consistently most accurate (lowest PICP) and approximate the probability distributions best (highest CRPS), showcasing 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 | 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 | 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.

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 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.

When comparing the Regional ISEFlow and Emulandice (Table 5), Regional ISEFlow performs better on PICP, indicating more reliable coverage of the true outcomes. However, Emulandice achieves narrower intervals (lower MPIW), but at the expense of significantly worse PICP. This suggests that Emulandice tends to underestimate uncertainty, leading to intervals that fail to capture the true values consistently. The CRPS values are very close between the two models, but Regional ISEFlow still slightly outperforms Emulandice.

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


When comparing the ISEFlow-AIS and ISEFlow-GrIS to both Regional ISEFlow and Emulandice, ISEFlow-AIS and -GrIS show significant improvements across nearly all metrics. It achieves the lowest CRPS and Winkler scores, indicating more accurate and sharper 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, 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 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 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 inclusion of ISM characteristics as input features, as well as the incorporation of all available ISMIP6 climate forcings.

A primary goal of this study is to determine the ability of ISEFlow to capture the correct underlying processes, and whether the emulator'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.

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


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 scenarios. ISEFlow could be employed to emulate sea level projections for intermediate Shared Socioeconomic Pathways (SSPs), 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. 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 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 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 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 the ice sheet model ensemble simulation process.

The findings of this study underscore the importance of incorporating a diverse set of input forcings and model characteristics 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

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


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 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 modeling.

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 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.

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. Furthermore, 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, external forcing uncertainties, or potential feedbacks between processes not captured by the dataset. Addressing these limitations in future work, 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 projections of future sea levels. We compare ISEFlow to state-of-the-art emulators and show that ISEFlow creates more accurate projections, is faster to run, and is able to capture climate sensitivity to carbon 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's predictions but also offers valuable insights into ice sheet model characteristics and their impact on sea level projections. Another significant contributions 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's ability to quantify both data coverage and emulator uncertainty

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


separately ensures that areas of sparse data are properly accounted for, reducing the risk of overconfidence in projections where data is limited. 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 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 to future climate scenarios. By applying ISEFlow to a broader set of forcings, we may uncover new insights into the potential differences between emission scenarios that were previously underrepresented. 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 forcings contribute most to quantified uncertainty.

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'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 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.

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


*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 – 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 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 feedback on the manuscript. This is ISMIP6 contribution number 34.

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


## 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/10.1145/3292500.333070, 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.

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.

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


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.

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.

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


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, 2023a.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023b.

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.

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

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


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/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.

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.

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

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


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.

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


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.

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.

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


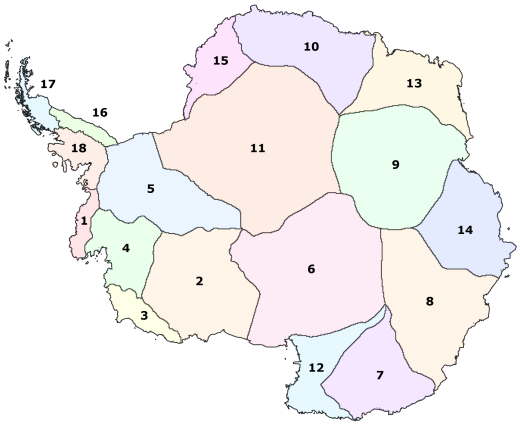


**Figure A1.** Map of Antarctic Ice Sheet sectors as used in ISMIP6 AIS (Seroussi et al., 2020).

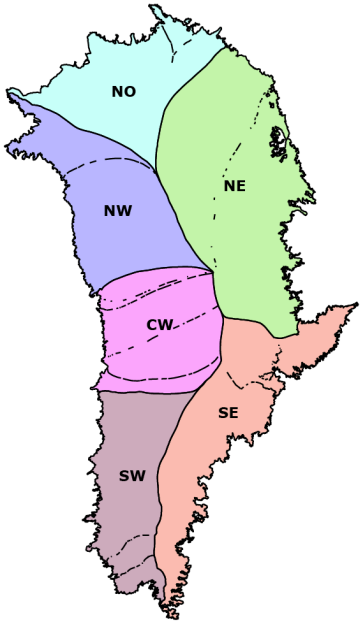


**Figure A2.** Map of Greenland Ice Sheet sectors as used in ISMIP6 GrIS (Goelzer et al., 2020)

## 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.

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


**Table A1.** Ice Sheet Model Characteristics for AIS, per Table 3 of (Seroussi et al., 2020).

| 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, and Pine Island Glacier-Larsen (PIGL) |
| Ice Shelf Fracture | Inclusion of ice shelf fracturing | True or False |

## A2 Ice Sheet Model Characteristics

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


**Table A2.** Ice Sheet Model Characteristics for GrIS, per Table A1 of (Goelzer et al., 2020).

| 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 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 |

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


**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 |

## Appendix B: ISEFlow

### B1 ISEFlow Architecture

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

https://doi.org/10.5194/egusphere-2025-4914
Preprint. Discussion started: 1 December 2025


### B2 Emission Scenario Predictors

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 $1 \times 10^{-3}$. 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 |