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

Peter Van Katwyk, Baylor Fox-Kemper, Sophie Nowicki, Hélène Seroussi, and Karianne J. Bergen

We are genuinely thankful to all three reviewers, Chris Smith (#1), Denis Felikson (#2), and Tamsin Edwards (#3), for the depth and care of their reviews of this manuscript. The reviews have pushed us to substantially rethink several aspects of the manuscript, and the paper will be considerably stronger for it. The most significant changes across the revision include a thorough revision of language throughout the manuscript to accurately frame ISEFlow's contributions as improvements in emulating the ISMIP6 ensemble rather than claims about real-world predictive accuracy, a more careful and honest treatment of the GP baseline, including clear documentation of the ways our implementation differs from the original Emulandice, and a renaming to "GP emulator" throughout to reflect this, and a substantive reworking of the Discussion to acknowledge the trade-offs involved in different emulator design choices, the limitations of the ISMIP6 ensemble itself as a basis for policy-relevant projections, and the additional steps that would be required to produce projections suitable for a context like the IPCC.

In this author reply, we have carefully addressed each comment and suggestion. Reviewer comments are shown in blue, and our responses are provided below. Line numbers refer to the original submitted manuscript.

Response to Reviewer #2

1.) Additional discussion is needed on the topic of distinguishing between scenarios.
a.) The introduction needs additional text to clearly define the metrics that will be used to judge whether ISEFlow performs better in terms of distinguishing between scenarios.
We agree that the introduction should better prepare the reader for the metrics used to evaluate scenario distinction. In the revised manuscript, we will add text to the introduction defining the KS D statistic and explaining how it is used to evaluate the separation of emulated distributions under different emission scenarios.

b.) The text in Section 3.3 uses the KS statistic and claims that ISEFlow is able to distinguish between scenarios better than emulandice. However, Table 4 shows that emulandice has larger KS values than ISEFlow for both ice sheets. My understanding is that a larger value for emulandice indicates that it produces distributions that are further apart than for ISEFlow. This seems to contradict the claim that ISEFlow is better able to distinguish between scenarios and this must be addressed.

We appreciate the reviewer drawing attention to this. For the AIS, ISEFlow has a KS D of 0.16 compared to the GP's 0.12, indicating that ISEFlow does produce greater separation between the RCP 2.6 and RCP 8.5 distributions for Antarctica (higher D statistic → greater distinction between distributions). For the GrIS, the GP does show a larger KS D (0.26 vs. 0.15). However, distinguishing between scenarios for the GrIS has historically been less problematic; even in the IPCC AR6, the Emulandice-based projections were able to separate GrIS scenarios (Fox-Kemper et al., 2021). The primary concern motivating this analysis was the AIS, where previous emulators produced nearly indistinguishable scenario distributions. We acknowledge that our framing overstates the case for ISEFlow's superiority across the board, and we will revise the text to present the KS D results more precisely, making clear that the improvement is most notable for the AIS.

c.) I also wonder about whether this lack of ability to distinguish between scenarios is actually being caused by the underlying ISMIP6 ensemble and not by the emulators themselves. I suggest adding text on this in the Results and Discussion sections and clearly stating whether the ISMIP6 ensemble can distinguish between scenarios, maybe by showing the KS values for that ensemble (if it is valid to use that statistic for that ensemble).

This is a great point, and we thank the reviewer for bringing this up. We will add the metrics for the distinction between scenarios for ISMIP as well, as the emulators are not expected to be able to distinguish better than the original ISMIP6 ensemble. We will also add text to the Results and Discussion addressing this.

d.) Please also add some discussion on whether the differences in the KS D statistics in Table 3 are statistically significant and what values of KS indicate a "strong" ability to distinguish between distributions (as stated in Section 3.3). Qualitatively, it's not immediately obvious from the box-and-whisker plots in Fig 4a that ISEFlow can distinguish between the two scenarios and it's not clear whether a KS value of 0.16 indicates that these two distributions strongly differ.

Thank you for this comment. We agree that more context is needed on both the statistical significance of the KS D differences between emulators and the interpretation of the absolute D values. To address the former, we will compute bootstrap confidence intervals on each D statistic by resampling the RCP 2.6 and RCP 8.5 predictions independently, and will report these alongside the D statistics in the appendix. Regarding the interpretation of absolute values, there is no universally accepted threshold for what constitutes a "strong" KS D statistic, as its practical significance depends on sample size and context. We will accordingly remove the

characterization of ISEFlow's D statistics as indicating a "strong" ability to distinguish between scenarios, replacing this with more measured language that notes the relative differences between emulators while acknowledging that the absolute magnitudes are modest. We note that the qualitative picture in Figure 4a should also be interpreted in light of the reviewer's comment 1c: differences in the KS D between emulators partly reflect the uneven sub-ensemble designs of ISMIP6, not scenario signal alone, and this caveat will be incorporated into the revised text.

2.) Throughout the Results section, there are results reported for the NN emulator and the GP emulator (e.g., line 299) and these are seemingly referring to different architectures of ISEFlow. Although the Methods section (paragraph beginning on line 170) discusses different ISEFlow architectures, it doesn't describe a "GP" architecture for ISEFlow. I suggest adding text to the Methods section that provides more detail about the GP architecture and describes the differences between the NN architecture and the GP architecture for ISEFlow.
The text will be updated to clarify these distinctions. The multiple realizations of ISEFlow and the GP-emulator focus on differing inputs rather than architectures for the most part. For example, Table 3 shows the same ISEFlow architecture for the “NN” architecture and the “GP” architecture (essentially the Emulandice architecture) except with different forcing inputs. We will, however, add much more text surrounding differences, especially around the GP (Emulandice), as our implementation of Emulandice isn’t strictly the same, so we will clarify (see Reviewer 3 responses).

3.) The text in Section 3.2 reports the SHAP analysis for a "classification model" (line 226). It is not clear how these results demonstrate that ISEFlow can distinguish between scenarios, given that this is a completely different set of inputs, outputs, and a retrained model. Text should be added (either in Methods or elsewhere) to address this.
The reviewer is correct that the classification model is a separate model from the ISEFlow emulator, and this distinction was not made sufficiently clear. The purpose of the classification SHAP analysis is not to demonstrate that ISEFlow itself can distinguish between scenarios, but rather to be a diagnostic tool to help identify which input features are most informative for separating emission scenarios in the ISMIP6 data. This provides insight into why including a broader set of forcings improves scenario distinction in any emulator. We will clarify this distinction in both the Methods and Results sections.

b.) Given (a), the utility of these results is not clear. I am perhaps oversimplifying but these results are showing that, for example, surface air temperature over the GrIS and the AIS is a strong predictor of scenario. In other words, the emulator is able to correctly identify the scenario given the temperature. This is a valid result but what does this information tell us about the climate models, ice sheet models, or the ISEFlow emulator? There should be text added to the Discussion section to address this.

The classification SHAP analysis serves to identify which forcing variables carry the most scenario-distinguishing information. The key insight is that while temperature is the single strongest predictor, the combined contribution of all other forcings is comparable in magnitude (for the AIS, the average SHAP value of non-temperature features combined is 0.434 vs. 0.214 for temperature alone). This supports the design choice of including a broader set of climate forcings in the emulator, rather than relying on temperature alone. We will add discussion text clarifying this interpretation and connecting it to the implications for emulator design.

4.) Throughout the manuscript, "emulator uncertainty" is the term used to refer to the uncertainty on the output of the emulator. Does this capture both the uncertainty introduced by the emulation process and the uncertainty within the underlying input data (i.e., the ISMIP6 ensemble itself)? If so, I suggest explicitly stating this somewhere (possibly in the paragraph on lines 129–139).
We will add clarifying text in the paragraph on lines 129-139 to explicitly define the relationship between these sources of uncertainty. In our framework, “emulator uncertainty” refers to the uncertainty introduced by the emulation process itself (captured by the deep ensemble variance), while “data coverage uncertainty” reflects the uncertainty in the training data itself, (captured by the normalizing flow). The spread of the ISMIP6 ensemble is inherently reflected in the training data, and is therefore captured in data coverage uncertainty. We will also add text addressing how these two components relate to the total uncertainty and whether they should be combined for any given application.

5.) The use of LARMIP in the paper is a bit unclear to me and it stems from the "training time" reported in Table 1. Am I interpreting correctly that it takes 20 minutes to generate 20,000 random samples using the LARMIP process (but with ISMIP6 AIS GSAT as inputs instead of what the LARMIP paper used)? Or was there something else introduced as part of the process that caused it to take a long time? I think that I'm just surprised that generating 20,000 samples takes so long but I've also not thought too deeply about the computations involved in LARMIP. I just wanted to check that I'm understanding this part of the paper correctly. If my interpretation is incorrect, please add a clarification to the text to explain how the LARMIP emulator was used and how it differs from what was done in the original LARMIP paper.
The reviewer’s interpretation is essentially correct. The 20-minute training time reflects the generation of the 20,000 basal melt forcing time series ensemble as specified in Levermann et al. (2020), using ISMIP6 AIS GSAT as input. We will clarify this in the revised text to make the LARMIP procedure and its computational cost more transparent.

6.) Text should be added to the introduction to the paragraph on lines 75–84 to specify how this study will determine whether the emulator is "learning correct physical principles". The Results section uses a SHAP analysis of the emulator inputs and compares this against previous studies that have used other methods to partition uncertainty. This should be explained in the introduction for added clarity.

We will expand the introduction to specify that the SHAP analysis of input variable importance is the primary method used to assess whether ISEFlow's learned relationships are consistent with the physical understanding of ice sheet dynamics established in the literature.

7.) The input variables used for training ISEFlow (for all of the different trainings done in this paper) need to be more clear. This can be done in the main text or supplement.
a.) Provide a table that lists all input variables used to train the emulator that produced the results in Table 1.
b.) Provide a table that lists the input variables used for the "All", "Temperature", and "SMB" versions of the emulator shown in Table 3.
We will add comprehensive tables to the supplement listing all input variables used for each model configuration: the full ISEFlow model (Table 1 results), the sensitivity test configurations (Table 3: All, Temperature, SMB), and the Regional ISEFlow model used for comparison with the GP (Table 5).

Minor Comments

Line 21: Here, you cite DeConto and Pollard (2016) in the explanation of sources of uncertainty for ice sheet projections. I suggest expanding on this just a bit to make it clear that one major source of uncertainty is MICI, which reflects "deep uncertainty," and this is separate from the other sources of uncertainty.
We agree and will add text distinguishing the deep uncertainty associated with marine ice cliff instability (MICI) from other sources of uncertainty, such as climate forcings, ice sheet model parameters, and initial conditions. We will also note the Structured Expert Judgement (mentioned below) process used in AR6 to address these deeply uncertain contributions, as this provides important context.

Line 24: Change "climate simulations" to "ice sheet simulations". Line 25: Change "climate projections" to "ice sheet projections". Line 26: Change "small-scale" and "large" to "computationally efficient" and "computationally expensive". Line 28: Change "climate model" to "physical model". Line 29: Change "climate system" to "physical system".
We agree with all of these suggested changes and will implement them in the revised manuscript. The revised language better reflects the scope being discussed and avoids ambiguity.

Lines 46–51: In addition to ISMIP6 and LARMIP, AR6 used the results from the Structured Expert Judgement (SEJ) process to produce the high-end, deeply uncertain projections for the ice sheets. You can call that out here.
We will mention the SEJ process alongside ISMIP6 and LARMIP in this paragraph for completeness.

Line 55: Specify that these fixed parameters resulted from choices made in the ensemble design of ISMIP6 and not limitations imposed by Emulandice itself.
This is an important distinction. We will revise to clarify that the fixed parameters reflect the ISMIP6 experimental design rather than limitations of the Emulandice framework.

Lines 114–120: This paragraph could be clarified by stating the total number of projections available for each ice sheet from ISMIP6. Additionally, please clarify what is meant by the word "full" when stating "136 full projections."
We will state the total number of projections available for each ice sheet and clarify that “full” refers to 86-year projections.

Lines 129–139: To obtain the "true" uncertainty, should the "data coverage uncertainty" and "emulator uncertainty" be combined? I suggest adding some text on whether this is the case.
Yes, for a complete picture of total uncertainty, both components should be considered. For this reason, when assessing ISEFlow’s ability to quantify uncertainty, we use the total uncertainty (Table 5). We will add text explaining how these two sources relate and how they might be combined for downstream applications.

Line 158: Should this read: "areas where the data is highly variable or missing"?
Corrected. We will fix this in the revised manuscript.

Line 185: Specify what additional data is needed from ISMIP6 and how it should be combined with the output from the LARMIP function to produce SLE values.
We will add a description of the additional ISMIP6 data required and how it is combined with the LARMIP output.

Line 233: There's a typo here: "that was help out of the original ISMIP6 ensemble". Please rephrase.
This should read “that was held out of the original ISMIP6 ensemble.” We will correct this.

Line 237: Please clarify what an "individual projection" is.
An individual projection refers to a single ISMIP6 experiment, i.e., the output of one ice sheet model driven by one set of climate forcings under one emission scenario with one model configuration for a single 86-year span. We will define this explicitly in the text.

Line 240: How are MSE and MAE calculated? Is the error calculated as the difference between (1) emulated and (2) mean ISMIP6 SLE at each time step? Or just at 2100? Please explicitly state how these metrics are calculated.
MSE and MAE are calculated as the average error across all time steps (2015–2100) for each projection, and then averaged across all projections in the test set. The error at each time step is

the difference between the emulated SLE and the ISMIP6 SLE for that individual projection. We will state this explicitly in the Methods.

Table 1: Add units to all reported values in the table, either within the table or in the header.
We will add appropriate units (mm² SLE for MSE, mm SLE for MAE, dimensionless for KLD and JSD, minutes and seconds for time) to all table entries.

Line 283: I suggest adding "the emulator would need to be re-trained and" before "the rankings are likely to differ." Line 284: I suggest replacing the word "additional" with "individual climate". Line 289: Would it be a bit more specific to replace "offers more information for modeling SLE" with "is a better predictor of SLE"?
We will incorporate these changes in the revised version.

Lines 295–316: I suggest moving and combining the text in these paragraphs with what's in Section 3.3.
We agree that consolidating this material with Section 3.3 would improve the flow of the manuscript and will restructure accordingly.

Line 315: The phrase "is insufficient for producing accurate future sea level projections" should be changed to something like "results in less accurate sea level projections."
Agreed. The original phrasing is too strong given that the temperature-only configuration still produces informative results. We will revise as suggested.

Line 317: Add text to clarify that the results shown in Figure 3 are for the "All" set of training inputs.
We will add this clarification.

Figure 4: It would be useful to show the RCP 2.6 and RCP 8.5 ISMIP6 projections in two different colors.
We will color the ISMIP6 projections by scenario using the same colour scheme as the emulated distributions, while keeping thinner lines to distinguish them from the emulated results.

Line 390: Change "widely used emulators" to "previous emulators such as emulandice".
Line 394: I suggest changing the phrase "capture the correct underlying process" to something like "capture the physical processes that have been shown by previous studies to be the dominant drivers of future mass change".
We will revise as suggested.

Line 412: In this paragraph, you should mention that emulandice was also previously used to fill in the different SSPs that weren't modeled by ISMIP6.

We will add this context.

Line 434: This sentence states that ISEFlow can be a "valuable as a tool in designing ISMIP7 configurations." Please add a sentence or two to describe how the outputs of ISEFlow could be used in ISMIP7 design.
We will expand this to explain that ISEFlow can be used to predict how different ISMIP7 experimental configurations (e.g., model choices, forcing selections, parameterization decisions) would affect the resulting sea level projections, thereby helping the community prioritize simulations that help better understand or reduce uncertainty.

Lines 438–440: The text states that ISEFlow can be used to predict ice sheet model output for any CMIP7 forcing. This reads like it is restating what was written in the paragraph on lines 412–420.
We will revise this paragraph to avoid redundancy and more clearly distinguish between the two capabilities: (1) filling in SSPs not simulated in ISMIP6, and (2) evaluating new CMIP7 forcing products for inclusion in the ISMIP7 experimental design.

Line 477: Change "contributions" to "contribution".
Corrected.

We once again thank all three reviewers for their careful and constructive comments. We are confident that addressing these points will result in a substantially improved manuscript, and we look forward to submitting the revised version.