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 #1

From what I can tell, ISEFlow can emulate ISMIP models well, which I believe were from a small number of SSP/RCP projections. What would be very useful indeed would be the ability to produce projections for emissions scenarios not run in CMIP/ISMIP models. Is this possible at the moment in ISEFlow?

This is possible in principle, and we will clarify this point in the text. However, it requires compatible climate forcing inputs. Because ISEFlow uses regional atmospheric and oceanic forcings rather than GMST alone, extending it to scenarios outside the original ISMIP6 ensemble requires corresponding forcing data that must be provided in the same format as the training inputs. We will clarify this distinction in the manuscript and note it as an important direction for

future work. This kind of work is a natural next step, we are working on running an additional 24 CMIP models through the ISEFlow-AIS emulator to produce sea level projections for forcing scenarios not included in the original ISMIP6 ensemble. We will add a brief note to the Discussion clarifying this capability and talk about this as a future work.

Figure 4: I think the SLR anomaly axis has the wrong sign, by tracing the origin of this figure back to IPCC (we have mass loss, so SLR anomaly should increase). I also think it would be nice to be consistent with the majority of literature and switch the colours of RCP2.6 and RCP8.5.
We thank the reviewer for this observation. Following the ISMIP6 convention (e.g., Seroussi et al., 2020; Goelzer et al., 2020), sea level equivalent (SLE) is reported as the change in ice sheet mass expressed in mm SLE, where negative values indicate mass loss and thus a positive contribution to sea level rise. We acknowledge that this sign convention can be counterintuitive when compared with figures in the IPCC report that show sea level rise as positive. We will add a clarifying note in the caption. We will also switch the colours of RCP 2.6 and RCP 8.5 to align with the conventions used in the majority of published studies.

In all cases where comparative metrics are used such as MSE, MAE, KLD, JSD, etc. do they have units? I would be surprised if MSE and MEA didn't.
MSE and MAE are calculated on SLE values and are therefore reported in units of mm^2 SLE and mm SLE, respectively. KLD and JSD are dimensionless divergence measures. We will add units to all reported values in the revised tables and captions.

Line 5: LSTM – introduce acronym
We will introduce the full term “Long Short-Term Memory (LSTM)” at its first occurrence.

Line 25: "climate projections": if I was being pedantic, the references in this sentence are pertaining to projections of sea level rise from ice sheet loss rather than the whole climate.
Thank you, we will revise this to “ice sheet projections” to be more precise.

Line 66: "accurately … projects future sea level": again pedantic, it's probably not correct to suggest that this model accurately projects the future sea level since we don't have this observation; would it be better to say it accurately emulates the sea level rise components from ISMIP models?
We agree. This is an important distinction, and one that Reviewer #3 also raises in some detail. We will revise this and similar language throughout the manuscript to clarify that ISEFlow accurately emulates the ISMIP6 model projections, rather than claiming accuracy of real-world future sea level.

Line 88: Coupled (not Climate) Model Intercomparison Project
Corrected. We will update this throughout the manuscript.

Line 89: "yearly-averaged atmospheric and oceanic forcing anomalies": it would be nice to have a list of these forcings, if it isn't too long

We will add a comprehensive list of the input forcing variables used in ISEFlow to the supplement. This will include all climate forcings and ISM characteristics used across the different model configurations presented in the paper.

Line 97: signpost to figures A1 and A2 on the ISMIP6 regions somewhere in this paragraph.

We will add references to Figures A1 and A2 in this paragraph.

Line 115: 635 projections in the training set, 136 projections in the validation set. How were these numbers decided? And what makes up the total of 771 projections? Presumably this is some number of ice sheet models taking forcing data from a number of CMIP models under some number of scenarios?

The total projections correspond to the full set of ISMIP6 simulations available, which arise from the combination of multiple ice sheet models, each driven by forcings from different CMIP climate models under different scenarios (RCP 2.6 and RCP 8.5) and with varying model configurations (e.g., different basal melt parameterizations, ocean sensitivities). For the GrIS, for example, a 70/15/15 training/validation/test split was used, resulting in 635 training and 136 validation and 136 test projections. We also (see Reviewer #3) will update the values for the AIS, which are currently incorrect. We will clarify the composition of these projections and the rationale for the split in the revised manuscript.

Line 141: "another model": "other models"?

Corrected.

Line 200: 256 GB RAM, I assume

Yes, this is correct. We will ensure the units are clearly stated.

Lines 189–190: This is a very opaque sentence for somebody not versed in machine learning.

We will rewrite this sentence to be more accessible to a broader audience, explaining the GP kernel structure in plain language.

Line 284, related to my first question: This paragraph reports that including variables beyond temperature improves emulations, which isn't surprising. (Can you confirm whether this is global mean temperature or local temperature)? However, the mainstream climate emulators would generally give you only global mean surface temperature from emissions scenarios, which would allow a user to produce climate projections from any emissions scenario and not just the ones run by CMIP/ISMIP models. Therefore, can SLR projections from the GrIS and AIS components be produced from ISEFlow which are "good enough", even if not ideal? Figure 5, if I interpret it

correctly, seems to suggest so. This would really help to find a valuable use case for this model by piggybacking off GMST projections by emulators.
The temperature used in Table 3 is local surface air temperature as provided by the ISMIP6 forcings, not global mean surface temperature (GMST). We will clarify this in the revised text. Regarding whether ISEFlow could produce projections using only GMST: this is an interesting direction. While the sensitivity tests (Table 3) demonstrate that using local temperature alone yields lower accuracy than using all available forcings, the results do suggest that temperature-only projections are still informative. This is indeed a promising use case for coupling ISEFlow with climate models such as FaIR or FACTS, and we will note this explicitly in the Discussion as a direction for future work.

Line 301: D statistic of 0.158. Is this good? I have no feeling of what a good value is. Are there units here?
The KS D statistic is dimensionless where lower values indicate similar distributions (with 0 being identical), and higher values indicate less overlap between distributions. A value of 0.158 indicates a modest but detectable difference between the RCP 2.6 and RCP 8.5 projected distributions. We will add context in the text to help readers interpret this value, including a brief explanation of the KS test and its range.

Caption to figure 5: you can drop the word "carbon" to be more general and accurate.
We will change “carbon emission scenarios” to “emission scenarios” throughout the manuscript.

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.