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

Reviewer #1

In my opinion this is a useful model that will add to the rapidly growing toolbox for modelling sea level rise and its components. 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?

ISEFlow, in its current form, requires regional CMIP climate forcing inputs (surface air temperature, SMB components, ocean properties, etc.) and is therefore limited to scenarios for which such forcing data have been generated by a GCM. This differs from emulandice, which uses GSAT as its sole climate input and can be coupled to simple climate models such as FaIR to generate projections for any emissions scenario. We have added a discussion of this trade-off in the Discussion section (Section 4). We discuss how a temperature-only version of ISEFlow could be developed and coupled to simple climate models such as FaIR or FACTS to achieve broader scenario coverage, while acknowledging that this would entail a reduction in emulation accuracy relative to the full ISEFlow configuration. We have also clarified in the Discussion that ISEFlow is intended as a complementary tool to existing emulators, each serving different use cases, rather than as a replacement.

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.

Thank you for pointing this out. The sign convention used in this study follows the ISMIP6 convention, in which SLE anomalies are reported as deviations from a control experiment, with negative values indicating ice mass loss relative to the control (i.e., a positive sea level contribution). We have added a note in the Figure 4 caption clarifying this convention. Regarding the colors, we have updated Figure 4 so that RCP 2.6 is shown in blue and RCP 8.5 in orange/red, consistent with the convention used in much of the sea level literature.

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.

You are correct. MSE is reported in mm² SLE and MAE in mm SLE. KLD, JSD, and the KS D statistic are dimensionless. Units have been added to the Table 1 caption (and all relevant tables throughout the manuscript) to make this explicit.

Line 5: LSTM – introduce acronym

The acronym LSTM (Long Short-Term Memory) has been introduced on first use in the abstract. Specifically, the abstract now reads "a deep ensemble of recurrent neural network models" (avoiding the acronym in the abstract) and the full definition is given at first use in the Methods section (Section 2.2): "a Deep Ensemble of Long Short-Term Memory (LSTM) models."

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. Many examples of climate projection models are given in Romero-Prieto et al., accepted (https://egusphere.copernicus.org/preprints/2025/egusphere-2025-2691/). Or you could say ice sheet/cryosphere emulations to be specific.

The phrase "climate projections" has been changed to "ice sheet projections" at this location (now in the second paragraph of the Introduction). Similarly, "climate simulations" has been changed to "ice sheet simulations" at the preceding line.

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?

This is an important point. The language throughout the manuscript has been revised to consistently state that ISEFlow "accurately emulates" ISMIP6 projections, rather than "accurately projects future sea level." The abstract, introduction, conclusions, and results sections have all been updated accordingly. We are careful to distinguish between emulation accuracy (assessed against held-out ISMIP6 simulations) and predictive accuracy for the real world.

Line 88: Coupled (not Climate) Model Intercomparison Project

Corrected.

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

A complete list of input variables for both ISEFlow-AIS and ISEFlow-GrIS has been added as Appendix A2 (Tables A2 and A3), including variable names, units, temporal structure, and data type. The main text now refers readers to these tables. Additionally, the sensitivity test input configurations (Table A4) and the Regional ISEFlow input configuration (Table A5) have been added to the Appendix for full transparency.

Line 97: signpost to figures A1 and A2 on the ISMIP6 regions somewhere in this paragraph.
A reference to the sector maps (Figures A1 and A2 in the Appendix) has been added to the data description paragraph in Section 2.1, where the spatial partitioning of the AIS into 18 sectors and the GrIS into 6 sectors is introduced.

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?
Thank you for identifying this ambiguity in the original preprint. The numbers were not well explained and the updated manuscript provides a more complete description. The dataset consists of projections that each correspond to a combination of ice sheet model, CMIP climate forcing model, emission scenario (RCP 2.6 or RCP 8.5), and experimental configuration. For the AIS, the full dataset contains 3,180 projections, and for the GrIS, 907 projections. The training set comprises 70% of projections (2,226 for AIS; 635 for GrIS), with validation and test sets each containing 15% (477 for AIS; 136 for GrIS). The manuscript now clearly states these total numbers and the basis for the splits in the Data section (Section 2.1). The larger AIS dataset relative to the GrIS reflects the larger number of experiments and model configurations in the ISMIP6 Antarctic ensemble, including additional sensitivity experiments.

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

Line 200: 256 GB RAM, I assume
Yes, the text now reads "256 GB RAM" (referring to the compute node memory).

Lines 189-190: This is a very opaque sentence for somebody not versed in machine learning.
The sentence describing the Gaussian Process kernel has been rewritten in the revised manuscript to be more accessible. The text explains that the power exponential kernel measures similarity between inputs as a smoothly decaying function of their distance, and that the nugget kernel adds a small independent noise term to each observation to account for variability due to factors not included as inputs.

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 forcing used in the sensitivity tests is local (sector-level) surface air temperature, not GSAT. This has been clarified in the manuscript. As noted above in our response to the first comment, we have added a discussion in Section 4 of the possibility of developing a temperature-only (or GSAT-based) version of ISEFlow that could be coupled to simple climate models. The sensitivity test results (Table 3) show that the temperature-only configuration, while less accurate than the full configuration, still meaningfully outperforms the GP emulator in projection accuracy for the AIS (MSE 2.38 for NN vs. 3.03 for GP). This suggests that a future temperature-only or GSAT-coupled ISEFlow version could provide a "good enough" emulation while enabling projections for any emissions scenario.

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, and its interpretation depends on context. We have added explicit context in the revised manuscript. To provide an empirical ceiling, we now report the KS D statistic applied directly to the ISMIP6 simulations themselves (D = 0.162 for AIS, D = 0.172 for GrIS), which represents the maximum scenario separation achievable by any emulator trained on this data. ISEFlow achieves D = 0.16 for AIS and 0.15 for GrIS, close to these ISMIP6 values, suggesting it captures nearly all of the scenario signal present in the training data. We also report bootstrapped 95% confidence intervals for all D statistics to support the interpretation of their significance and magnitude.

Caption to figure 5: you can drop the word "carbon" to be more general and accurate.
The word "carbon" has been removed from the figure caption. The caption now refers to "emissions scenarios" rather than "carbon emissions scenarios."

Reviewer #2

1 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.
Thank you for this suggestion. The Introduction has been revised to clarify this point and to signal that the KS D statistic will be the primary metric for this purpose. The Introduction also clarifies that the SHAP-based classification analysis provides supplementary insight into which input variables carry the most scenario-distinguishing information.

1 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.

The GP emulator achieves higher KS D statistics for the GrIS in particular, but this is partly an artifact of the GP emulator's underestimation of the ISMIP6 ensemble spread (as evidenced by its much larger KLD and JSD values in Table 1). When the GP collapses the predicted distribution into a narrower range than the true ISMIP6 distribution, the two RCP sub-ensembles (which are already different in size and scope in ISMIP6) appear more separated in the emulated space than they actually are in the training data. A wider separation in the emulated distributions is not necessarily a better representation of the true scenario difference if that separation is partly an artifact of poor distribution approximation. We have reworked Section 3.3 (Scenario Sensitivity) to address this. We now report the KS D statistic computed on the raw ISMIP6 simulations themselves as an empirical reference point. ISEFlow's D statistics closely match this reference, while the GP emulator's higher GrIS D statistic exceeds the ISMIP6 ceiling---a sign that it is over-separating scenarios due to underestimating the ensemble spread. We have revised the language throughout Section 3.3 to avoid claiming that ISEFlow "outperforms" the GP across the board on scenario distinction; instead, we describe the nuanced trade-offs that Table 3 and Table 1 together reveal. We also note that Table 4 was removed and the KS D Statistic was moved to Table 1.

1 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).

We now report the KS D statistics computed directly on the ISMIP6 simulations (D = 0.162, 95% CI: 0.151–0.176 for AIS; D = 0.172, 95% CI: 0.164–0.181 for GrIS) as an empirical ceiling for scenario separation. We note that these values are statistically significant but modest, indicating that even the true ISMIP6 ensemble does not cleanly separate RCP 2.6 from RCP 8.5 at year 2100. We also add a note that the KS test compares two unequal sub-ensembles (the ISMIP6 RCP 2.6 subset is smaller and samples different portions of the uncertainty space than RCP 8.5), so the test cannot cleanly isolate genuine scenario dependence from structural differences in the ensemble design. Studies with more balanced ensemble designs have found limited scenario differences in ice sheet projections before 2100, and we now cite these in the Discussion.

1d.) 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 suggestion. Bootstrapped 95% confidence intervals for all KS D statistics have been added to the Results section. The revised text states, for example, that ISEFlow-AIS achieves a D statistic of 0.161 (95% CI: 0.154–0.175), compared to 0.124 (95% CI: 0.119–0.142) for the GP emulator, and notes that the difference is statistically distinguishable under bootstrap resampling ($\Delta D = 0.033$, 95% CI: 0.020–0.049). We also further clarify the magnitude of D = 0.16 relative to the ISMIP6 reference value, and explicitly caution that even the true ensemble shows only modest scenario separation, consistent with findings in the broader literature. Language referring to a "strong" ability to distinguish has been softened throughout.

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 Methods section has been revised to make this distinction clearer. The GP and NN emulators appearing in Table 3 (forcing sensitivity) are both trained without ISM characteristics, holding the input set constant so that architectural differences are the only source of variation. Neither is the full ISEFlow model. The full ISEFlow model (with all forcings and ISM characteristics) is the one reported in Table 1. We have also added a new Appendix (Appendix C) documenting the key differences between our GP implementation and the original emulandice, and revised Table 1's caption to refer to the emulator as the "GP emulator (following a similar design to emulandice; see Appendix C)" throughout, to avoid any confusion.

3.) The utility of the "classification model" is not clear to me. The results from this model (lines 344-354) are framed as showing the ability of ISEFlow to distinguish between emissions scenarios. However, in my mind, although the emulator architecture is the same, this "classification model" is a completely distinct emulator from the regular ISEFlow that is used to predict SLE. Two major issues should be addressed:

a.) 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 framing of the classification model has been clarified substantially. Both the Methods section (Section 2.4) and the Results section (Section 3.3) now state that the classification model

is a diagnostic tool, entirely separate from ISEFlow, with the sole purpose of identifying which input variables most strongly encode emission scenario information. The classification model does not directly test ISEFlow's scenario sensitivity; rather, it provides guidance on which variables in the climate forcing data are most useful for distinguishing scenarios, thereby informing future emulator design choices. The results from ISEFlow's scenario sensitivity are based on the KS D statistics reported from ISEFlow's own predictions, not from the classifier.

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.

Thank you for pressing on this. We have added discussion in Section 4 clarifying the scientific value of the classification SHAP analysis. The key insight is not only that temperature predicts scenario (which is expected), but that the combined contribution of all other forcings is comparable in magnitude to that of temperature alone. This finding supports the design choice to train ISEFlow on a broad set of climate forcings. It also provides guidance for future emulator development, even if GSAT-based emulators capture the primary scenario signal, additional regional forcing variables encode scenario-relevant information that can improve both emulation accuracy and scenario sensitivity. We also note in the Discussion that the fact that the classifier can distinguish scenarios at all (77.8% for AIS; 87.9% for GrIS) provides indirect evidence that the ISEFlow input features do contain the scenario signal---the question is how much of it each variable carries.

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).

This has been addressed in the uncertainty framework section (Section 2.2). The text now clarifies that "emulator uncertainty" refers specifically to epistemic uncertainty arising from limitations in the emulator's architecture, training process, or coverage of the input space---it does not capture the uncertainty in the ISMIP6 ensemble itself (i.e., the spread across ISMs and CMIP forcing models), which is instead reflected in the "data coverage uncertainty" component. The total predictive uncertainty combines both components and represents the full spread of outcomes the emulator attributes to a given set of input conditions. We have also added text clarifying that for most applications, the total predictive uncertainty is the relevant quantity, while the decomposition is primarily useful for diagnostic purposes such as identifying poorly constrained regions of the input space or guiding the design of future ISMIP experiments.

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.
Your interpretation is correct. For LARMIP, there is no separate training phase, the 20.3 minutes reflect the wall-clock time required to generate all 20,000 samples by convolving the 20,000 randomly generated basal melt forcing time series with the pre-computed linear response functions for each basin. The LARMIP "training time" in Table 1 is therefore the total time to generate all predictions, and we have clarified this in the Table 1 caption. We have also updated the Table 1 caption with this point, and added a sentence in the Methods section (Section 2.3) stating that for LARMIP, this time represents the full prediction generation process rather than a training phase. We also now more clearly state in the Methods section that we use ISMIP6 AIS GSAT as the climate forcing input, following the methodology of Levermann et al. (2020), and that we do not add an SMB correction as was done in the AR6 assessment.

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.
The Introduction has been revised to explain the approach earlier. The paragraph introducing ISEFlow and the goals of the study (Section 1, final paragraph) now specifies that the assessment of whether ISEFlow's learned relationships are physically consistent with prior understanding will use SHapley Additive exPlanations (SHAP), applied to both ISEFlow's projections and a diagnostic classification model. The text notes that SHAP allows us to identify which input variables most influence emulated sea level projections, and that we assess these against domain knowledge from prior ice sheet modeling studies.

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 (or appendix by adding to the existing Table A1 and A2).

a.) Provide a table that lists all input variables used to train the emulator that produced the results in Table 1.

Complete input variable tables for ISEFlow-AIS (Table A2) and ISEFlow-GrIS (Table A3) have been added to the Appendix, listing all features including names, units, temporal structure, and number of dimensions. These are referenced in the main text.

b.) Provide a table that lists the input variables used for the "All", "Temperature", and "SMB" versions of the emulator shown in Table 3.

A sensitivity test input variable table (Table A4) has been added to the Appendix, showing which features are included in the "All", "Temperature", and "SMB" configurations for both AIS and GrIS. The table uses checkmarks to indicate inclusion/exclusion of each variable for each configuration. The "All" configuration is clarified as corresponding to the same climate forcings as the full ISEFlow emulators but without ISM characteristics.

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 (climate forcings, ice sheet model parameters, initial state, etc.).

The Introduction has been updated to mention that the deep uncertainty associated with Marine Ice Cliff Instability (MICI), as discussed in DeConto and Pollard (2016). The text (Discussion) notes that there are other uncertainties beyond those that are quantified here that can be topics for future research.

Line 24: Change "climate simulations" to "ice sheet simulations"
Line 25: Similarly, change "climate projections" to "ice sheet projections"

Corrected.

Line 26: I suggest changing "small-scale" and "large" to "computationally efficient" and "computationally expensive," however I defer to the authors on this. I'm suggesting this because "small-scale" and "large" can be used to refer to spatial scales but I think what's actually being discussed here is the computational scale.

The text has been changed to "computationally efficient models that approximate computationally expensive, more complex physical models."

Line 28: Change "climate model" to a more general term like "physical model"
Line 29: Change "climate system" to "physical system"

Corrected.

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 (this is related to my first minor comment above). This doesn't

need any sort of detailed explanation - it's not your responsibility in this paper to explain all of these nuances - but I think it's worth mentioning for completeness.
The AR6 description in the Introduction has been updated to mention "structured expert judgment" alongside ISMIP6, emulandice, and LARMIP as inputs to the AR6 ice sheet assessment.

Line 55: Specify that these fixed parameters resulted from choices made in the ensemble design of ISMIP6 and not limitations imposed by Emulandice itself.
The revised text now clarifies that emulandice uses fixed parameters for glacier retreat, sub-shelf basal melt, and ice shelf collapse "that result from the ISMIP6 experimental design" rather than from limitations in the emulandice framework itself.

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."
The total number of projections is now stated explicitly (3,180 for AIS, 907 for GrIS) and the train/validation/test split is described in terms of both percentages and absolute counts. The word "full" has been removed; the text now states "136 86-year projections" to clarify that each projection spans the complete 2015–2100 time period.

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, and this is now stated in the revised manuscript. The uncertainty section (Section 2.2) now includes a paragraph explaining that the total predictive uncertainty is the relevant quantity for most downstream applications and reflects the combination of data coverage uncertainty (from ISMIP6 variability) and emulator uncertainty (from the emulator's own limitations). The text clarifies that the decomposition is primarily of diagnostic value and that the two components are combined in the uncertainty metrics reported in Table 5.

Line 158: Should this read: "areas where the data is highly variably or missing"?
The text in this entire section has been reworked so that it is more clear. The sentence describes data coverage uncertainty as arising in regions "where the training data exhibits variability or disagreement."

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.
The Methods section (Section 2.3) now specifies that we use ISMIP6 AIS global surface air temperature (GSAT), derived from the six CMIP models used in the ISMIP6 Antarctic ensemble, as the climate forcing input to LARMIP. The output of the LARMIP convolution is the dynamic AIS sea level contribution only. We also explicitly state that no SMB correction is applied (in

contrast to the AR6 assessment), and note the implications of this for the LARMIP results shown in the paper.

Line 233: There's a typo here: "that was help out of the original ISMIP6 ensemble". Please rephrase.
Corrected.

Line 237: Please clarify what an "individual projection" is. Is it the projection for one ISMIP6 experiment (i.e., one emissions scenario, one climate model forcing, and one set of model parameters) or something else?
The term "individual projection" is now defined in that section. The text clarifies that each projection corresponds to one ISMIP6 experiment: a unique combination of ice sheet model, CMIP forcing model, emission scenario, and experimental configuration, producing a single 86-year SLE time series (2015–2100).

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.
A dedicated paragraph has been added to Section 2.3 (now between the GP emulator description and the SHAP section) that defines MSE and MAE precisely. The error for each projection is calculated as the difference between the emulated SLE anomaly and the corresponding ISMIP6 SLE anomaly at each of the 86 time steps; these per-step errors are averaged over the full projection period to yield a per-projection error. MSE and MAE are then the mean of the squared and absolute per-projection errors, respectively, averaged over all projections in the test set.

Table 1: Add units to all reported values in the table, either within the table or in the header. If MSE and MAE (or other metrics) are reported in units of mm SLE, please make sure to include the "SLE".
Units have been added to the Table 1 caption: MSE is in mm^2 SLE, MAE is in mm SLE, KLD, JSD, and KS D statistic are dimensionless, training time is in minutes, and prediction time is in seconds.

Line 283: I suggest adding "the emulator would need to be re-trained and" before "the rankings are likely to differ."
The suggested text has been incorporated.

Line 284: I suggest replacing the word "additional" with "individual climate"
Corrected

Line 289: Would it be a bit more specific to replace "offers more information for modeling SLE" with "is a better predictor of SLE"?

The text in this section has been substantially improved and rewritten based on feedback here (next comment) as well as from other reviewers. This statement no longer exists in the updated text.

Lines 295-316: I suggest moving and combining the text in these paragraphs with what's in Section 3.3.

The discussion of scenario sensitivity results from the forcing sensitivity ablation (Table 3 KS D results) has been moved to Section 3.3, where it is discussed alongside the full ISEFlow vs. GP emulator scenario sensitivity comparison. The two sets of results (Table 3 and Table 1) are now framed as complementary, with Table 3 holding inputs constant to isolate architectural effects and Table 1 comparing the full designs.

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." Although the Temp- or SMB-only results are less accurate, it hasn't been shown that they are insufficient.

Corrected. While this specific text no longer exists in the manuscript, the overall tone has been softened around these points.

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

A clarifying sentence has been added to the Figure 3 (SHAP all-variables) caption and in the Results text, noting that the SHAP values are computed using the full ISEFlow models with all input variables and ISM characteristics.

Figure 4: There are a couple of things that are unclear with the ISMIP6 time series that shown:

a.) It would be useful to show the RCP 2.6 and RCP 8.5 time ISMIP6 projections in two different colors. This can probably be accomplished by using the same blue/orange colors but keeping the thinner lines for ISMIP6.

Figure 4 has been updated to show RCP 2.6 and RCP 8.5 ISMIP6 projections in different colors (blue and orange, respectively), using thin lines to distinguish the ISMIP6 ensemble from the emulator medians, which are shown as thicker lines. There are, however, so many lines that it is still somewhat difficult to separate, but the colors were changed.

b.) It's not clear whether the AIS ISMIP6 projections all start at 0 SLE in 2015. Some of the lines look like they're offset from zero. These lines should be shifted or, if they do start at 0 but then "jump" instantly to another value in 2016, this should be mentioned in the caption.

A note has been added to the Figure 4 caption explaining that some ISMIP6 projections do not begin at exactly zero anomaly in 2015, which reflects differences in model initialization and reference states within the ISMIP6 ensemble. The projections are reported as anomalies relative to a control simulation per ISMIP6 convention, and the small offsets at 2015 are a feature of the original data, not a post-processing artifact.

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".
Corrected.

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.
The Discussion paragraph on ISEFlow's role in the CMIP7/AR7 cycle has been updated to note that emulandice was previously used to fill in intermediate SSPs in AR6, and that ISEFlow could serve a similar complementary role.

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.
Additional text has been added to the Discussion section elaborating on how ISEFlow can contribute to ISMIP7 design. Specifically, ISEFlow can predict how different model characteristics affect projected sea level, helping to prioritize which model configurations or experiments are most informative for reducing projection uncertainty. It can also be used to emulate what the existing ISMIP6 model configurations would project under CMIP7 forcing data, enabling early screening of CMIP7 climate models for inclusion in the ISMIP7 forcing ensemble.

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. If this is the case, this duplicate text here should be removed. If I am misunderstanding and this text is introducing a different idea, it should be rewritten to be more clear.
The duplicate text has been removed from the Conclusions section. The Discussion section now contains the full discussion of ISEFlow's potential role in CMIP7/AR7, and the Conclusions section summarizes this concisely without repeating the detail.

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

Reviewer #3

L3, etc: Model Land. See General comments. "Accuracy" of future sea level projections is impossible to achieve, or test, for multiple reasons – reword everywhere to accuracy/success in emulating ISMIP6 model projections. As you say at L138, the degree to which climate and ice sheet models simulate the real world is not the focus here.

We thank the reviewer for this important point. Throughout the manuscript, language has been revised to consistently refer to "emulation accuracy" or "accuracy in emulating ISMIP6 projections" rather than "accuracy of future sea level projections." The abstract, introduction, results, discussion, and conclusions have all been reviewed and revised accordingly. For example, the abstract now states that ISEFlow is "designed to accurately and efficiently emulate sea level contributions from both ice sheets," and the conclusions state that ISEFlow "creates more accurately emulated ISMIP6 ice sheet model projections." We have been careful to maintain this distinction throughout.

L10, etc: Scenario distinction. The ability to capture the effects of varying emissions scenarios is not robustly tested here, because it is convolved with the uneven ISMIP6 design (see later). It is also repeatedly asserted that emulandice should have given very different projections for different scenarios. LARMIP showed only a little more difference, after the SMB correction was added for AR6. For equivalent projections, emulandice showed -1cm difference between medians, and LARMIP2 +2cm, and the AR6 assessed that there is "no consensus" on scenario difference between climate and ice sheet models (without marine ice cliff instability, MICI). Recent studies that have identical ensemble designs for these two scenarios, rather than uneven designs like ISMIP6, show little difference before 2100 (Lowry et al., 2021; Coulon and Klose et al., 2025). Of course, some uncertainty in the emulandice projections was introduced by using GSAT as the climate forcing, discussed elsewhere, as well as from the GP itself. Edwards et al. (2021) also demonstrate how uncertainty in the lower scenarios could have been reduced with a more even ISMIP6 design (see end of our Methods). But there is still insufficient evidence from previous studies or this one of a 'true' scenario dependence in ISMIP6, or in the slightly larger LARMIP ensemble, or in other non-MICI models, to evidence the assertions that there should be a clear difference between scenarios in probabilistic projections that evenly sample uncertainties. See also the interpretation of the Figure 4 results below.

We have substantially revised the language around scenario distinction throughout the manuscript. The manuscript no longer asserts that there should be a clear difference between emission scenarios; instead, it describes ISEFlow's ability to reproduce the degree of scenario separation that exists within the ISMIP6 ensemble itself. We now acknowledge the limitations of the KS D test when applied to ISMIP6's uneven ensemble design, citing Lowry et al. (2021) and

Coulon and Klose et al. (2025) as evidence that more balanced designs show limited pre-2100 scenario differences. We calculate the KS D statistic applied directly to the ISMIP6 simulations as a reference, and note that ISEFlow approximates this value, suggesting the emulator is a similar degree of scenario separation as is present in the training data. We also note that the absence of an SMB correction in our LARMIP comparison (unlike the AR6 assessment) contributes to LARMIP's stronger apparent scenario separation in Figure 4. The language about scenario distinction has been revised from "strong ability to distinguish" to "reproduces the scenario separation observable in ISMIP6," and we have softened the language comparing ISEFlow to the GP emulator on this metric.

L13-14, etc: Policy-focused projections. See General comments. These uses are not supported by the current tool and manuscript, for various reasons described below – revise throughout.
We agree that policy-relevant probabilistic sea level projections require additional steps beyond the scope of this work, and the revised manuscript has been revised to reflect this. The abstract and introduction no longer claim that ISEFlow directly "supports climate risk assessments" or "informs policy decisions" in the final product sense. The Discussion now lists the additional steps required before ISEFlow could contribute to policy-relevant projections. ISEFlow is presented as a tool for efficiently exploring the ISMIP6 ensemble and as an input to the research process, consistent with a multi-emulator assessment approach.

L54, etc: GSAT vs regional climate forcings. See General comments. "Relies on" GSAT. There is little to no mention in the manuscript of why this choice was made or needed, including:
1. to include correlation of land ice regions induced by global climate change;
2. to be used within the sea level calculation framework FACTS, which requires all modules to use GSAT trajectories from FaIR for the 'book-keeping' across sea level components;
3. to use a more comprehensive and defensible prior for climate change projections (FaIR probabilistic projections using the AR6 assessed climate sensitivities (ECS and TCR), rather than the CMIP6 ensemble of opportunity;
4. to rapidly and easily update projections under any emissions scenario (e.g. current NDCs, new scenarios).

Continued…
We fully agree that the original manuscript unfairly characterized emulandice's use of GSAT as a limitation without acknowledging the deliberate design reasons behind it. The Discussion has been rewritten to present the trade-offs between GSAT-based and regional-forcing-based emulators in a balanced way. The revised Discussion now includes the four reasons the reviewer identifies for emulandice's GSAT choice. The text also notes that ISEFlow's regional forcing approach introduces dependencies on a specific set of GCM simulations, limits applicability to scenarios with available forcing data. ISEFlow is described as complementary to emulandice, appropriate for different use cases.

L71: Emulandice could have used more input variables: computational efficiency was not the reason. The main reason was time, to deliver for AR6 at short timescales while the ISMIP6 datasets – and their documentation and publications – were still evolving: I wanted to focus on the ISMIP-wide parameters that were defined for all (or at least most) models, and adding other inputs would have made it more complex to design, validate and justify in the time we had, especially given that the multi-model emulation approach was new. Here, using more information is plenty of justification in itself, and a strength of this study.
We apologize for the inaccurate characterization. The text now states that emulandice was designed around a smaller set of ISMIP-wide input variables that were more consistently defined across models and more tractable to validate in a new multi-model emulation framework, given the limited time available to deliver projections for AR6 and the continued evolution of the ISMIP6 datasets and documentation at that time. The claim that computational efficiency was the limiting factor has been removed.

L85-98: The manuscript is very unclear on the simulation data used, in particular:
* that only the RCP projections (Seroussi et al.) are used, not the later SSP projections (Payne et al.);
* that the additional simulations used by Edwards et al. (2021) are not included (113 for Antarctica and 22 for Greenland), the majority of which were designed to improve its accuracy by exploring parameter extremes and interactions, and the low emissions scenario;
* the additional simulations that seem to be added, as above;
* exactly which climate variables are used in the main version of ISEFlow: I eventually found this in Figure 3, but they should be clearly listed in the text.

Continued…
The Data section (Section 2.1) has been revised to address all these points. It now states that the dataset used for ISEFlow consists of the RCP-based ISMIP6 projections (Seroussi et al., 2020 for AIS; Goelzer et al., 2020 for GrIS) and does not include the additional supplementary simulations used by Edwards et al. (2021). The total number of projections for each ice sheet (3,180 AIS, 907 GrIS) is stated explicitly. Complete input variable tables have been added to the Appendix (Tables A2 and A3) listing all climate forcings with units, temporal and spatial averaging, and data origin (GCM vs. RCM where applicable). The ocean forcing variables are described as depth-averaged and spatially extrapolated to sub-shelf cavities following Jourdain et al. (2020), and this is noted in the table caption.

L99: Again, computational efficiency is not the reason emulandice did not use higher spatial resolution: the main reason was to ensure independence of the projections (to avoid ignoring or modelling any correlation of model errors between basins), and to reduce the complexity of the analysis.

Corrected. The text now states that the coarser spatial aggregation in emulandice was chosen primarily to ensure independence of projections across regions (avoiding the need to model spatial correlations of model errors between basins) and to reduce the complexity of the analysis.

L101: Basins are not "the maximum amount of spatial information possible", grid cells are.
The phrase has been revised to "a resolution closer to the original resolution" and "finer spatial partitioning" rather than claiming it represents the maximum possible spatial information.

L112: ISEFlow could predict the results of new combinations of existing models and inputs, but not the results from new model updates or parameters.
The text has been revised to reflect this more precisely. The sentence states that ISEFlow "can approximate projections for new combinations of existing model characteristics and forcings within the sampled ISMIP6 design space," rather than implying it can extrapolate to genuinely new model configurations or parameters outside that space.

L115: You seem to have the same number of projections for training and testing for each ice sheet, which I can't believe is right..? It also needs explaining why the total numbers are more than double those in Edwards et al. (2021), given that Payne et al. SSP projections and the extras in that study are also not used. It doesn't seem to be multiplied by basin, as dividing by 18 or 6 doesn't give a round number. Is it all the extra simulations, such as ocean-only forcings? Please clarify, and document the experiments if they are not published elsewhere.
Thank you for identifying this inconsistency in the preprint, the numbers in that version were indeed incorrect and have been corrected in the revised manuscript. The full dataset contains 3,180 projections for the AIS and 907 for the GrIS, with a 70/15/15 train/validation/test split. The projections are directly processed from the ISMIP output repositories (via GHub) and represent projections for each available experiment (of all Tiers) for the 18/6 sectors. These are consistent with the published ISMIP6 Antarctic and Greenland dataset. It doesn’t perfectly divide by 18/6 sectors due to the removal of a small number of outliers as described in Van Katwyk et al., 2023.

L135: I don't find this to be very clear: you mention incomplete data as a contributor to emulator uncertainty, but also as the source of data coverage uncertainty ("where the training data is [sic] sparse or unevenly distributed"). And you don't show any separation of these in the manuscript, even though their separation is mentioned as a strength. Can you clarify whether data coverage uncertainty is a part of what others would usually include in emulator uncertainty? And can you provide any quantitative examples or illustrations of this? If not, or the paper would get too long, I suggest shortening this section and demonstrating these things in another paper.
The uncertainty framework section (Section 2.2) has been substantially rewritten to clarify differences between uncertainty types. We now clarify that in the ML uncertainty quantification literature, data coverage uncertainty is typically treated as a form of aleatoric uncertainty

(irreducible variability in outcomes given the inputs), while emulator uncertainty is epistemic. However, we acknowledge that in the context of emulators, where the "data" are generated by other models rather than observation, the distinction is not perfectly clean, and incomplete training data coverage can contribute to both components. We have shortened the technical discussion of uncertainty decomposition and focused it on what is actually demonstrated in this paper: the total predictive uncertainty is what is reported and evaluated (PICP, MPIW, CRPS, Winkler Score in Section 3.4), while the decomposition into data coverage and emulator components is described as a diagnostic capability. We do not show explicit separation of the two components in the quantitative results, and we have revised the text to not overstate this as a key demonstrated result, noting instead that the architecture enables this separation in principle and that a more detailed investigation is left for future work.

L146: And also because of the structural difference between models – many combinations of models and inputs/choices do not physically exist.
Thank you for this suggestion. While this text has been substantially revised and no longer exists within the manuscript, this sentiment is reflected in the Discussion section where it mentions that there are many sources of uncertainty that cannot be accounted for fully within these two buckets.

L175: Do the KLD and JSD metrics here include the emulator uncertainty estimates, or just the mean values? How are you ensuring that the emulator is not over-fitting (e.g. too many inputs; testing data too similar to training) and/or under-estimating its uncertainties? I would expect coverage metric(s) such as those in Section 3.4 to be part of the validation.
The KLD and JSD metrics are computed using the emulator's predicted values (not including the uncertainty distribution), evaluated against the distribution of ISMIP6 projections at year 2100. This is now stated in the Methods. Regarding overfitting, ISEFlow is evaluated on a held-out test set that was withheld throughout training and hyperparameter optimization. The PICP and Winkler Score results in Section 3.4 confirm that the emulator's uncertainty intervals are well-calibrated on this unseen data, providing evidence against overfitting. A sentence to this effect has been added to the sensitivity testing discussion.

L181: Call these "approximations to" the previous emulation tools, or similar, because you are not running the actual code or (at least for emulandice) reproducing the methods exactly.
The revised manuscript consistently refers to our baseline emulators as "an approximation to the emulandice design" (GP emulator) and "LARMIP linear response functions applied with ISMIP6 inputs." Table 1 now labels the GP baseline as "GP emulator (following a similar design to emulandice; see Appendix C)."

L182: It is important to state clearly that you are not adding any SMB component to LARMIP here, given your context of the AR6 (discussing scenario distinction, policy-relevance

projections, etc). Estimating only the dynamic component produces stronger scenario dependence in LARMIP (e.g. in Figure 4), because of the opposing dynamic and SMB responses to warming.
This has been addressed in the Methods section (Section 2.3), which now states that no SMB correction is applied to the LARMIP output. The text also notes that as a consequence, the LARMIP results shown in this paper represent only the dynamic component of AIS sea level contribution and likely show stronger scenario dependence than the AR6 LARMIP results, due to the opposing dynamic and SMB responses to warming. The Figure 4 caption has also been updated to note this.

L184-187: This is not clear or accurate. LARMIP is not only for understanding basal melting; you have called it an emulator throughout until now (which I think is fine: it is being used for this purpose); you do not say what additional data is used from ISMIP6; you do not say what you mean by comparing alongside emulandice (e.g. the AR6 main assessment compares them with the SMB correction added to LARMIP, which is not applied here).
The LARMIP description in Section 2.3 has been rewritten. We now describe LARMIP as a "linear response model" (rather than "emulator" in the general sense) that estimates the dynamic AIS response to ocean forcing, clarify the ISMIP6 GSAT inputs used, explain why no SMB correction is applied, and note the precedent from the IPCC AR6 for including LARMIP in this comparison.

L188: You do not implement or describe the emulandice architecture correctly…
A dedicated Appendix (Appendix C) documents these differences in detail. The text in the main Methods section (Section 2.3) introduces the GP emulator as "a GP emulator following a similar design to emulandice" and refers readers to Appendix C for the full list of implementation differences. Table 1 and all other tables have been updated to refer to this emulator as "GP emulator" rather than "Emulandice." The main text also refers to it as "GP emulator" consistently.

L205: SHAP has already been used for ISMIP6 Greenland (Rohmer et al., 2022, The Cryosphere), in more depth: for example, analysing the evolution of the most important parameters through time. This should be cited here and the results compared with in the Discussion.
Rohmer et al. (2022) has now been cited in the Results and Discussion. The Discussion section now includes a direct comparison of ISEFlow's GrIS SHAP results with those of Rohmer et al. (2022), noting that both studies identify retreat sensitivity (κ) and initialization method as key controls on projected GrIS sea-level contribution, while other characteristics such as numerical scheme have comparatively little impact.

L212-223, L259-294: These tests have very useful potential, but are poorly explained and not always well-justified.

Continued…

The sensitivity testing section (Section 3.2) and the associated tables have been revised for clarity. Table 2 now clearly describes the comparison as NN emulators trained with and without ISM characteristics (all characteristics listed in Appendix A3/A4), with a note clarifying that the "characteristics" in this context refer specifically to the full set of model-level ISM parameters listed in Tables A1 and A2, which are distinct from the three ISMIP6-wide parameters (basal melt parameterization, retreat sensitivity, and ice shelf fracture) that are included in all emulator configurations (including the sensitivity tests in Table 3). Table 3 and its caption now state which inputs are held constant across configurations (the three ISMIP6-wide parameters are included in all Table 3 configurations) and which vary (the climate forcing set). We also acknowledge in the Discussion that the small performance gap between "All" and "SMB" for the AIS is somewhat unexpected given the importance of ocean forcing to AIS dynamics, and offer a possible explanation, but include that this warrants further investigation. Regarding the categorical treatment of the continuous parameters, the updated text clarifies that the basal melt and retreat parameters are treated as categorical in our GP emulator because they take on only a small number of discrete values as sampled in ISMIP6. Continuous encoding is not used because the discrete set of values sampled in ISMIP6 does not define a meaningful continuous scale for extrapolation purposes.

L222: These sensitivity tests aren't "essential for accurately assessing emulator behaviour and performance": they assess the effect of choosing to use more or different information. One can accurately assess the behaviour and performance of a single emulator design. Rephrase around optimisation/choice/trade-offs of emulator design. And where is the counter-discussion of using too many inputs, with the risk of over-fitting?

The text has been revised to frame the sensitivity tests as assessing the trade-offs and design choices involved in selecting input variables, rather than as being "essential" for assessing performance. A sentence has been added acknowledging that using more inputs raises the risk of overfitting, and noting that ISEFlow's performance on the held-out test set and the calibration metrics in Section 3.4 provide evidence against this in the current implementation.

L239: The previous paper did not demonstrate superior performance compared with emulandice, but a GP that was even further from the emulandice design, nor did it compare with LARMIP: please rephrase to clarify.

The reference to Van Katwyk et al. (2023) has been rephrased. The text now notes that Van Katwyk et al. (2023) demonstrated that NN architectures effectively approximate ice sheet dynamics through incorporation of more input variables and explicit modeling of temporal data, without claiming it demonstrated "superior performance compared to emulandice."

L252 Have you tested this statement about fewer regions? Isn't it also likely to be the much smaller number of inputs? You would need to show training times for one (or the same number of) regions across all emulators to support this statement.
The statement about training time has been revised to attribute the GP emulator's faster training to both the smaller number of regions and the smaller number of input features, without making a stronger causal claim. The revised text reads: "The GP emulator trains faster than ISEFlow, attributable to both the smaller number of regions and the smaller number of input features in the GP design."

L297: The KS D is distinguishing between the two RCP sub-ensembles, but this is not the same as distinguishing between the two scenarios. The RCP2.6 ensemble in ISMIP6 is far smaller, sampling far fewer climate and ice sheet model uncertainties. An RCP comparison would require identical designs. I can see no subsetting of the RCP8.5 ensemble to try and do this. It's an interesting enough test to include, but the test and its interpretation need to be described much more precisely.
This is a great point, thank you for bringing this up. The updated manuscript now states that the KS D statistic compares the two RCP sub-ensembles as they exist in ISMIP6 (which have different sizes and sample different portions of the input uncertainty space), and that this test cannot isolate genuine scenario dependence from structural differences in the ensemble design. The text is now much more precise in how it characterizes what the test measures.

L305: It is at first unclear how L305-6 and 308-312, which say that the GP is similar or better at distinguishing scenarios than the NN, is then followed by precisely the opposite conclusion (L330: "the NN emulator outperforms the GP and emulandice [sic] in...scenario separation"). The difference between the two sets of tests and their interpretation needs clearer explanation, e.g. at L328 (see also next comment).
Section 3.3 now explains that these two sets of results (Tables 3 and 1) are complementary, not contradictory, and that their apparent inconsistency arises from different experimental designs. Table 3 holds inputs constant, showing that the GP achieves higher or comparable D statistics when architectures are compared with the same inputs. Table 1 compares the full ISEFlow design (all forcings + ISM characteristics) against the GP emulator as deployed following the emulandice design (GSAT only), under this comparison, ISEFlow achieves higher D for the AIS. The text discusses both comparisons and notes that the GP emulator's higher D statistics in Table 3 partly reflect its underestimation of ensemble spread rather than superior scenario resolution.

L309, L331: It would be more accurate to say that the GP always excels in the KS D statistics when ISM characteristics are not included, not just when using (local) temperature forcings, as it also out-performs the NN for the All and SMB cases in Table 3.

The text has been corrected. The revised wording notes that across all forcing configurations in Table 3 that do not include ISM characteristics, the GP emulator achieves comparable or higher D statistics than the NN, not only for the temperature-only configuration.

L309: This GP is not equivalent to emulandice: not only for the reasons described elsewhere, but also because this test is (if I understand correctly) using local SAT, not global.
Confirmed, and now clearly stated. The sensitivity test GP configurations use local (sector-level) surface air temperature, not GSAT as in emulandice. This distinction is now stated in the Table 3 caption and in the relevant section of the Results.

Figure 3: Why are only 8 ISM characteristics listed in the figure (blue bars), when 12 are named in Table A.1? Are they not all used here? Or dropped from the figure if small? Please explain.
A note has been added to the Figure 3 caption clarifying that all ISM characteristics listed in Tables A1 and A2 are included as inputs to ISEFlow, but only the top 18 features by overall SHAP importance are shown in the figure for clarity. The remaining ISM characteristics were ranked below the displayed threshold. This is now stated in both the figure caption and the main text.

L326, Figure 4: It is surprising to me that the authors do not comment on, or explain, the very different results they show for the GP here compared with the results in Edwards et al. (2021) and AR6, which show far wider uncertainty ranges in projections and far better emulation of ISMIP6.

Continued…
Thank you for this important observation. Our GP emulator performs substantially worse than published emulandice results (54.1% vs. 94.1% PICP for GrIS), which is primarily due to the implementation differences documented in Appendix C. The most consequential of these are the different training datasets (our GP uses individual ISMIP6 CMIP forcings rather than the additional simulations designed by Edwards et al. to improve coverage and performance) and different hyperparameter estimation (scikit-learn MLE vs. RobustGaSP marginal posterior mode). These differences have been acknowledged in Section 3.4 with a reference to Appendix C. Regarding Figure 4: the 5th/95th percentiles shown for all emulators (including the GP) represent the distribution of projected SLE values across projections, not uncertainty intervals from the emulator's uncertainty quantification. This is now stated in the Figure 4 caption ("the 5th/95th percentiles and medians shown represent the emulated projection distributions only, without additional sampling of emulator uncertainty"). A new Appendix (Appendix D) has been added containing emulator vs. simulator validation figures (scatter plots and error histograms).

L340: The LARMIP scenario difference in Fig. 4 will be partly due to the lack of SMB correction, as discussed earlier: this should be explained.

The Comparison with Previous Emulators and Scenario Sensitivity sections now note that LARMIP's stronger apparent scenario separation in this figure is partly attributable to the absence of an SMB correction, which, when added as in the AR6 assessment, reduces the scenario difference due to the opposing sign of dynamic and SMB contributions.

L328, Table 4: If I understand correctly – and it is not very clearly explained – the difference between tests in Tables 3 and 4 is that in the latter the GP uses GSAT (not local temperature), and ISEFlow additionally uses the other ISM characteristics, i.e. they are closer to their chosen designs. If so, it is even less surprising that ISEFlow outperforms the GP, because the differences are partly arising due to the different ensemble sub-designs (as mentioned before), and it is using much more of that information.
We agree with both points. The relationship between Tables 3 and 1 (formerly Table 4) is now clearly explained in the text, and Table 4 (the scenario comparison table) has been merged into Table 1, adding a KS D Stat column. This consolidation reduces redundancy and makes the relationship between projection accuracy and scenario sensitivity easier to assess at a glance.

L345: Can you explain or suggest why using only temperature forcing best explains the scenario difference (Figure 5), when it performs worse than using only SMB in Table 3?
This apparent inconsistency has been addressed in the text. The classification model (Figure 5) asks which variables best distinguish the emission scenario label itself, a question where temperature excels because it is the most direct expression of the forcing difference between RCP scenarios. The emulation model (Table 3) asks which variables best predict ice sheet response, a question where SMB performs well because it integrates precipitation, melt, and runoff into a compact summary of the climate state most directly relevant to surface mass changes. The two analyses are measuring different things, and the combination of both provides a more complete understanding of the role of each variable than either alone. This is now explained in the revised Section 3.3.

L358, Table 5: is this GSAT? If not, i.e. if it is temperature over WAIS/EAIS/Peninsula, then it is not the same dataset as "emulandice" (aside from the other differences previously mentioned).
The Regional ISEFlow uncertainty comparison (Table 5) uses local surface air temperature (averaged over the three AIS regions and the single GrIS region), not GSAT. This is now clearly stated in both the table caption and the Methods text. The distinction from emulandice's use of GSAT is also noted, as part of the broader set of implementation differences documented in Appendix C.

L375: How does Regional ISEFlow perform better on PICP for AIS? The value is lower, not higher (assuming the target is 95%, though this is not defined – see Technical corrections).
The PICP target of 95% has been added to the Table 5 caption and explained in the Section 3.4 text. Regarding the AIS comparison: Regional ISEFlow achieves PICP of 0.826, while the GP

emulator achieves 0.918. The GP emulator's PICP is closer to the 95% target for the AIS. The updated text in Section 3.4 now accurately states that "Regional ISEFlow achieves higher PICP for the GrIS but lower PICP for the AIS relative to the GP emulator," which is the correct characterization.

L377, Table 5: The authors do not comment on the fact their estimate of PICP for the Greenland GP is very different to that shown in Edwards et al. (2021): a rather miserable 54.1% coverage, compared with 94.1% in our study.
This discrepancy is now explicitly acknowledged in Section 3.4 and attributed to the implementation differences documented in Appendix C, primarily the different training datasets and GP implementation.

L378: How does Regional ISEFlow perform better on CRPS than the GP, when both of its scores are higher? A typo? Are the two rows exchanged?
For CRPS, lower values indicate better probabilistic performance. The table rows were correct; the error was in the text, which stated "ISEFlow is more accurate (higher PICP)" but made an incorrect claim about CRPS. The text now correctly states that "Regional ISEFlow achieves higher PICP for the GrIS (0.858 vs. 0.541)" and that "CRPS values are close between Regional ISEFlow and the GP emulator, with Regional ISEFlow slightly worse for the AIS (0.463 vs. 0.371) but comparable for GrIS."

L383: How does a higher PICP for ISEFlow indicate sharpness? 99% coverage is pretty high – I have no problem with that as a result (again, assuming 95% is the target), but it's a bit of a stretch to call it sharp.
The reference to "sharpness" in connection with high PICP has been removed. The text notes that ISEFlow-AIS achieves 98.6% PICP, which is above the 95% target; while this reflects good coverage, it also indicates intervals that are somewhat conservative (wider than strictly necessary). The Winkler Score is the metric that best captures the overall balance between coverage and sharpness, and ISEFlow's lower Winkler Score relative to the GP emulator reflects that the additional coverage comes at an acceptable cost in interval width.

L375-386: I would describe these larger uncertainties as an advantage, not a disadvantage – see General comments – because I think it's much more important to try and account for some of the under-estimation of uncertainty about the real world by ISMIP6, rather than having the sharpest possible prediction of each individual simulation.
The text in Section 3.4 now notes that larger uncertainty intervals, while penalized by accuracy-sharpness metrics relative to the 95% PICP target, may be viewed as an advantage in the context of sea level projections, where the ISMIP6 ensemble is known to underestimate real-world uncertainty. We have softened the language framing wider intervals as a "trade-off" and instead note this as a contextually appropriate feature.

L390: Presumably LARMIP is included in this conclusion (e.g. Table 1) and should be cited too?
L404-6: And the fact you are using three times more regions in AIS than GIS…?
LARMIP has been added alongside emulandice in the Discussion and Conclusions references to the comparison set. The difference in number of regions (18 for AIS, 6 for GrIS) is noted in the Discussion where spatial information is discussed as a driver of the AIS emulator's results.

L415: Making projections for intermediate scenarios appears to be stated as a novel purpose, when it was the primary motivation for developing emulandice (as you say later) and is also a capability of LARMIP. Please clarify – fine to justify this with the higher accuracy, spatial resolution etc – but if going down the route of potential future projections, then some of the questions, issues and trade-offs described in General comments need to be discussed.
The Discussion has been revised to acknowledge that filling in intermediate scenarios was a primary motivation for developing emulandice and is also a capability of LARMIP. ISEFlow's potential contribution in this area is framed as a complementary tool offering higher spatial resolution and accuracy within the ISMIP6 ensemble, while acknowledging the additional steps required before it could be used for policy-related projections.

L442: I'm not sure it's fair to say anyone "assumed" that GSAT alone is sufficient for "accurately" capturing ice sheet responses – we did say in Edwards et al. (2021) that "using regional climate variables would improve the signal-to-noise ratio for the emulator". It would be more accurate to say that different choices have different trade-offs, or might be made for pragmatic reasons (like time and complexity, especially when being a somewhat new approach in the field). This study is an important advance in emulating multi-model ice sheet projections: it doesn't need to dismiss everything that came before it to justify itself.
We appreciate this feedback. The text in the Discussion has been revised to acknowledge that the choice of GSAT reflects deliberate trade-offs rather than an assumption of sufficiency, and to present the two approaches as complementary tools with different strengths and limitations. The Discussion makes clear that ISEFlow is not intended to replace GSAT-based emulators but to add a complementary capability.

L444: I'm not sure it's a good idea to make such a sweeping statement about providing insight to other fields, with so few references, when there is long and broad history of statistical emulation and uncertainty quantification for complex models.
The broad claim about insights to other fields has been revised to be more modest and specific. The revised text briefly notes that a similar emphasis on multi-variable input sets and hybrid uncertainty quantification has been explored in emulators for glaciers, weather, and ocean circulation models, with specific citations, rather than making a general methodological claim about the field.

L467: ISEFlow does partially account for model uncertainties, by emulating a multi-model ensemble forced by multiple GCMs. On the other hand, it is essentially impossible to fully "address these limitations": we can only aspire to keep improving our assessments of our lack of knowledge.
The limitations paragraph has been revised accordingly. The text now acknowledges that ISEFlow partially accounts for structural model uncertainties through its multi-model, multi-forcing training, while noting that complete characterization of these uncertainties remains aspirational.

Technical corrections
L5, L127, L161: Expand acronym LSTM (in all three places) – needed for this journal audience.
L46: IPCC 6th -> Sixth
LSTM has been expanded to "Long Short-Term Memory (LSTM)" at all first-use locations. We leave LSTM as an acronym throughout the rest of the manuscript, as we believe that it is more commonly known by its acronym than the spelled-out version. "6th Assessment Report" has been changed to "Sixth Assessment Report" throughout.

L47: To me, land ice includes global glaciers. I would rephrase, e.g. "ice sheet evolution and sea level projections".
Corrected. Any instance of “land ice” has been changed.

L48: Small point, but emulandice is lower case (see GitHub and also Kopp et al., 2023, GMD).
Corrected throughout. "Emulandice" has been changed to "emulandice" at all occurrences (besides beginning of sentence).

L50: An assessment, not a comprehensive view. Comprehensive is an aspiration! Also cite Kopp et al., 2023, GMD, describing the use of FACTS in AR6. Add "to 2100" and delete second "future".
L52: I think model uncertainties/inputs/choices would be a clearer choice than "data".
L54, L71 etc: I would say "ice dynamics" is much more than changes in sea level contribution – emulandice doesn't begin to aspire to represent e.g. spatial patterns of ice velocity, ice shelves – how about ice sheet change or evolution, or just end at ice sheets?
L55: Add ice shelf collapse.
All of these changes have been made.

L67, 224 etc: Emissions scenarios are not just carbon – delete this word throughout.
The word "carbon" has been removed from all occurrences of "carbon emissions scenarios" throughout the manuscript.

L94: Split or otherwise rephrase this sentence, as it currently sounds like the forcings are also deviations from the control experiments.
The revised text now states "The key emulator prediction target is the Sea Level Equivalent (SLE) contribution...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."

L131, L143, L158, L481, Figure 1 caption: data are.
Corrected throughout. "Data is" has been changed to "data are" at all noted locations.

L169: I would add "and models" to "these systems".
L183: Describe more precisely: using GSAT from the CMIP5 GCMs used by ISMIP6 (Seroussi et al.). They are not ISMIP6 temperatures, and AIS is confusing because it implies a spatial average.
L192-3: Global temperature change, and I would add 'parameter' after melt.
L233: Sentence has gone wrong.
All of the mentioned changes have been made. The last was a typo ("that was help out") that has been fixed to "that was held out."

L244 It would clearer for this audience to use "prediction", not "inference", throughout. L275: consistent with
"Inference" has been changed to "prediction" throughout, and "consistent to" has been changed to "consistent with."

L286: Presumably this is SAT averaged over each ice sheet basin? I would remind the reader in case they think of GSAT.
The text now refers to "local surface air temperature (averaged over each ice sheet sector), distinct from GSAT" to avoid this ambiguity.

L289-316: This needs restructuring. I would discuss MSE first, then KS D. It's confusing to summarise both results at L291 before explaining the metrics. Then end with the monthly/daily temperature statement – or delete it, as aren't ice sheet models forced with annual means?
This entire section has been substantially revised based on your comments as well as other reviewers'. The paragraph has been restructured to discuss MSE results first, then KS D. The monthly/daily temperature sentence has been removed, as the ISMIP6 ice sheet models are indeed forced with annually averaged forcings and this point is not supported by the current analysis.

L297: Rephrase "projection accuracy", as in the general point about model land.

"Projection accuracy" has been changed to "emulation accuracy" at this location and throughout the section.

L364: Presumably 95% intervals for the PICP? Define. How are the WAIS, EAIS and Peninsula results combined – averaging? (This probably needs clarifying in other parts of the manuscript).
The target coverage probability (95%) is now defined in the Table 5 caption and in the Section 3.4 text. The AIS PICP and MPIW values are computed by pooling all sector-level predictions for the three regions (East Antarctica, West Antarctica, Antarctic Peninsula) rather than averaging region-level PICP values, and this is now stated explicitly.

Table A.1,2: Table A.2 is missing the GIS retreat parameter. I would also highlight the three ISMIP6-wide parameters in A.1 and A.2 (and Fig. 3) in bold, and use consistent phrasing (collapse or fracture, not a mixture of the two). The PIGL parameterisation is an acronym for "Pine Island Grounding Line": it does not use any information from the Larsen ice shelf. Please also state whether only using the PIGL median value, or also the 5th and 95th percentiles.
The GrIS characteristics table (Table A2) has been updated to include the Ocean Sensitivity (retreat parameter κ) and Ocean Forcing (open vs. standard framework) characteristics, which were missing. A clarifying comment has been added about the three ISMIP6-wide parameters in the text. Consistent terminology ("ice shelf fracture" rather than "ice shelf collapse") is used throughout. The description of Ocean Sensitivity in the AIS table has been corrected, it now refers to "Low, medium, and high ocean sensitivity corresponding to the 5th, 50th, and 95th percentile values of the 'MeantAnt' distribution (Jourdain et al., 2020)". For GrIS, the retreat sensitivity values used (25th, 50th, and 75th percentile of κ, as defined in Slater et al., 2020 and Goelzer et al., 2020) are now specified.

General – for the AIS tests where LARMIP is not included, it would be helpful to explain why not.
A brief note has been added to the Table 3 caption (forcing sensitivity ablation) and to the Table 5 caption (UQ comparison) explaining that LARMIP is not included in these sensitivity analyses because it uses a fundamentally different methodology (linear response functions) and accepts only a single input (GSAT), which does not lend itself to the same input variable sensitivity analysis performed here. For the UQ comparison, LARMIP does not provide calibrated uncertainty intervals and therefore cannot be evaluated using the same coverage-based metrics.