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

General comments

This study examines the performance and results of a new machine learning emulation method applied to the ISMIP6 multi-model projections for the Greenland and Antarctic ice sheets to 2100. This kind of study is very important – independent methods in this area are much needed for comparison with the few existing other methods – especially probabilistic approaches that estimate uncertainties of the emulator, that can cope with big data (more simulations, and using more information from those simulations), and that aim to explain and improve understanding of the most important uncertainties in the underlying models and the emulator itself. The study yields a useful new emulation approach that performs very well, and provides information about driving processes/uncertainties, as well as highly useful information quantifying the benefit of using additional information from ice sheet model choices and climate forcings in the emulators. The methods are mostly sound and the validation thorough, with some exceptions described below, and the study represents an impressive piece of work.

There are, though, various problems that need to be addressed. The study has multiple errors in the implementation and description of emulandice, at least one of which was mentioned in my review of the original submission, which substantially change the authors’ conclusions relating to emulandice itself. Some technical aspects might be straightforward to address (e.g. emulator structure) and others less so (e.g. training on the original dataset). Either way, the method should still be referred to as a GP using many/most of the emulandice choices, as in the lead author’s previous study: it is not the emulandice code, and misleading to describe it as such. It is also unclear if the GPs’ uncertainties are included in all predictions, or only used in the coverage tests. Several results in validation (Table 5) and prediction (Figure 4) seem to be completely different to emulandice, and the authors do not note or explain why or consider whether this could indicate a problem with their implementation.

There are other problems: the text has other scientific errors, appears to contradict their quantitative results in some places, and makes some claims that are at best confusing, or at worst poorly-evidenced or contradictory – particularly around scenario-dependence.

Finally, the study lacks sufficient perspective to situate the work in the wider field when comparing with previous studies and discussing the future potential of ISEFlow. This new emulator is indeed a powerful and useful tool, which outperforms the GP and LARMIP on several fronts, but that does not in and of itself mean it is a suitable and desirable replacement for making equivalent sea level projections for policymakers. The manuscript gives a very thorough description of validation and sensitivity analysis (SA) – the first stages in building and trusting an emulator to make real world projections – but it only makes predictions under the original input choices, and does not fully consider how to make projections that incorporate a much wider sampling and inclusion of uncertainty information and expert judgement than the ISMIP6 ensemble, i.e. the usual purpose when building emulators for quantifying model uncertainties. When discussing future potential, it discusses little of the context that is critical for using emulators to answer broad and complex questions such as projections for the IPCC: why certain choices were made in the previous emulator designs that were judged desirable for making sea level projections for policymakers, even if they sacrificed some aspects of performance in reproducing simulation results (i.e. validation metrics). This narrow view – that validation and SA are the only measures of success and suitability – is very understandable, as it is the norm in machine learning. But real world sea level projections using emulators do require other considerations, such as:

1. How should we choose, derive and defend prior probability distributions for the uncertain inputs when making real world projections?
2. How can we try and account for the fact that the uncertainties of the ISMIP6 ensemble itself are underestimated?
3. How can we combine our ice sheet projections with those for other contributions to sea level change, with all relevant 'book-keeping' and uncertainty propagation?
4. How can we incorporate the correlation of responses between the ice sheets induced by global climate change?
5. How can we update our projections using new emissions scenarios, such as revised climate policies?

Several of these questions are at the core of the statistical field of uncertainty quantification for complex models (e.g. Kennedy & O'Hagan, 2001), but are still too rarely considered in the machine learning field. This foundational problem is encapsulated by the repeated and (apologies, but) absurd claim that ISEflow is far more *accurate* at *projecting future sea level* changes – not at reproducing the simulations of ISMIP6. This is not purely semantics, but represents a viewpoint that Erica Thompson has described as living in 'Model Land', which could perhaps be adjusted here to 'Machine Learning Land'.

The manuscript should therefore substantially revise the interpetation and discussion of how ISEFlow could be used in future, and the implications of using different choices compared with previous work. There may well be potential to use it for the purposes of real world projections for decision-making, as discussed below, but this is beyond the scope of the current study and tool. The necessary trade-offs between precision and practicality should also be more clearly communicated when describing the previous emulators, to avoid a biased representation of their choices.

Overall, then, the manuscript requires major revision of claims and interpretation, and potentially also of the GP methods. But I do strongly support the overall intent and methods of the study – and agree with the majority of the Discussion section, which is mostly more considered about possible future uses – and I hope the authors find these (very) detailed comments useful to strengthen and clarify the manuscript.

Tamsin Edwards

Specific comments

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.

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.

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.

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:

a) to include correlation of land ice regions induced by global climate change;
b) 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;
c) 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;
d) to rapidly and easily update projections under any emissions scenario (e.g. current NDCs, new scenarios).

This is relevant when considering ISEFlow's choice to use regional climate forcings, and how this limits or alters its potential use in sea level projections in decision-making. If we were to use CMIP models as the prior for ISEFlow sea level projections, as proposed in the Discussion, then this implicitly includes (a), but precludes (b), and limits (c,d). For (c), the CMIP ensemble is not technically defensible as a prior on our uncertainty about future climate (for example, the IPCC inflates the ensemble variance for likelihood assessments), even if all GCMs saved all available climate forcings used by ISMIP6, which they may not; of course, the fact it is an ensemble of physical models does give it a different advantage over FaIR trained on these models. For (d), ISEFlow would have to wait for new GCM simulations (e.g. to run AR7 scenarios), just as ISMIP itself does, which can be a major disadvantage in itself. It would also limit the emissions scenarios that can be used, and the prior sample size for each (e.g. some scenarios are run by far few GCMs). This does not even consider the point that the Greenland SMB forcings were taken from one Regional Climate Model – not mentioned in the manuscript – for which there is a far more limited availability of simulations, and even fewer from other RCMs. For a trade-off in model simplicity and accuracy, FaIR rapidly provides probabilistic (large sample) GSAT projections for any emissions scenario, including updated NDCs and net zero targets, using assessed climate sensitivity distributions.

If instead we were to statistically supplement CMIP6 (or use a purely statistical method) to generate large sample probabilistic prior projections on regional climate fields at the basin level evolving through time, then the trade-offs and requirements would be: for (a), to ignore or explicitly model the correlations between polar climate variables induced by global climate change; for (b), to statistically model the relationship between global and regional climate (as LARMIP does for basal melt); for (c), to ignore or statistically model the correlations between each climate variable and each basin, if present. (For (d), I think it would be a similar case as using GSAT: i.e. assuming all relationships hold for new scenario trajectories.). All are possible: but substantial statistical analyses, that are well outside the scope of this study.

These trade-offs would be worth exploring in future, and this kind of emulator is absolutely useful for making comparisons of using GSAT vs regional climate: but it is not equivalent in application. By only describing the advantages of ISEFlow, and not the disadvantages, the manuscript gives a biased view when comparing emulator designs: broadly equivalent to stating that a high resolution earth system model is simply "better" at predicting global mean temperature than a climate model of intermediate or low complexity, while never discussing why the latter might be developed or used. Why do we use FaIR, MAGICC and HECTOR for making some of the policy-relevant projections in the IPCC, and not just CMIP or even HighResMIP...? For some of the same reasons as choosing to use a simpler emulator of sea level components over a more complex one.

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.

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.

I would add another table for the climate variables, adding the spatial and temporal averaging period for each, and whether they are derived from GCMs or RCMs, and ideally adding rows for the later sensitivity tests (which combinations of, or additional, climate variables are used, e.g.

"All" in Table 4). The ocean forcings are not averaged over the ice sheet basins shown in Figure A1/A2, so this needs a brief mention of the boundary extensions, as well as the (presumable) depth-averaging and/or depth selection.

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.

L101: Basins are not "the maximum amount of spatial information possible", grid cells are.

L112: ISEFlow could predict the results of new combinations of existing models and inputs, but not the results from new model updates or parameters.

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.

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.

L146: And also because of the structural difference between models – many combinations of models and inputs/choices do not physically exist.

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.

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.

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.

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).

L188: You do not implement or describe the emulandice architecture correctly:

- As mentioned in my review of the initial submission, emulandice also includes a categorical variable for open vs standard retreat parameterisation for Greenland. Also, do you use the same method to impute the retreat parameter for the open models? It's not stated.
- You describe emulandice as being trained on FaIR projections (e.g. L195, L484-486), but these are only used for *predictions*. The description of the CMIP5 and CMIP6 simulations used for training emulandice is in the Methods section named "Global climate model simulations", and they can be found in the climate forcing CSV files on my GitHub as rows marked CMIP5/6 in the first column.
- From memory, we did not have information on the exact GCM variant used by ISMIP6 at the time, which is why we used the mean of all realisations: note that this smooths the GSAT time series, reducing noise compared with the single simulation method you use, and this might affect the GP's performance.
- As noted above, you also do not train the GP on the additional simulations that were designed to improve emulandice's performance, or the SSPs, and appear to add many new simulations: such large changes in datasets might have led to choosing a different design, such as the covariance structure, when validating the GPs. Gaussian Processes, like all non-parametric and machine learning methods, are – or should be – carefully checked and validated for the datasets they use: it is not a like-for-like comparison to train the same method on very different data and expect it to give the same result or perform as well, especially on such a sparse and unevenly designed dataset as ISMIP6. Fine to test this combination, but again: it is not the same GP as in the emulandice code.
- There may also be differences when using RobustGaSP, which uses marginal posterior mode estimation, vs the scikit-learn GP package implemented here, which uses the more usual Maximum Likelihood Estimation.

Please clearly document these differences, here and in summary elsewhere (e.g. Table 1 caption), and revise the description to simply "Gaussian Process / GP" throughout the paper (including tables etc), to avoid confusion.

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.

L212-223, L259-294: These tests have very useful potential, but are poorly explained and not always well-justified. Table 2 and L214-5, L262-271 appear to describe tests comparing using climate forcings only vs climate forcings and ISM inputs – is that right? Is this useful? Is it a comparison aimed at LARMIP? It doesn't help understand the limitations of emulandice, which would need a comparison of using the ISMIP6-wide parameters (i.e. retreat/melt + collapse) with all ice sheet model inputs (this would be interesting). Or perhaps – see Table 3 discussion below – you do include the 3 ISMIP6-wide parameters in both? You should also remove the citation to Edwards et al. (2021) at L215, because it is not accurate, as emulandice of course does not use climate forcings "in isolation" for either ice sheet. The tests in Figure 2 are very useful, but then it's unclear what Table 3 shows: the text (L285-7) says that ice sheet model parameters are used in these sensitivity tests, but the Table 3 caption seems to say they are not, then that they are – after a few readings I realised you don't count the three ISMIP6-wide parameters as "ISM characteristics" – this needs more careful explanation and terminology. Related: basal melt and collapse are listed within the AIS charactistics Table A.1, but retreat is not listed in the GIS Table A.2. It's also not clear why Table 3 says the three ISMIP6-wide parameters have categorical representations, when 2 of the 3 are continuous (and in fact collapse is essentially treated as a continuous fraction in emulandice too) – if true, e.g. if ISEFlow does treat them all as categorical (e.g. using the discrete values sampled by ISMIP6), this point isn't explained in the text, and is presumably not the case for the GP anyway. Once explained more clearly, this will be a very useful assessment of how much using additional climate information improves emulation. I'm surprised that ocean forcing doesn't improve "All" for AIS more, over "SMB", especially for ISEFlow – can you explain or suggest why? Or add an ocean-forcing only test for each? It would also be very interesting to add a comparison using GSAT, not local SAT, if possible – and see comment below.

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?

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.

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.

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.

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).

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.

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.

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.

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. For the projections, there would be multiple reasons for this: potentially the inclusion of the GP uncertainties (they may be included here too, but if so the sampling method is not described) and, of secondary importance to this figure but still important for general interpretation later, the priors on climate forcing and ice sheet model inputs, which broaden the sea level projections to capture much more uncertainty. The authors should clarify whether the ISEFlow and GP emulator uncertainties are included in Figure 4 (and elsewhere), as they should be for a fair and relevant comparison.

As these projections are only for the ISMIP6 inputs, a closer comparison can be made with the validation aspects of Edwards et al. (2021). If the GIS results in this study's Figure 4 are correct, then our ED Figure 1 would show an extremely flat set of results, where the emulator drastically under-estimates both low and high extremes in the ensemble. It does not: so why does this GP? Implementation differences, or a bug? Why are these basic comparisons not made as a sanity check? For the 3 AIS sectors, our validations are somewhat flat at the top end, from underestimating the highest ~1% SL contribution simulations – mainly SICOPOLIS, under extreme high basal melt values (see 'Evaluating the emulators'). But there is not the drastic underestimation of ensemble spread visible in this study's Figure 4. Why?

As well as these discrepancies with emulandice, the GP gives only a 20% smaller difference between the RCP medians for AIS than ISEFlow (16% smaller for GIS). I may be biased, but this difference seems rather over-stated in the manuscript (e.g. "unable" to distinguish, Fig. 4 caption; "difficulty differentiating", L326, compared with statements of ISEFlow's success), especially given the KS D results in Table 3, which show a similar ability to the NN under reduced inputs (though note my caveat that neither test solely evaluates scenario difference, because they are convolved with the very different ISMIP6 designs).

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.

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. Again: it is a useful test, but it does not only test scenario distinction. Please clarify this interpretation. It would make sense to merge Table 4 into Table 1: one column is already identical (MSE mean), and the KS D could go alongside the KLD and JSD distribution metrics. Then Table 3 would be a simpler discussion of the forcing tests.

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? It would also be interesting to see the ISM inputs added here (as in Fig. 3), if possible, but I understand if this is beyond scope.

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).

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).

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. They are also slightly lower for AIS (91.8%, compared with 92.4-94.8% across the three sectors in our study). My first suspicion would be the absence of one of the emulandice GIS inputs (open vs standard retreat parameterisation); my second would be the more restricted datasets for both ice sheets, without the extra simulations to improve the emulator, and adding others that might have changed the emulator design choices. Or a bug? This undermines the conclusions of the paper regarding emulandice and its AR6 projections: this could be rectified without changing the analysis by clarifying the differences between this GP and emulandice (assuming not a bug), and making it clear the comparisons are with this specific GP. Validation figures such as the standard emulator vs simulator plot or a histogram of standardised errors (also shown in Edwards et al. 2021, ED Figure 1) would also help to confirm these results, as well as check for any systematic biases: is there any reason these standard figures are not shown?

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?

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.

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.

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…?

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.

L442: I'm not sure it's fair to say anyone "assumed" that GSAT alone is sufficient for "accurately" capturing ice sheet reponses – 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.

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.

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.

I (heartily!) agree with the rest of the discussion, which is a much more sound description of the strengths and capabilities of ISEFlow. There are some very interesting possibilities here, in particular from using the additional climate forcings: the point about early prediction of ISMIP7 from CMIP7 is an excellent one. This is a different problem to generating probabilistic policy-relevant projections, but would be a valuable contribution to this effort by potentially informing the design and prioritisation of GCMs and scenarios as soon as each CMIP7 simulation is completed. As you say in the Conclusions, training ISEFlow on ISMIP6 simulations to 2300 would potentially allow ISMIP6-style projections to be made for all CMIP6 climate projections to 2300, to inform the choice of GCMs for ISMIP7. Using the extra information from ISM characteristics may be a bit more complex or challenging to complete and interpret, given that models may change each generation and can take time to document their methods, but there is good potential for input to design here too.

(But please do revise the conclusions in line with the above comments, particularly around accurate sea level projections, scenario distinction, and description of emulandice).

Technical corrections

L5, L127, L161: Expand acronym LSTM (in all three places) – needed for this journal audience.

L46: IPCC 6th - > Sixth

L47: To me, land ice includes global glaciers. I would rephrase, e.g. "ice sheet evolution and sea level projections".

L48: Small point, but emulandice is lower case (see GitHub and also Kopp et al., 2023, GMD).

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.

L67, 224 etc: Emissions scenarios are not just carbon – delete this word throughout.

L94: Split or otherwise rephrase this sentence, as it currently sounds like the forcings are also deviations from the control experiments.

L131, L143, L158, L481, Figure 1 caption: data are.

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.

L244 It would clearer for this audience to use “prediction”, not “inference”, throughout.

L275: consistent with

L286: Presumably this is SAT averaged over each ice sheet basin? I would remind the reader in case they think of GSAT.

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?

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

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).

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.

General – for the AIS tests where LARMIP is not included, it would be helpful to explain why not.