We sincerely appreciate the editor and reviewers for taking the time to review our manuscript. We carefully reviewed all the comments and suggestions and made revisions based on their comments. We would like to emphasize several major revisions we made to improve the quality of our manuscript.

(1) **Remove CNN (FCN) comparisons:** We agree with the editor and all the reviewers that it is not a fair comparison to use FCN as a baseline emulator because it has a different nature from GNN. Therefore, in the revised manuscript, we remove the comparison with FCN and only include the comparisons among GNNs (i.e., GCN, GAT, and EGCN). We focus more on "using computationally efficient GNN emulators to find optimal parameterization for computationally heavy numerical models" rather than "showing the superiority of GNN over FCN as the emulator of finite-element modeling."

(2) **Adjust the input & output variables for better generalizability:** The reviewers are concerned about the general applicability of the previous GNN emulators. To address this issue, we changed the input and output variables and re-trained the GNN emulators to ensure better generalizability. First, now our emulators predict the next time (t) ice conditions *purely* from the previous (t-1) ice conditions; all the input variables are from the t-1 time period. Second, instead of using the level set ice mask as the direct output variable, our emulator now predicts *tensile stress (sigma)* as the output variable. Then, the level set ice mask is derived from this tensile stress condition with a certain σmax value using equations 1, 2, and 3 (Fig. 2). This allows us to (i) improve the generalizability of emulators in any ice conditions, time frame, or σmax scenarios and (ii) facilitate finding the physical meaning of σmax in calving modeling.

(3) **Add comprehensive sensitivity tests:** The reviewers have concerns about the robustness of the emulator for various σmax cases because we only selected two σmax as test datasets and used the others as training datasets. To test the robustness of our emulators depending on the selection of different training samples (i.e., different σmax as training datasets), we conduct four sensitivity tests. We divide the 9 σmax simulation data into training and test datasets in four different ways (please see Table 1): (A) even-value samples (0.75, 0.90, 1.05 MPa σmax as training dataset), (B) low-value samples (0.70, 0.75, 0.80 MPa σmax as training dataset), (C) medium-value samples (0.85, 0.90, 0.95 MPa σmax as training dataset), (D) high-value samples (1.00, 1.05, 1.10 MPa σmax as training dataset). By evaluating the model performance for these four different training sets, we check how the model performance is sensitive to the selection of training samples. The sensitivity test results show that the emulators have less than 10-20 % uncertainty in the selection of training samples (Fig. 3 and Appendix A), which supports that the emulators trained with a certain set of σmax can represent the general cases for all σmax values.

(4) **Add discussions about the physical implications of σmax:** The optimal σmax value we obtained from the updated GNN emulator agrees well with the retreat and advance of the calving front (Fig. 6). When we adopt this temporally varying σmax in ISSM numerical

simulations, this simulation result shows better agreement with observations compared to the constant σmax setting, which has been used in most previous studies. Therefore, we highlight the need for fine-tuning of σmax to improve the accuracy of glacier mass loss prediction. Additionally, in section 6, we add more discussions about how to connect this σmax setting with external environmental conditions, such as mèlange, ocean tide, ocean circulation.

Please see our detailed responses to the reviewers' comments below. The first-round comments are highlighted in green, the second-round comments are in black, and our responses are highlighted in red. All line numbers are from the clean version manuscript.

Reviewer 1

L269 'out of sample' is perhaps a bit of an overstatement- the degree of correlation between neighboring σ values would very likely be quite high- as a check, it would be interesting to see what error is induced by comparing model predictions made using σmax = 0.8 to the withheld test set values for σmax = 0.75 (or something like that). I expect the metrics would be similar because there is little difference between such small variations in the parameter. A more useful test would be to train on just the two extremal values (0.7,1.1) and see if it can still interpolate well.

This reviewer comment was similar to mine, and I'm not sure that the experiments added in the appendix address this. What would be nice to see (and what I think this reviewer suggested) was to re-train the GNN with just 0.7 and 1.1 values and determine how well the resulting GNN performs at predicting the rest of the simulations conducted. More broadly, a more comprehensive set of sensitivity experiments (randomly leaving two out of training then testing on those two, also randomly leaving more and more out of training to see how the performance metrics degrade as a function of the size of the training set). I'm afraid that without these experiments included, it would be challenging to know whether this method truly is robust or simply is memorizing the training set in an idiosyncratic way. I wouldn't be able to endorse publication without seeing these extra experiments.

- Thank you for your comments. We agree that a comprehensive sensitivity test is necessary to argue the robustness of our emulator. In the revised version, we add a set of sensitivity tests to check if the emulator works well with any σmax threshold settings. We divide the 9 σmax simulation data into training and test datasets in four different ways (please see Table 1): (A) even-value samples (0.75, 0.90, 1.05 MPa σmax as training dataset), (B) low-value samples (0.70, 0.75, 0.80 MPa σmax as training dataset), (C) medium-value samples (0.85, 0.90, 0.95 MPa σmax as training dataset), (D) high-value samples (1.00, 1.05, 1.10 MPa σmax as training dataset). By evaluating the model performance for these four different training set settings, we would like to check how the

model performance is sensitive to the selection of training samples. The sensitivity test results show that the emulators have less than 10-20 % uncertainty in the selection of training samples (Fig. 3 and Appendix A), which supports that the emulators trained with a certain set of σmax can represent the general cases for all σmax values.

- Please see section 4.4 for the details of this sensitivity test setting and section 5.1 for the results.

Sec. 5.2 It is frankly absurd to not include even a mention of the computational cost associated with increasing the training data, which- so far as I can tell- must be repeated for any new geometry or parameter or location. Ignoring this cost does of course lead to much more impressive speed-ups, but these are not real. I would expect this problem to become considerably more severe when trying to use this technique to emulate models that are a function of more than a single parameter- the curse of dimensionality still applies. Furthermore, the notion that a GNN will be more suitable than a CNN for higher resolution modelling is another strawman because it ignores the fact that generating the cost of generating the training data (which is presumably more expensive than any network training) is also going to scale proportionally with resolution. I would encourage the authors to revisit this entire section with a more sober perspective aimed at delivering a factual assessment of the present work's utility.

The point is that the workload is running ISSM. This is required before you can train the NN emulator. I think the point the reviewer is making here is that this may not be hard to do for a handful of σmax values for a single glacier, but would be very challenging to do computationally for a large region or the whole ice sheet. So, giving an accurate accounting of the costs of running ISSM to generate training data is important to defining for readers how computationally expensive it is to apply this method to a new setting.

- Thank you for clarifying the previous comments, and we admit that the workload of running ISSM to train the GNN emulators is significant. In our case, it takes 948 seconds to run one ISSM transient simulation with a single σmax value (i.e., ~8,500 seconds for all 9 σmax values). According to our new sensitivity test (Fig. 3), three ISSM simulations (from three different σmax values) are enough to train the GNN and achieve a certain model fidelity, especially in the even-value sampling (set A). Hence, the time for simulation data collection can be less than 8,500 seconds if appropriate σmax values are selected as training data. We add a mention about this in L343-348:

  "It should also be noted that collecting various ISSM numerical simulations with various σmax values is required to train GNN emulators: in our case, approximately 8,500 seconds are required to collect all 9 simulation results with different σmax values. Nevertheless, given that the emulators show sufficient performances only with 3 training σmax values, fewer simulations would be sufficient for training. Furthermore, we emphasize that these GNN emulators can be efficiently applied to this target glacier with

a short implementation time once they are trained, which can accelerate the search for optimal parameterization settings that align the numerical model with observations."

- Above all these, we would like to highlight the computational efficiency of GNN emulators in "finding optimal (temporally varying) σmax calibration." Given that it takes ~8,500 seconds to collect transient simulations with 9 σmax values, it would be extremely time-consuming to find the temporally varying σmax values using numerical simulations. However, we can save a lot of time in this procedure by using 30-34 times faster GNN emulators. Additionally, whereas the ISSM simulations are implemented on super computing clusters, our GNN experiments can be easily and quickly implemented on local machines. We add a mention about this in L391-395:

  "Moreover, the use of GNN emulators results in a dramatic increase in computational efficiency, with speeds 30-34 times faster than the ISSM simulations. This speed-up allows us to find the temporally varying optimal σmax values quickly. Given that running numerical simulations every time to determine the optimal σmax setting is extremely time-consuming, our fast GNN emulators reduce the workload associated with calving calibration."

The accuracy metrics and differences therein are not very convincing. Interpreting a difference between 0.997 R value and 0.999 is not good statistics, particularly without assessing significance of these statistics on the training data. Similarly, I'm not sure how different a calving front accuracy of 98.6% vs. 99.4% is. I'm guessing both are significant at some very high level and so reading much into the difference beyond that isn't very meaningful. What happens if you drop some of the training data? Does the accuracy degrade? This is a common way to determine whether the NN has learned anything about the underlying dynamics of the system vs. acting as a fancy interpolator of the training data.

Moving to this less-inflated metric is useful, but I think the points still stand that comparing R values is not statistically meaningful without a sense for the significance. Statistical inference is a way for answering a binary questions, usually whether two samples are significantly different. It is useful that you have shown that one number is smaller than another, but there is a possibility that this could happen by chance. Significance testing allows you to rule out that this is by chance alone. If you would like to be able to conclude, as you do, that one NN performs significantly better than the other, you need to show that the performance metrics you choose are significantly different.

- Thank you for your comments. We understand that the reviewers still have concerns about R values. We removed R values in evaluating the model fidelity and only use RMSEs as the evaluating metrics.
- In evaluating model fidelity (section 5.1 and Appendix A), we calculate statistical significance to check (i) if a certain model is *significantly* different from the others in

terms of RMSE and (ii) if a test RMSE is *significantly* different from a train RMSE (see Fig. 3, Table A1, and Table A2).

L243: Related to point #2 above - it seems that you have chosen test cases non-randomly, and I wonder what would happen if you chose 0.7 as a test case instead (with 0.7 not in the training/validation data)?

See point above about this. If two different data sets (the NN outputs here) are significantly correlated to a benchmark dataset (the ISSM outputs), then the conclusion is that neither can be statistically distinguished from the benchmark, regardless of whether one is slightly higher on a performance metric since this could occur due to random chance.

- Thank you for your comments. We agree that a comprehensive sensitivity test is necessary to argue the robustness of our emulator. In the revised version, we add a set of sensitivity tests to check if the emulator works well for a wider range of σmax threshold settings. Please see our previous points about the sensitivity test and sections 4.4 and 5.1 for more details.

L271: remarkable in what sense? This is related to point #3 above - what is your benchmark that you are comparing to? Significance at 0.95 or above?

See point above about this. If two different data sets (the NN outputs here) are significantly correlated to a benchmark dataset (the ISSM outputs), then the conclusion is that neither can be statistically distinguished from the benchmark, regardless of whether one is slightly higher on a performance metric since this could occur due to random chance.

- As you suggested, we calculate statistical significance to check (i) if a certain model is *significantly* different from the others in terms of RMSE and (ii) if a test RMSE is *significantly* different from a train RMSE (see Fig. 3, Table A1, and Table A2).

L391: how are these emulators promising for parameterizing future behavior? They provide no way of constraining σmax without observations and you haven't demonstrated that they can 19 extrapolate outside the temporal sample of the training data. Perhaps they could be used to do uncertainty quantification since they enable cheap MCMC sampling of parameters space.

Predicting one time step in the future does not constitute the ability to do future projections in ice sheet models, since ice sheet projections almost always entail many time steps of a model. I think that you need to be very clear in the conclusions that what you have demonstrated is that the GNN emulator works well over a period in space and time where some ice sheet model simulations have already been done and a cheap emulator is needed for parameter calibration or UQ. However, you have not shown that these emulators can perform well without ice sheet model simulations already existing. You should be very specific about this point so that readers do not get the sense that this is meant to replace ice sheet models in any way.

- Thank you for pointing out this. In this study, what we ultimately want to do with this GNN emulator is "to find the optimal (*now* temporally varying) σmax for accurate modeling of glacier calving." Our GNN emulator makes it possible to find this calving calibration setting that aligns with observations orders of magnitude faster than numerical simulations. Then, we conduct numerical simulations with this optimal σmax value, and the simulation results agree well with the observations compared to conventional constant σmax settings. In the revised manuscript, we highlight that the optimal σmax settings derived from the GNN emulator help accurate modeling of the calving front (see section 5.3 for more detail). Additionally, we add a comprehensive discussion about the implication of this temporally varying σmax in section 6 (L372-395).

  "In this section, we discuss the scientific implications of σmax variation and the significant role of GNN emulators in improving the physical modeling of glacier calving. Our results reveal that the fine-tuning of temporally varying σmax leads to better performance in calving prediction compared to traditionally used constant σmax parameterization. As a parameter indicating how difficult calving can happen, a greater σmax means a lower rate of calving, and a lower σmax means a higher rate of calving. Previous studies have shown that the physical condition of mélange around the ice front, such as ice thickness and concentration, changes the frequency of calving events; a weaker or disappearing mélange area leads to an increase in calving events, which corresponds to lower σmax (Wehrlé et al., 2023; Xie et al., 2019; Meng et al., 2025). Although this study is focused on finding optimal σmax values using GNNs, exploring the linkage between optimal σmax and mélange conditions can contribute to the accurate parameterization of σmax for calving models. Additionally, given that this mélange and sea ice conditions can vary along the ice front, assigning spatially varying σmax can address the discrepancies in calving front migration for two flow lines shown in Figures 7 and 8. It would also be valuable to investigate how σmax interacts with other environmental factors, including tidal forcing and deep water circulation, which are known as significant drivers for calving events (O'Neel et al., 2003; Bassis and Jacobs, 2013; Slater and Straneo, 2022). Consequently, finding the optimal calving parameterization setting and connecting it to external environmental drivers will contribute to the accurate modeling of glacier mass loss and sea level rise beyond the conventional constant calving calibration.

  Given the scientific implication of temporally varying σmax, the significance of this study lies in the (i) fidelity and (ii) computational efficiency of using GNN architectures, especially EGCN, in emulating numerical ice sheet models operating on unstructured meshes to find the optimal calving parameterization. By applying GNNs directly on raw unstructured meshes, we can keep high resolution around the fast ice area to delineate the calving front accurately. In particular, EGCN shows the best fidelity in predicting ice thickness due to its equivariance concept throughout the graph structures. Moreover, the

use of GNN emulators results in a dramatic increase in computational efficiency, with speeds 30-34 times faster than the ISSM simulations. This speed-up allows us to find the temporally varying optimal σmax values quickly. Given that running numerical simulations every time to determine the optimal σmax setting is extremely time-consuming, our fast GNN emulators reduce the workload associated with calving calibration."

Reviewer 2

Summary

I reviewed this paper as part of a second-round review; I was not involved in the first round of review. The paper demonstrates the application of graph neural networks as emulators for ISSM unstructured-grid simulations of Helheim Glacier, showing that the GNNs are able to accurately emulate the model outputs. The authors then go on to show that the much quicker run time of the emulator allows them to easily determine the required σmax parameter in a von Mises calving law to reproduce observed calving-front positions at Helheim between 2007 and 2020. They also compare their results to those achieved using a convolutional neural network, of the type previously used in ice-flow modelling.

- We sincerely appreciate your time and effort in reviewing our manuscript, even though you were not involved in the first round of the review. Thank you for your insightful and constructive suggestions. We carefully reviewed all your comments and did our best to address your concerns.

I am honestly unsure what to make of this paper. I think there is a good GMD article in there about the application of GNNs as emulators for numerical glaciological models that use an unstructured grid – the machine-learning part is well-executed and makes more sense following the clarifications added in response to the first-round reviews – but, glaciologically, the paper fails to prove anything much. The authors appear to have a hang-up on proving that their approach is better than using a CNN, which is self-evident, as using a CNN to emulate an unstructured-grid model would be a poor choice to start with, but the authors perform a bad-faith comparison between their approach and using a CNN to do just that, and then keep mentioning at every available opportunity how the CNN was much worse than their GNNs. Both the previous reviewers pointed this out and I'm pointing it out again: either do the comparison in a fair manner or delete it entirely. As it stands, it makes the authors seem absurdly competitive about something no one else was competing over.

- Thank you for your suggestion, and we agree that it is not necessary to include the comparison with FCN. We removed the comparison with FCN from the revised manuscript. While we wanted to show why we chose GNN instead of other common NN

architectures, we now focus more on the scientific implications (i.e., finding an optimal evolution of σmax from a glaciological point of view) we obtain from GNNs.

On the glaciological side, the main finding seems to be – I think unintentionally – that the von Mises calving law isn't very good as a physical basis for calving. This might also be considered a little obvious: no existing calving parameterisations do a good job. I am also unconvinced that the emulator would do well at predicting the calving-front position at another glacier, or at Helheim at a different time, further limiting its glaciological relevance. At least, the authors provide no information or examples showing that their method would yield good results in such a case. I again feel I'm not saying anything particularly new compared to the first-round reviewers, but it bears restating.

- This is a very good point for the reviewer. While the paper focuses on calving, we essentially present a new approach to efficiently constrain physical parameters that traditional approaches would struggle with. We agree that calving is a poorly constrained process, but at this point von Mises is "the best" calving law we currently have. When new, better calving laws are proposed, the method described in this paper can still be employed to efficiently calibrate calving.

My recommendation would be to take out the comparison to a CNN and the glaciological interpretation, which is limited, unconvincing, and feels like an afterthought, and submit the core paper about the technical advance of applying GNNs successfully to an ice-flow model for the first time to GMD. If the authors want this to be published in a disciplinary journal, there is a substantial amount of work that needs to be undertaken, and I think it would be a case of revise and resubmit, as it would be too much to do within a major revisions timeline. I would also like to record my disappointment that I'm having to essentially restate many of the points raised by the first-round reviewers, as the authors seem to have not properly engaged with the review process beyond clarifying their own method, which was an important point raised by the initial reviewers, but by no means the only one, nor the one that was most damaging to the paper. I've consequently left this review in a more aggrieved tone than I would usually adopt in the hope that it communicates to the authors that they cannot brush these issues off and that substantial work is needed to properly consider and address them. Page and line numbers refer to those in the clean version of the submitted manuscript.

- We apologize that our previous round of revisions have not addressed the reviewers' concerns completely. We did not mean to overlook the reviewers' comments, but we admit that we were too excited about showing the effectiveness of GNN architecture compared to other existing NN architectures rather than showing any scientific significance, which a journal like The Cryosphere really pursues.
- In the manuscript, we removed the comparison with CNN (FCN) and now focus more on how to interpret the results we obtained from the GNN emulators to make it suitable for The Cryosphere. We hope this revision addresses your concerns sufficiently. Please see

the detailed responses below and our updated manuscript (particularly 5.3 and 6 for scientific discussions).

Major Comments

Comparison to FCN: see my increasingly tetchy comments below, but, as it stands, the comparison to the FCN tells us nothing about the relative performance of it or the GNNs and, worse, gives the whole paper an overly competitive tone that reflects badly on the authors and detracts from the more sensible bits of the paper. The comparison either needs to be done in such a way that it isn't just proving interpolation reduces numerical accuracy, or it should be abandoned. I would argue for the latter because I don't think it's needed: it's clear that GNNs are well-suited to this application because they can run on an unstructured grid, and, for that very reason, one wouldn't try to apply a CNN of some kind.

- Thank you for your comments. We agree that the comparison with CNN is not necessary. Instead, we focus more on comparing GNNs and how to interpret the calving parameterization we derive from the GNNs.

Glaciological interpretation: The current interpretation of the results is unconvincing and extremely superficial. It also contradicts the expected behaviour of σmax as outlined by the authors themselves in the methods section, but no real explanation or discussion of this is provided in the paper (it is in the response to Reviewer 2, but the paper hasn't been changed to reflect it). As I've gone into in more detail below, I'm fairly certain the underlying problem is that the von Mises law is fundamentally quite bad as a physical representation of calving, and that therefore attempting to find a physical explanation for changes in σmax is a wild goose chase. It may be that the authors can come up with a convincing physical explanation, but they certainly haven't yet, so they either need to put the work in to do so, or abandon the glaciological interpretation and send this off to GMD as a technical modelling paper.

- Thank you for your comment, and we apologize that our previous manuscript did not address this issue appropriately. We recognize that calving is an active area of research in the glaciology community and that new and improved calving laws may be developed in the future, but many models use this calving law since it is "the best we have". Based on the physical background of this calving law, we added some discussion of the physical explanation of σmax and the implications of this parameter linked to external environmental conditions (mélange, ocean tide, ocean circulation). You can find details in sections 5.3 and 6, but here we would like to provide a brief summary of our findings:
- (1) Temporal variation of optimal σmax agrees well with the calving front migration (correlation coefficient > 0.5; p-value < 0.01) (Table 3).
- (2) We added a parameter verification section: we performed an additional numerical simulation (based on VM calving law) with the fine-tuned σmax calibration, which

reproduces the observed calving front movement well compared to the traditional constant σmax parameterization (Figs. 7 and 8).

- (3) By connecting the σmax parameter (i.e., a control on calving rate) with external environmental conditions (e.g., mélange, ocean tide, ocean circulation), we can further determine how to parameterize the appropriate σmax depending on environmental conditions. This will eventually contribute to the accurate prediction of future glacier mass loss and sea level rise. This study demonstrates that GNN can accelerate finding the optimal calving parameterization in this workflow.

Generalisability of the emulator: the authors are quite cagey about this, but all the results as presented in this paper prove is that the GNNs do a good job of emulating the specific ISSM simulations used to train them (or, to be fairer, ISSM simulations at Helheim within the parameter space, or very close to the parameter space, defined by the training data). I get no sense of whether they would perform well if applied to Helheim at a different time, or to another location, which again limits the glaciological interest of the paper. They may well perform well, but the paper doesn't show this. Again, unless the authors are prepared to do some substantial additional work showing a second application and proving that the emulator still does a good job, this paper really should go to GMD as purely a technical advance in applying GNNs to glaciological simulations. As a related issue, if, as the authors themselves seem to admit in their response to Reviewer 2, and as the results of this paper seem to show, σmax is not in fact actually physically meaningful but just a numerical fudge factor to get the model to agree with observations, how can the emulator be generalisable as there is quite possibly no consistent underlying pattern to learn that could be extrapolated to another time or place (particularly to the future where there would be no observations)? If the purpose of the emulator is to find the best value of σmax, but it has to be retrained for each new glacier or period, then, glaciologically, what is the usefulness of the emulator if you've got to run the numerical simulations anyway? I agree GNNs seem to be promising as a way of emulating ice-sheet models – that is a nice technical advance – but this particular application of them does not seem overly useful as presented.

- Thank you for your constructive comments. As we want to keep our manuscript in the review process of The Cryosphere, this is a very important point we need to take very seriously. We recognize that calving is an active area of research in the glaciology community and that new and improved calving laws may be developed in the future, but many models use this calving law, since it is "the best we have". We would like to emphasize that our paper is centered on an efficient approach to constrain a complex physical law (that involves moving boundaries!) and is not restricted to this specific application. We also note that σmax has been found to be around 1MPa consistently across the coast of Greenland, with a few exceptions.

- In order to address this issue regarding generalizability, we made the following main revisions to our GNN emulator and conducted additional experiments for sensitivity tests:

(1) Adjust the input & output variables for better generalizability: We changed the input and output variables and retrain the GNN emulators to ensure better generalizability. First, now our emulators predict the next time (t) ice conditions purely from the previous (t-1) ice conditions; all the input variables are from the t-1 time period. Second, instead of using the level set ice mask as the direct output variable, our emulator now predicts "tensile stress (sigma)" as the output variable. Then, the level set ice mask and calving front for the next time step are derived from this ice velocity and tensile stress condition with a certain σmax value using equations 1, 2, and 3 (Fig. 2). By employing these new input-output settings, we can purely infer how the previous-time step ice conditions (ice thickness, ice velocity, SMB, etc.) change the next-time step ice conditions (ice thickness, ice velocity, tensile stress). Then, the movement of the calving front is determined purely by the physical laws of VM calving law with an adequate σmax value. This new approach helps us (i) improve the predictability and accuracy of our emulators and (ii) explore the true physical meaning of σmax by directly applying this value to VM-based physical equations. Indeed, now we achieve the binary calving front error of less than 1 %, corresponding to > 99 % of agreement with the ISSM simulations.

(2) Add comprehensive sensitivity tests: To test the robustness of our emulators with the selection of different training samples (i.e., different σmax as training datasets), we conduct four sensitivity tests. We divide the 9 σmax simulation data into training and test datasets in four different ways (please see Table 1): (A) even-value samples (0.75, 0.90, 1.05 MPa σmax as training dataset), (B) low-value samples (0.70, 0.75, 0.80 MPa σmax as training dataset), (C) medium-value samples (0.85, 0.90, 0.95 MPa σmax as training dataset), (D) high-value samples (1.00, 1.05, 1.10 MPa σmax as training dataset). By evaluating the model performance for these four different training set settings, we would like to check how the model performance is sensitive to the selection of training samples. The sensitivity test results show that the emulators have less than 10-20 % of uncertainty in the selection of training samples (Fig. 3 and Appendix A), which supports that the emulators trained with a certain set of σmax can represent the general cases for all σmax values.

- We also added more discussions about the physical meaning of σmax, focusing on (1) how the temporally (and potentially spatially) varying σmax can contribute to more accurate ice sheet modeling and (2) its connectivity to external environmental conditions that determine the calving rate, such as mélange and ocean forcings. Please see sections 5.3 and 6 for more details.

Minor Comments

p.6, l.125: Delete ‘merely’ because a) it doesn’t really make sense here and b) it makes it seem as if the authors are saying models on regular grids are rubbish, which is perhaps not an ideal tone to strike.

- Done.

p.12, l.284-286: Don’t think this paragraph really needs to be here, unless the authors are taking an extremely dim view of the intelligence or memory of their readers!

- Thank you for pointing out this. We remove this paragraph.

Table 1: Sorry to bang on about this, both of the first-round reviewers having raised this point, but is the R metric really showing us anything useful, beyond all three GNNs are better than the FCN (which is obvious from the RMSE and calving-front accuracy anyway)? The R numbers for the GNNs are all so similar as to be quantitatively meaningless and I’d be wary about placing too much emphasis on very slightly different values for the third decimal place.

- We understand that the reviewers have a lot of concerns about R values. We remove R and only include RMSE as the evaluating metrics. We also test a statistical significance to determine whether one GNN model is *significantly* better than the others from the statistical point of view (see Appendix A).

p.13, l.298-303: Yes, these are fair criticisms of the fixed grid used by a CNN, but as both the previous reviewers state, the FCN is being set up to fail because the training data is interpolated onto the fixed grid at the start, introducing errors, and the results are then interpolated back onto the unstructured ISSM grid for the comparison, introducing more errors. In that situation, the FCN is mathematically virtually guaranteed to do worse. I do not doubt that the GNNs are a more natural fit for an unstructured grid and perform better on it, but if the authors want to compare the performance of the GNNs to an FCN, the FCN needs to be trained with data produced on and results evaluated on a structured grid. Otherwise, this comparison boils down to ‘interpolation is bad for numerical accuracy’, which isn’t the most striking finding in the world. To be honest, is there even any need to compare to a fixed-grid CNN? The rest of Table 1 shows that the GNNs are doing a good job on the unstructured grid, which is the important thing; trying to prove that they’re somehow inherently better than a CNN seems unnecessary, especially when the application presented here is clearly not one a CNN would be used for, because ISSM runs on an unstructured grid. Even if the authors have some particular animus against CNNs (I can’t help feeling there’s maybe a little bit too much of an attempt to prove that the method here is inherently better than IGM’s, which is an unhelpful attitude), I would strongly suggest scrapping this comparison entirely, and limiting it to the existing earlier remarks about how GNNs are a natural fit for an unstructured grid and that a CNN would be inappropriate for this application because it would require a fixed grid.

- Thank you for your suggestion. In the previous version, we were too excited about the comparison between FCN and GNN itself just to convince the reason why we chose GNN for our emulators. However, we agree that it would be better to remove this FCN-GNN comparison entirely and only briefly mention some remarks about the fact that GNNs are a natural fit for an unstructured grid. We refer to our recent paper (Koo and Rahnemoonfar, 2025) for this. Instead of comparing FCN and GNN, we now contribute more to the physical interpretation of the optimal σmax value we derived from the GNN emulators.

p.15, l.324: Yes, but compared to what? Presumably, given what follows, actually solving the model on CPUs, but then this advantage is true of all neural networks, not specifically GNNs. Some rephrasing might be needed here to make it clear which advantages are generic to neural networks and which are specific to GNNs.

- We rephrase this paragraph as follows (L317-319):

  "The most significant advantage of GNN emulators is their ability to reduce this computation time by leveraging GPUs while maintaining finite element structures, which is not completely available via other non-graph neural network architectures (Koo and Rahnemoonfar, 2025)."

- Additionally, we add detailed comments about the significance of GNN in section 6 (L387-395)

  "Given the scientific implication of temporally varying σmax, the significance of this study lies in the (i) fidelity and (ii) computational efficiency of using GNN architectures, especially EGCN, in emulating numerical ice sheet models operating on unstructured meshes to find the optimal calving parameterization. By applying GNNs directly on raw unstructured meshes, we can keep high resolution around the fast ice area to delineate the calving front accurately. In particular, EGCN shows the best fidelity in predicting ice thickness due to its equivariance concept throughout the graph structures. Moreover, the use of GNN emulators results in a dramatic increase in computational efficiency, with speeds 30-34 times faster than the ISSM simulations. This speed-up allows us to find the temporally varying optimal σmax values quickly. Given that running numerical simulations every time to determine the optimal σmax setting is extremely time-consuming, our fast GNN emulators reduce the workload associated with calving calibration."

p.15, l.332: I'm assuming there isn't any interpolation time being counted in the FCN stat here? Otherwise, again, not a fair comparison.

- The FCN comparison is removed.

p.17, l.346: ...possibly, the FCN is taking longer because the data it's being fed are inherently worsequality because they've been interpolated? Though I admit that the magnitude of the speed-up in training is such that I don't doubt it's real, but my point again is that the comparison is essentially meaningless as presented

- The FCN comparison is removed.

Figure 3 and Figure 4: They're a bit much. I might suggest reducing the information overload by just having a nice six-panel figure of ISSM field, best-performing GNN field, difference, for the two parameter values, showing that the GNNs are doing a good job, and then sticking the 72-panel monsters in the supplementary information so that readers don't just glaze over with eye strain in the main paper itself. Something on the scale of Figure 5 is a much nicer presentation!

- Thank you for pointing out this. We reduce the number of panels to 24 to simply show the spatiotemporal errors of 3 GNN emulators.

p.17, l.349-354: I really think the CNN-bashing is getting a bit silly here, especially given how flawed the comparative basis in the paper is. See my earlier comments, but unless the authors are going to put in the work to do an actual fair comparison, statements like this are built on sand and make the whole paper seem weirdly aggressive. Do the work or delete the comparison entirely. If, after a fair comparison, GNNs do just turn out to be better, bash away, but right now, this is an untenable claim. The most that can be said would be something like 'While we have not conducted a full-scale comparison between GNNs and CNNs, owing to the difficulties introduced by the fundamentally different grid requirements, our successful emulation of ISSM shows that GNNs are inherently well-suited to replicating the results of finite-element models that use an unstructured grid, an application where a CNN, with its requirement for a structured grid, would struggle.'

- The FCN comparison is removed.

p.18, l.371-376: This feels like a bit of an afterthought. And also doesn't hold water. A higher σmax value should mean stronger, more stable ice (calving is more difficult), as the authors state on p.5, l.101-2. Here, σmax is increasing after 2014 as the calving front retreats rapidly, which is a contradiction in terms. As a related point, the ISSM simulations include ocean thermal forcing, so the authors should be able to say confidently if that's important here too, and it should be easy from remote-sensing observations to work out if there was more mélange after 2014, rather than the current weak formulation. Regardless, either the authors have to come up with an explanation of why a rapidly retreating calving front is, against all expectations, actually one more resistant to calving, or admit that σmax is not really physically based and is just a tuning parameter that is compensating for other errors and processes not included in the model, in which case interpreting changes in it is worthless. The latter is essentially what the authors do in their response to the same point raised by Reviewer 2, but the text here (and the wider paper) should be changed to

reflect that, rather than the current presentation of the issue, which attempts to present the parameter as physically meaningful despite the evidence of this paper's own results and the authors' response to the previous round of reviews.

- Thank you for your constructive comments. We agree that a deep discussion about the physical meaning of σmax is critical in a disciplinary journal like The Cryosphere. To highlight the scientific meaning of σmax calibration, we updated the optimal σmax with the updated GNN architecture (we have changed the input and output settings to make the emulator represent the physical meanings of σmax better). Moreover, we now discuss the physical meanings of this σmax in sections 5.3 and 6.

p.18, l.378-388: I'll let the authors interpolate my comment from those above: this is a use case where obviously using a CNN is a silly idea, and the comparison isn't fair, so there's really no need to keep sniping at them constantly. At this point, an equivalent argument is 'we proved this screwdriver was better at driving screws than this dead fish.' I mean, great, but people might think it was a bit odd that the dead fish was considered a sensible comparison in the first place.

- We agree with your point. We remove the comparison with CNN and also remove this paragraph in the updated manuscript.

p.20, l.406: This is probably why the earlier interpretation of the changes in σmax doesn't make sense.

- We update the optimal σmax with the updated GNN architecture (we have changed the input and output settings to make the emulator represent the physical meanings of σmax better), and we discuss the physical meanings of this σmax. Please see sections 5.3 and 6 for details.

p.20, l.408: Do they? It's an emulator: by definition, it's not going to tell anyone anything much about the underlying processes or mechanisms. I certainly can't say that I feel I've found out anything about calving mechanisms so far.

- We rephrase this sentence (L410-411): *"Thus, although our GNN emulators can provide valuable insights into how to select appropriate σmax for VM calving law, they rely on our current imperfect physical understanding of calving."*

p.20, l.409-411: The emulator isn't really emulating calving processes, because this ISSM setup is not modelling calving processes. The emulator is emulating a parameterisation of those processes, a parameterisation that explicitly ignores all the processes going on in favour of having a single easy parameter to play about with. The VM calving law is not itself a calving process, it's merely a (flawed) representation of underlying processes that are being ignored. Please be more careful with the language here and make it clear what the emulator has done and can do (emulate ISSM solutions and determine the correct parameter values to match observed

calving front positions), and what it can't do (provide any information at all about any of the underlying processes, especially if σmax isn't really all that physical)

- This paragraph is removed. Instead, we update section 6 to highlight the significance and limitations of our GNN emulator. Please see our previous responses and section 6.

References

Koo, Y. and Rahnemoonfar, M.: Graph convolutional network as a fast statistical emulator for numerical ice sheet modeling, Journal of Glaciology, 71, e15, https://doi.org/10.1017/jog.2024.93, 2025.