Reviewer 2

This is a review of the pre-print by Koo et al. in The Cryosphere titled "Calibrating calving parameterizations using graph neural network emulators: Application to Helheim Glacier, East Greenland". This study describes the use of Graph Neural Networks (GNNs) of various types to emulate the behavior of ISSM, a finite-element ice sheet model. There is also extensive comparison of the use of GNN to more traditional FCNs, which require input data to be on a uniform, rectangular grid. Once the NN is trained, it can be used to predict ice thickness, velocity and terminus position (through an ice mask) at a subsequent time step based on these fields at the prior time step and a parameter governing calving behavior in a model parameterization (sigma_max).

The most novel aspect of this study is the use of GNNs to emulate a finite-element ice sheet model. The study makes a good case for why this type of NN makes sense for emulating a model on a non-uniform mesh, though I'm not sure the way in which its accuracy was compated to FCNs is completely fair. In that sense, I found this the most compelling potential advance of this study to be the development of a general purpose ice-sheet-model emulator (like IGM, but with some advantages). I am less convinced that we have necessarily learned much about calving from the use of this new method. I explain some of my major issues in this regard below and then a list of smaller suggestions below that.

- Thank you for your constructive comments. We carefully reviewed all your comments and will handle them to improve our manuscript. Please see the below responses to your comments for details.

Major points:

1. As I read through the study, I found myself unclear about the scientific use of this methodological advance. The GNN will emulate ISSM at high accuracy and significantly lower computational cost. What questions will that help us to answer that isn't possible with conventional methods? This is a particularly important question to answer since this is submitted to The Cryosphere, a disciplinary journal, as opposed to a more methods oriented journal like GMD or JGR:MLC.

- The objective of this study is to facilitate the assimilation of numerical models and observations using GNN emulators. Since it is quite challenging to find optimal parameterizations of numerical models that are consistent with observations because of the computational demands of numerical models, we propose to use the GNN emulators to find appropriate parameterizations of numerical models. In the revised manuscript, we will highlight the objectives of this study and how this can contribute to scientific findings.

Once I got to the end, I saw that the main application this new emulator was used for was essentially something like transient parameter estimation (using the low cost of the NN to enable an exhaustive grid search for sigma_max at each time step). But then the result of this application didn't make physical sense. The calving front retreats while sigma_max increases, which is sort of opposite what should happen *if* calving drives retreat (which it may not). The text pushes off the explanation on "other processes" without much investigation of whether the methods may be at fault, or other potential explanations. Ultimately, this is a challenge of using completely data-driven ML without further investigation of the latent space of the NN - the emulator is a black box, so it is challenging to diagnose what is happening in it that causes this counter-intuitive result.

- It is true that the calving front retreats while sigma_max increases in this case, but we can also see the ice velocity increase during the same period. Mathematically speaking, as shown in Eqs. (1) and (2), the calving rate can increase with increasing ice velocity or increasing tensile stress. Meanwhile, decreasing sigma_max will increase the calving rate. In general, lower sigma_max means the ice front is easier or more susceptible to calve. However, as we showed in this manuscript, as the ice front retreats further upstream, sigma_max increases, which means the calving rate may decrease. We agree that the von Mises calving law is a simplification of reality, and this law certainly misses some important internal or external feedbacks. The result we showed in this study is the best parameterization to reproduce observations. We meant "other processes" to express the potential imperfection of the VM method. Although we chose the VM calving law for Helheim Glacier because the previous studies showed that this law fits the Helheim Glacier, this calving law would not be 100 % perfect to describe every detail of the calving process. In the revised manuscript, we will add discussions about the limitations of the VM calving law.

2. The study, as it stands, has not convinced me that the GNNs trained as they were in this study, generalize at all outside of the very limited training data. The test data is completely within the interior of the limited parameter/state space on which the GNNs are trained. If I simply used linear interpolation to generalize from the training data to the test data, how accurate would that be in comparison? It would certainly be computationally cheap.

- This is an interesting idea. In general, though, calving fronts do not always respond linearly to changes in sigma_max, and we find tipping points or thresholds. We agree that this is a good idea though and will add some experiments to the Appendix by applying a linear interpolation with neighboring σmax values as suggested, to see if the GNN is more accurate than a simple interpolation.

More importantly, the GNNs have not been tested on any cases that are out of the temporal or spatial sample of the training data. If the aim is to narrowly train the model to do a really good job learning what Helheim did from 2007 to 2020, thats OK, but state that narrow expectation explicitly. There are places in the study where you say that these GNNs could be used to replace an ice sheet model more generally, or in future simulations, but you haven't really shown the ability of the GNNs to do that, since they haven't been tested outside of this very narrow place and time period.

- At this stage, we are just looking at a proof of concept that GNNs are viable and useful tools for parameter search. We only focus on how our GNN emulators work, specifically for Helheim Glacier, in the time periods from 2007 to 2020. In this study, we highlight the potential of EGCN "network architecture" to replicate the finite-element numerical ice sheet modeling, especially in delineating the calving front in the dynamic ice sheet system. Since this GNN or EGCN have not been used explicitly in ice sheet modeling, our study introduces a new useful tool to the field. However, we agree that the generalizability and time-invariant of the emulator are extremely important for further general applicability. This would be our next step based on our current results, and we will modify the network architecture to guarantee the generalizability of our emulator.

3. 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.

- This is a good point that was also mentioned by the first reviewer and will be revised. We found that the R values are too inflated because we calculated this metric for the entire glacier domain. However, the significant differences are only near the ice front and ice stream, while the other regions with slow ice show insignificant differences. We will recalculate these metrics only near the ice front and ice stream.

Additionally, the way that you train and then assess the accuracy of the FCN does not provide a fair comparison to the GNNs. By interpolating from the finite-element mesh to a uniform rectangular mesh, you've done two things: lowered the resolution of the training data in the finest parts of the grid and inflated the relative weight of the coarse parts of the grid by increasing the number of grid points in these areas. The places with the finest resolution in ISSM are also places where velocity is the highest and where the ice mask is changing (i.e. near the terminus) which will tend to make errors more important. effectively, after interpolating you have given the FCN worse training data than the GNNs. The least you can do is interpolate the FCN training data onto a uniform grid with resolution equal to the finest resolution in the ISSM mesh. Additionally, using some knowledge about where errors are likely to be the largest, you can apply weights in the FCN training loss function which are proportional to the finite-element grid resolution. In that way, you will be "fixing" the mis-weighting that has occured by interpolating the training data that you then assess accuracy on.

I get that in some sense your whole point is that FCNs are not natural fits for finite-element training data, but with the relatively minor differences in accuracy you find, its hard to discern whether this is due to the NN being superior at capturing the data vs. the training data just being different due to interpolation artifacts. These are very different claims.

None of this changes the fact that GNNs are likely to be much more efficient at natively training and then running on the finite element mesh. I believe your case that they are computationally superior, but I'm not sure I see much difference (or a fair comparison of differences) in the accuracy. My suggestion is simply to focus on the fact that emulating finite-element models (which most modern ice sheet models are) is more natural using GNNs since it doesn't require interpolation and that the computational advantage of GNNs over FCNs is massive. The GNNs do a great job accurately emulating the model by any objective measure, so emphasize this.

- The need for additional interpolation from the finite-element mesh to a uniform rectangular mesh is one of the main reasons we claim that FCNs are not “natural” for emulating finite-element models, such as ISSM. As you point out, this additional interpolation brings significant problems in replicating finite-element models: (1) loss of detailed resolution in the finest element area with fast ice velocity (primarily near ice front and ice stream), (2) allocation of unnecessary computational loads to coarse-resolution areas where ice velocity is slow. We emphasize these limitations of FCN and the advantages of GNN in replicating “unstructured-mesh” simulations.
- Our current loss function, mean square error (MSE), already works as a sort of “weighting” loss function because this loss function weights the large error area by squaring the errors. Moreover, the modification of model weighting of FCN does not really solve the two problems of the FCN for finite-element model: (i) loss of details and (ii) inefficient allocation of computational resources. These problems are caused by the “fixed-resolution” nature of FCN for all locations, and increasing FCN resolution would bring an exponential increase in computational cost. However, in our preliminary research, we found that the GNN architecture is a “flexible-resolution” approach, which is not dependent on changing resolution. That is, even if we change the spatial resolution of unstructured meshes, the performance of GNN is not affected by mesh resolutions. The preprint version of our preliminary paper is available at this link: https://arxiv.org/abs/2402.05291
- In this aspect, we argue that the fact that we do not need to interpolate the model results on a regular grid is a significant advantage of GNNs.

Minor suggestions:

L1: Increasing calving has been linked to the retreat

- Yes, higher calving rates can lead to retreat. However, calving is also somehow responsible for the acceleration and thinning of glaciers, so we would like to keep them "separate" in the text as well.

L3: have been used to simulate ice

- Done.

L10: reproduce the observed evolution

- Done.

L22: total ice sheet mass loss

- Done.

L28,30: optimal in what sense?

- We meant the optimal parameterizations that match observations. We clarify it in L28.

L35: as a boundary condition in numerical

- Done.

L41: necessitate using high-performance

- Done.

L56: the training of emulators

- Done.

L60: outlet glaciers in Greenland

- Done.

L79: The migration rate of the ice front

- Done.

L82: ice front migration rate (velocity is confusing here because it could refer to other things)

- Done. Thank you for your suggestion.

L87: VM has not been defined as an acronym

- VM is defined in L29.

L87: correlates with weaker ice

- Done.

L89: many observational studies have found tensile strength as low as 100 kPa (Vaughn 1993 is a particularly well known paper), so I'm not sure where this lower bound is coming from.

- This lower bound value is from Morlighem et al. 2016 and Petrovic, 2003.

L91: is important to accurately reproducing observed glacier evolution

- Done.

L117: CNN cannot represent finite-element ice sheet models on their native grid

- Done.

L122: focused on calibrating calving parameterizations using

- Done.

L135: each transient simulation denerates of a total of 261 outputs between.

- Done.

L136: calibrated and held constant

- Done.

L136-140: the use of semicolons here is a bit challenging to read. Why not just write these as separate sentences?

- Done. We separate it into several sentences.

L146: adjacency matrices?

- Yes. We change it into adjacency matrices.

L151: we compare to remote-sensing

- Done.

L201: you aren't the first to develop an EGCN - you train a NN architecture that has previously been described in other papers

- We change “develop” to “adopt”.

L242: don't you mean validation instead of testing on this line?

- Yes, corrected.

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)?

- This training-testing split checks the applicability of emulators for various σmax values to match the numerical simulation and observations. By applying the emulators trained with various σmax values to out-of-sample σmax values (0.75 and 0.90), the current approach and results show that our emulators can represent the ice sheet dynamics and calving at out-of-sample σmax values within the range of 0.70-1.10 MPa.
- In the revised manuscript, we will add some experiments in the Appendix to get insight into how reliable GNNs are in replicating the simulation results for two σmax: 0.75 MPa and 0.90 MPa. We will apply a linear interpolation with neighboring σmax values and compare this with GNNs: comparing σmax = 0.8 and 0.7 to the test set for σmax = 0.75, and comparing σmax = 0.85 and 0.95 to σmax = 0.90.

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?

- We meant high R-values (significance at 0.99 for all cases) for replicating the spatial and temporal patterns of ice velocity and thickness. However, as we mentioned in point #3, these high R-values could be inflated because we calculated these metrics for the entire glacier domain. We will recalculate these metrics, covering only near the ice front and ice stream.

L305-315: it could be made clearer here that when tested on the exact same hardware, GNN are faster. Comparing wall time on two different processes or a different number of processors is not a fair comparison.

- Here, we should note that the ISSM runs on a high-performance computing (HPC) system because it is computationally demanding, while deep learning emulators (i.e., GNN and CNN) can be simply run on a local desktop. We would like to highlight that our deep-learning emulators have significant computational efficiency even on a local machine.

L350: I'm not sure I buy this argument partly because enough information hasn't been provided. Was ocean frontal melt included in the ISSM simulations? Do we know if melange increased at Helheim over these years? If so, it would presumably have an influence on calving rate, which could be captured effectively through sigma_max...This gets to the point above that this discussion here is entirely too brief and doesn't engage with any prior work on Helheim and its recent changes. If this paper is to be appropriate for TC, instead of say, GMD, then that discussion would be needed.

- Agreed. Since we only assume the VM calving law, we cannot really infer other factors besides the calving threshold. We will remove this part in the revised version. Instead, we will add some discussions about the limitations of the VM calving law, which we merely depend on to describe calving processes in this study.

L358: this begs the question: what would happen if you interpolated all the training data onto a rectangular grid, and then used that to train both the FCNs and the GNNs? This would be a fairer comparison than what you have now.

- Interpolating the training data onto a rectangular grid is not a good choice to delineate the calving front. It loses the detailed resolution at the calving front.

L364: CNN->FCN

- Done.

L373: why should GNNs be trained with numerical simulations?

- As "statistical emulators", GNNs should learn the statistical relationships between inputs and outputs represented in numerical simulations. We will add some explanation about this statement.

L385 with 13-year transient simulations of Helheim

- Done.

L391: how are these emulators promising for parameterizing future behavior? They provide no way of constraining sigma_max without observations and you haven't demonstrated that they can 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.

- Thank you for this point. Our emulator allows us to predict the next-time-step ice velocity and ice front conditions when the previous-time-step conditions are provided. Although we did not directly extrapolate the temporal sample of the training data out of 2007-2020, our emulator can predict the future behavior of ice sheets under certain conditions. As you suggested, we also agree that our cheap emulator enables uncertainty quantification via MCMC sampling.

References

- Choi, Y., Morlighem, M., Wood, M. & Bondzio, J. H. Comparison of four calving laws to model Greenland outlet glaciers. Cryosphere 12, 3735–3746, https://doi.org/10.5194/tc-12-3735-2018, 2018.

- Wilner, J. A., Morlighem, M. & Cheng, G. Evaluation of four calving laws for Antarctic ice shelves. Cryosphere 17, 4889–4901, 2023.
- Morlighem, M., J. Bondzio, H. Seroussi, E. Rignot, E. Larour, A. Humbert, and S. Rebuffi (2016), Modeling of Store Gletscher's calving dynamics, West Greenland, in response to ocean thermal forcing, Geophys. Res. Lett., 43, 2659–2666, doi:10.1002/2016GL067695.
- Petrovic, J.J. Review Mechanical properties of ice and snow. Journal of Materials Science 38, 1–6 (2003). https://doi.org/10.1023/A:1021134128038
- Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y.: Graph Attention Networks, https://doi.org/10.48550/arXiv.1710.10903, 2018
- Satorras, V.G., Hoogeboom, E. and Welling, M., 2021, July. E (n) equivariant graph neural networks. In International conference on machine learning (pp. 9323-9332). PMLR.