Review of 'Calibrating calving parameterizations using graph neural network emulators' by Koo et al.

July 2024

In this manuscript, Koo et al. describe the application of several variants of a graph neural network to emulate ISSM at Helheim Glacier. I am unsure as to exactly what this emulator does - whether it is similar to IGM in producing an approximate solution operator to the (in this case coupled) momentum and mass conservation equations or whether it is more similar to He (2023) in being geometry specific - but in general terms it is trained to reproduce the predicted ice velocity and thickness as a function of time and space.

I think that this paper has the potential to be a useful contribution to the literature, and the goal of coming up with emulators that operate naturally in the same discrete setting as FEM-based ice sheet models is worthy. Additionally, the saliency with which the neural networks either learn or memorize (I'm not entirely sure which) the model's behavior is impressive. However the manuscript needs significant clarification (and in some cases moderation) of its claims in order to assess their veracity and utility. In particular, the methods are unclear and not reproducible - as mentioned above, I am not clear what the features used for prediction actually are. Additionally, the paper does not include a fair representation of the computational costs of the proposed methods. Finally, the paper makes many claims that its proposed methodology is better than others without providing sufficient evidence to back up that claim.

Detailed comments are below.

L63 Downs (2023), which is already referenced in this paper, provides significant insight into this question, and also serves as another example of a surrogate model being used to infer calving dynamics at Helheim Glacier (though not a GNN).

L75 The acronym VM should be defined here.

L78 Physically, why should the stress threshold have to be calibrated on a glacier-by-glacier basis?

Sec. 2.3 This section should include some earlier literature. It would be worth looking at Tarasov et al. (2012, https://doi.org/10.1016/j.epsl.2011.09.010) and Brinkerhoff et al. (2021, https://doi.org/10.1017/jog.2020.112).

L123 I don't understand why GNN's would be particularly suited for ice front migration relative to other architectures.

L146 'adjacent' → 'adjacency'.

L149 Finite elements are, at their core, interpolants. Does bilinear interpolation here just mean that the FEM solution is evaluated at grid points? Or is some other interpolant introduced?

L179 the square root does not need to be defined.

L180 'of' → 'with'.

L185-187 I don't think that this description makes sense. The architecture *for some particular hidden layer* weights different adjacent nodes differently for an operation that is otherwise the same as vanilla graph convolution. The resulting convolutions are then stacked – graph convolution is stacked, but the attention operation is internal to that.

Eq. 6 I am not sure whether this equation is correct, but I am also not sure whether it's necessary - it seems like maybe something very detailed that appears in the reference (although it does seem weird to concatenate the projected feature vectors like this, which would seem to make the attention scores sensitive to the order of arguments). Maybe okay to forego?

Fig. 3 I don't think that this figure is helpful for illustrating how either of these models work.

Sec. 4.3 This section is really difficult to understand. One thing that I can glean from this is that this is operation is $O(n^2)$ in the number of graph nodes - does that have any implications for performance?

Sec. 4.5 This description of the model's inputs and outputs should appear at the beginning of the methods section. Furthermore, specifically what these features mean needs to be much more clearly described - as it stands, I cannot assess the quality of this work because, despite looking at both the manuscript and the linked code, I cannot tell what this emulator is building a mapping between. There are a few things implications to mention associated with this.

First, it appears that the time t is explicitly included as a feature. This then implies that the surrogate is *not* time-invariant and the mapping should be thought of as a tool for downstream analysis (like He (2023) or Downs(2023)) rather than as learning the solution operator for Stokes' equations (or Stokes plus continuity since the present work claims to predict thickness as well). This is fine - such models can be very useful - but it mandates a change in language to reflect the fact that it is unlikely that this method can generalize to other locations or times.

Second, the velocity components at some previous time (it's not clear whether this is from the model's previous time step or at the beginning of the simulation - this notation needs to be modified to be more clear) is used as a feature. This is an unfortunate choice because one thing that we know about ice physics is that the velocity is approximately diagnostic of the geometry - if you know the latter, you can predict the former. In the absence of that property, how can this model be started? Is it the case that ISSM has to be run first before this emulator can be applied?

What is a 'forwarding process'? What is a graph 'structure'? Is this just the collection of node/edge features?

This section is essential to understanding what is going on (more essential even than architecture choice), yet it's only two paragraphs long and does not have sufficient information to allow for reproducibility.

Eqs. 11, 12, 13 These are all standard definitions that do not need to be included here.

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 $\sigma_{max} = 0.8$ to the withheld test set values for $\sigma_{max} = 0.75$ (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.

Table 1 These metrics are all so inflated as to be useless. Is it possible to come up with relative metrics that use more significant digits? I also think it would be better to combine Tables 2 and 3 with Table 1 - it is useful to think of model accuracy relative to model size and expense.

Sec. 5.1 A single study does not establish the superiority of GNNs over CNNs for tasks such as these - it could (and very likely is) the case that the current results are incidental (or cherry picked) and that different researchers could find different conclusions. Furthermore, with respect to efficiency, there are many tricks that can be performed on CNNs to make them faster that have no analogue for an unstructured mesh, none of which were presumably included in the present analysis. The style of NeurIPS or similar notwithstanding, it is generally unhelpful to try to establish the primacy of one method over another in this way, and I would encourage the authors to either undertake a much more controlled and systematic comparison between methods or to reframe this as less of a competition.

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 proportionaly 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.

L351 This is only true if the model's behavior in response to variations in this new parameter is as well-quantified as it is to the parameter considered in this work, which may not be true, or may not be tractable.

Sec. 6 Again, I strongly urge the authors not to try to cast their work in terms of 'superiority' - the present work does not provide sufficient evidence for such blanket statements, nor is it necessary.

Fig. 7 This needs a more descriptive caption - I am struggling to see what these figures are showing.

L370–371 Is this statement really necessary?