Review of Koo et al. (2024) ‘Calibrating calving parameterizations using graph neural network emulators: Application to Helheim Glacier, East Greenland’

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 sigma_max 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.

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. 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.

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.

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.

- Glaciological interpretation: The current interpretation of the results is unconvincing and extremely superficial. It also contradicts the expected behaviour of sigma_max 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 sigma_max 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.
- 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, sigma_max 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 sigma_max, 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

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
- 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!
- 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.
- 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

- 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
- 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.
- p.17, l.346: ...possibly, the FCN is taking longer because the data it's being fed are inherently worse-quality 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
- 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!
- 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.'
- p.18, l.371-376: This feels like a bit of an afterthought. And also doesn't hold water. A higher sigma_max value should mean stronger, more stable ice (calving is more difficult), as the authors state on p.5, l.101-2. Here, sigma_max 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 sigma_max 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
- 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
- p.20, l.406: This is probably why the earlier interpretation of the changes in sigma_max doesn't make sense
- 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.
- 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 sigma_max isn’t really all that physical)