Review of C. Horvat and L. Roach — "WIFF1.0: A hybrid machine-learning-based parameterization of Wave-Induced sea-ice Floe Fracture."

The authors previously developed a physics-based model (SP-WIFF), which can capture wave-induced sea ice floe fracture and was used as a super-parameterization in the large-scale model CICE. However, including SP-WIFF increased the runtime of CICE by an order of magnitude. Here, the authors develop a neural network model (NN-WIFF), trained on the output of SP-WIFF, that can be used as a computationally efficient alternative to SP-WIFF in large-scale models. It significantly reduced runtime while still capturing the floe fracture patterns produced by SP-WIFF in CICE quite reliably. I found the paper was straightforward and easy to read (apart from section 2). It could be relevant for large-scale sea ice modeling, and I recommend it for publishing after revisions.

My main comment is that I couldn't really understand how SP-WIFF works just from reading the summary in the paper (section 2). I understand that it is explained in detail elsewhere, but it would help to add some more clarifications here so that it would be clearer what the baseline assumptions are — any error introduced by the neural network comes on top of this. Also, I feel that the claim that NN-WIFF is overall superior to physically based SP-WIFF-1, is somewhat unfounded (based on the presented evidence) — for example, under a climate change scenario the neural network could fail as the base state moves outside the parameter range of the training set, while SP-WIFF-1 could still be as accurate as before.

Major comments:

1. section 2: I had trouble understanding the basic assumptions of the model from this section. Here are some of the places where I got confused, but I would appreciate a more detailed discussion in general.
   (a) What is the timescale of Eq. 1 — does it evolve on the same timescale as the large-scale model (e.g. CICE), or is it considered to be infinitely fast compared to large-scale model evolution (i.e. for each time-step of the large-scale model, one finds a steady state distribution $f(r)$)? Do you consider Eq. 1 to be a part of SP-WIFF, or do just the steps S1-S3 fall under it?
   (b) Does this model explicitly ignore the possibility for ice floes to be advected between grid cells? Or is this possibility somehow included independently in the large-scale model?
   (c) At which timescales does the wave spectrum, $S(\lambda)$, evolve compared to the FSD? Is there a feedback with the FSD?
   (d) Why is $F(r, s)\mathrm{d}s$ independent of time, t, and of the duration $\mathrm{d}t$? I feel there is some implicit assumption here that I do not understand. Naively, I would think that the longer one waits and allows the ice to break, the higher fraction of the original floes would end up as small floes?
   (e) Why 10km in step S1? Is that the typical size of a grid cell?

(f) Does the floe size distribution, $f(r)$, enter the fracture algorithm, S1-S3? And, if so, how? It seems to me that it should since $\Omega(r,t)$, as defined at the beginning of this section, should be proportional to $f(r)$.

(g) Where does time enter S1-S3 (to give a time-dependent $\Omega(r,t)$)? Through $f(r)$ and $S(\lambda)$?

(h) Does S3 consist of repeating S1-S2 on the same floe (breaking it additionally with each step), or does each iteration start from a new 10km floe?

(i) I do not understand how steps S1-S3 yield the terms Ω and F as defined in the opening paragraphs of section 2. In the beginning, $\Omega(r,t)\mathrm{d}r\mathrm{d}t$ is defined as the fraction of the domain for which floes of size between r and $r+\mathrm{d}r$ fracture between times t and $t+\mathrm{d}t$. From Eq. 2, it seems that $\Omega(r,t)$ is the fraction of floes smaller than r that come from fracturing a very large floe (of 10km in size). How are these two definitions equivalent? Moreover, Ω as defined in the beginning has units of $\mathrm{m}^{-1}\mathrm{s}^{-1}$, whereas Ω in Eq. 2 is dimensionless. Likewise for $F(r,s)\mathrm{d}s$ — the introduction defines it as the fraction $\Omega(r,t)\mathrm{d}r\mathrm{d}t$ that breaks into floes of size between s and $s+\mathrm{d}s$, while Eq. 3 seems to suggests that it is the ratio of the number of floes of size between r and $r+\mathrm{d}r$ to those smaller than s. Again, I cannot reconcile these two definitions. Perhaps a more careful explanation of what $A(r)$ is and how it is related to Ω and F would help.

(j) Eq. 2: I believe it should be $\Omega(r,t)\mathrm{d}r\mathrm{d}t$ instead of $\Omega(r)$. Also, isn't $\int_0^\infty A(r)\mathrm{d}r = 1$, so that the denominator is unnecessary?

(k) Eq. 3: Shouldn't it be $F(s,r)\mathrm{d}r$ instead of $F(s,r)\mathrm{d}s$?

2. performance of NN-WIFF compared to SP-WIFF-1: A major drawback of neural networks (or any other "black box" method), is that we cannot rely on them in circumstances significantly different than those seen during training. Physically-based models, such as SP-WIFF-1, are not as susceptible to this. So, under a climate change scenario, SP-WIFF-1 could turn out to be a better choice. I feel this point was not really discussed much apart from a couple of sentences in the conclusions. Perhaps a short discussion (if not more investigation) about this would be useful. As a suggestion for future versions of this model (which I don't expect implemented here), it could perhaps be useful to include the possibility to flag data points that fall outside of the parameter range of the training set.

   (a) line 193: The model trained on 2009 data is tested on 2005 — did you compare with other years? Judging from white regions in Figs. 1d-f and 3, there are significant parts of the parameter space that are not visited during training, but perhaps could be visited under different conditions.

   (b) line 213: The difference between NN-WIFF and SP-WIFF-1 errors seems to be quite small, and typically much smaller than the spread of the error (e.g. in the Arctic the difference is 0.1% in the SAE metric). So, saying that NN-WIFF consistently outperforms SP-WIFF-1 seems like an overstatement to me. I would rather say that they are of quite similar accuracy, although NN-WIFF is significantly faster.

   (c) line 250: Again, I am not convinced that NN-WIFF is always more accurate than SP-WIFF-1.

   (d) paragraph of line 255: Perhaps you can expand on this discussion.

3. line 228: "This demonstrates that differences in WIFF implementation do not have an emergent effect on sea ice model state." — It could also be that ice fracture in itself does not have a major impact on the state of sea ice. Have you compared the sea ice state with and without ice fracture (here, or in some previous work)?

continued on next page. . .

Minor comments:

1. line 5: Can you explain what “bitwise reproducible” means?
2. line 17: Does $f(r)$ integrate to 1 or to sea ice concentration?
3. line 22: “Wave-affected sea-ice-covered regions are observed to be several million square kilometers in size in both hemispheres” — is this a significant fraction of the total sea ice area?
4. line 31: What is E exactly — total wave energy in a grid-cell? Or in some other area? E seems to be a normalized so that it has units m^2. Could you mention its units and say explicitly how it is normalized?
5. line 45: “it overall computation times by an order of magnitude” —“overall” → “increases”?
6. line 61: “unidirectional wave spectrum” — Can you explain what “unidirectional” means here?
7. line 91: What does 100x100 refer to? Hidden layers and nodes? That seems quite large.
8. line 92: “and a second with five hidden layers of 100 nodes each for generating fracture histograms.” — add a comma after “each”, this way it sounds like each node generates a histogram.
9. Figs. 1-3: Please add units to labels where appropriate (ice thickness, $\bar{R}$, etc.).
10. Fig. 1, caption: Please define “run rate” in the caption. Also, can you say what the white regions in panels d-f are.
11. Fig. 1c: There is a peak run rate at E between 0.01 and $0.1\mathrm{m}^2$. Does this mean that the floes break most at this energy, do you know if there is maybe a physical interpretation for this?
12. Fig. 2, caption: Could you please add references to equations 7 and 8 for $\bar{R}$ and SSE in the caption.
13. Fig. 2e: The way the figure is plotted, it is not clear what the bins are. Distributions look differently if the bins are linear or logarithmic, so it would be useful to show this somehow — e.g. add dots or plot bars.
14. Fig. 3f: Missing an x-label.
15. Fig. 3, caption: What seems white to me are the uncolored points, whereas $SSE = 10\%$ looks grayish-beige. Please either change the color-scale or explain the difference more carefully in the caption.
16. line 245: “Yet the “overall” accuracy of calls to NN-WIFF is 96.5%.” — This sentence comes before an explanation of what “overall” means, so it is a bit difficult to read.
17. Fig. 5, top row: Notation A for area could be confused with the distribution $A(r)$.