The manuscript by Bateson et al., "Sea ice floe size: its impact on pan-Arctic and local ice mass, and required model complexity" compares two of the main approaches for incorporating floe size distribution into a sea ice model (both using CICE) with observations in two ways. The first compares the floe size distribution with new FSD estimates from satellite imagery; the second is an evaluation of Arctic sea ice mean state over approximately the past 3 decades. With the newness of these models and the community focus on their implementation, this work is well justified. The manuscript is generally easy to read and complete. However, I have a number of concerns about how the comparison has been completed, and the presentation of the results.

<u>Major comments:</u>

- **Journal fit**. This fundamentally is presented as a model evaluation study. Is The Cryosphere the appropriate venue for this? Would GMD perhaps be a better fit? If published in TC, the authors should more clearly address and center what new science is presented.
- **Comparison with FSD observations**: I'm fundamentally a bit hesitant that floe size distribution models in global climate models are yet at a point where we expect them to match with observations (from specific location) as many other, more rigorously tested and developed features of models, can still not do so.
  To phrase this as a question: Why do you expect the models to represent realistic floe size distributions at a given point? Do you think there are other model factors that this representation is sensitive to, such as thickness distribution? Please discuss how other factors might impact this comparison
- **Comparison of sea ice mean state**. It is worth noting that CICE, as most models, has had parameters largely tuned to best represent current state. As a result, the comparison of model with no additional tuning to observations (of sea ice extent and thickness) seems poorly motivated. Would we expect it to improve representation of mean state without tuning of other variables? Additionally, the implementation in a forced, standalone setup is likely to see less response in sea ice state, without the possibility of atmosphere and ocean feedbacks.
  What I think is more interesting is changes to sea ice mean state between models, which can suggest something about how different physics and processes relating to floe size feedback onto other sea ice characteristics. However, this is hard to see in a forced (rather than coupled) model, where feedbacks are limited. I think it is worth focusing on differences in the seasonal cycle and maps between the model which may impact these feedbacks, in the absence of coupled runs. Do these suggest improvement in how ice evolves? In short, I suggest that in the absence of additional, fully-coupled runs, the conclusions should be re-framed.
- I was left confused how what appears to be substantial, meaningful changes in sea ice thickness in Fig. 11 (5-50 cm across much of the Arctic) corresponds to almost no change in volume in Fig. 5. Perhaps it is just an interpretation error on my part, but the presentation needs to be clarified to illuminate whether there are meaningful changes in ice state, or not.

Minor/specific comments:

- P1, L1: It would be helpful to provide a brief introduction to the range of floe ice sizes and key processes (and why it is useful to capture it with a distribution, as is often done for thickness)
- P1, L16: I would suggest that the sentence beginning with "Observations show…" should go before the sentence beginning with "Large-scale…" on L15.
- P1, L36: Is cluster of sea ice into larger floes really a process impacted by floe size, or is it primarily a process in determining floe size?
- P1, L37-38: May be worth considering additionally/alternatively citing Keen et al., 2021, which summarizes CMIP6 sea ice models and show that most use some derivation of CICE or LIM, which all have the same lateral melting parameterization
- P1, L39: Floe size is also not considered in dynamics.
- P2, L5: Why? Need to briefly describe the floe size distribution to justify why a power law is used (i.e., that typically more small floes). Replace "…generally fitted to a…" with "…summarized by fitting to a…"
- P2, L17: Please note to what degree power law does/doesn't fit observations summarized. Is it a manner of convenience, or do observations support its use?
- P2, L23-27: as noted above, it would be helpful to mention these processes earlier in introduction
- P2, L28: It would be helpful to include a transition sentence motivating introduction of brittle facture – that it is missing in most models, and may be important. Perhaps something like what is currently L42-43.
- P3, L3 and L4: replace "represents" with "is in"
- P3, L32: Would be helpful to also introduce the ITD, which is referenced in relation to the prognostic model
- P4, L11: Is there any possibility for ice-ocean feedbacks, such as albedo feedback, in this setup? Please specifically address in the text
- P4, L24: It would be helpful to be more clear in the description of CICE and general model that this is being used in a standalone setup.
- P4, L39: What is the L used in standard implementation? Is it 300 m, as used in lateral melt? Please define.
- P5, L19: In the introduction you present the WIPoFSD before the prognostic model. It would be helpful to be consistent about the order throughout manuscript. I suggest presenting WIPoFSD, then prognostic model, the brittle fracture scheme.
- P6, L31: change to "spatial and temporal scales" or "spatial scale and timescale" or similar
- P7, L1, 11, etc.: I'm not really sure I understand the physical implications of l_var. What is it intended to represent? What are the implications for observational comparisons?
- P7, L36: Is there anything that can be referenced to demonstrate that ERA-Interim wave product is reasonable to use for the Arctic? What is the treatment for waves in sea ice?
- P8, L9: Why 'best'? Unless you plan to show others runs involved in selection process, I suggest use of just 'WIPoFSD' and 'prog' for simplicity and clarity
- P8, L27: Perhaps simply "FSD observations" as title?

- P8, L29-30: I believe since this is included in contributions and acknowledgements, it is not necessary to include funding or names of contributors here.
- P8, L31: replace ‘samples’ with images
- L31: remove “three months”
- L39-40: What is the impact of this (as well as lower cutoff, L43+)? It seems like you could instead include the largest floes in the largest category, which may sometimes show a FSD with the uptick demonstrated by the prognostic model.
- P9, L11: change “is not novel” to “has been used previously”
- P9, L27: Is there an appropriate reference for this statement?
- P10, L8: Perhaps simply “Comparison of sea ice extent and volume” might be a clearer heading
- P10, L1-3: It would be helpful to show some results of when and where the brittle fracture scheme is implemented. Where is it most necessary? What does this suggest about what it corresponds to physically?
- P10, L 18-20: I think it is necessary to clarify here that these runs are done in a standalone setting, and that the possibility for feedbacks in a fully coupled climate model may give different results.
- P10, L23-24: It’s not clear what negative trends in the percent difference suggest. Does this suggest some sort of feedback in model?
- P13, L14-16: Maps show more substantial changes in representation of sea ice state. Do these suggest improvements?
- P14, L34: It may be worth mentioning that this is particularly relevant in a standalone sea ice model, as run here. In a coupled context (for which climate models are often used) the sea ice model is typically a small component of the total cost, and so the additional cost from the FSD is relatively not substantial.
- P15, L6: What is meant by ‘in-ice wave scheme’? Is it more accurate to say that the waves are forced with reanalysis?
- P15, L19: Future work to address the impact on Antarctic sea ice representation and comparison with observations may also be useful
- Figure 1: I find this figure really hard to interpret currently. A few reasons/suggestions… It might be better to use more realistic ‘floe diameter’ bounds, or to remove numbers from y axis, as it currently is hard to interpret these as actual bins. Only 2 examples are needed showing where redistribution is applied and where it isn’t (for example, far left and far right). For one where redistribution is applied, show the new floe size distribution resulting more clearly. It would also be helpful to add lines showing actual density gradient for comparison to dashed purple line.
- Figure 2: I’m not convinced that this is necessarily a “non-physical feature” of the model, as it is simply capturing floes that are potentially beyond the bounds here, and is not reported as such in Roach et al., 2018. A comparison to observations without largest floes removed may be helpful to show if this is ever observed in observations. Additionally, please add a label to this figure demonstrating that it is only for areas of SIC 15-80%

- Figure 3: Are these exact bounds of model areas? If not, a different symbol may better communicate that, as the boxes suggest that this is the exact selection of grid cells.
- Figure 4: Again, not necessary to name co-authors in the text.
- Figure 5: The change of prognostic models being compared is a bit confusing. Is it worth including other prognostic models somewhere as well, to show if/that there is little difference in sea ice state? Also, as the 'prog-best' doesn't include brittle fracture (right?) it would be helpful to note that in the short name for clarity.
- Figure 5-7: This feels like a lot of plots to show for almost no change between any of the models. Can this be simplified to one or two key panels, and then state there is no observable change in others?
- Figure 8: I might suggest to swap these plots around to show both models and same subplot, with top for sea ice extent, bottom for volume. This would then allow to show some comparison in difference of observations from reference. (e.g., Are model changes moving it in the right direction?)
- Figure 9: I am unclear how this figure is different from what is shown in Fig. 6, in terms of the take-aways. How do we know if this is improving the comparison if the scale of change is not comparable to the difference from observations?
- Figure 11: I think this is the most useful and interesting plot! But, I'm quite confused how what appears to be substantial changes in sea ice thickness in A(f) agree with what is in Fig. 5 – where almost no change is observed. Some thoughts: Could the difference in fractional ice area/thickness compared to observations also be shown? It would be helpful to place the effective floe size upfront (at the top) to set it apart from differences, and also make this more clear in the figure caption (meaning, that floe sizes are NOT a difference).
- Figure 12: I'm not sure what to take from these plots, based on the units show. Would it be more helpful to show standard deviation as a percent of mean value? Also, it would be nice to be consistent with the months shown in Fig. 11.
- Table 1: "CPOM-CICE" is not needed in model description, as all are the same. It might be helpful to separate technical details into finer resolution categories, such as brittle fracture (yes/no), # floe size categories; d_min/d_max (where applicable)

References:

Keen, A., Blockley, E., Bailey, D. A., Boldingh Debernard, J., Bushuk, M., Delhaye, S., ... & Wyser, K. (2021). An inter-comparison of the mass budget of the Arctic sea ice in CMIP6 models. *The Cryosphere*, *15*(2), 951-982.