Author Response to:

Christopher Horvat

This is my review of Denton and Timmermans (2021).

Author Comments are in blue.

Here the authors use image processing to obtain floe size distribution (FSD) measurements spanning a scale range from 50 m^2 -5 km^2, and over a wide time period in the Canada Basin. They find both that FSDs obey a power law relationship and that there is a connection between sea ice concentration and power law coefficient. Generally I find this paper to be an interesting contribution to the literature. It is well written and has a comprehensive review of existing work and does a good job of contextualizing the work here. I am also particularly excited that someone has characterized the GFL dataset! I have a few minor comments, and one important one about statistics that needs to be taken into account. The latter is a straightforward application of a method to data already presented in the MS, so it really is likely to be a minor revision, though it may have an impact on the results. So for the authors I'm recommending "major" revisions only to give more time to include that bit of analysis and discussion. Up to these changes, I'm happy to recommend this paper for publication.

Best,

Chris Horvat

——

Thank you, we appreciate your helpful review and overall suggestions related to power-law fitting statistics. We have taken into account each of your comments (inline below) to improve the manuscript.

Main comment.

L30, and generally - It is worth describing the power-law behavior of the FSD as a "hypothesis". As was pointed out by Herman (2010), there actually isn't strong evidence of a power law in most data, but the use of this to describe the distribution is because it is handy for multi-scale distributions.

Thank you for this comment. We now reference this paper and add the following clause to the sentence on lines 35-38 : "...although alternate distributions have been explored (see e.g., Herman, 2010, and the discussion by Stern et al., 2018a)."

As was pointed out in Stern's two 2018 papers, the appearance of power-law behavior can appear spuriously as a result of choosing a particular way of plotting the distribution. This hypothesis can be formally tested, and the mechanism for doing so is described in papers by Clauset (2007,2009,2014), who maintains code to perform fitting, goodness-of-fit tests, and p-value computations at https://aaronclauset.github.io/powerlaws/. When examining multi-scale distributions, one should report the p-value for fits, and the start of a power-law tail. There are many reasons for this!

Thank you, we now report *p*-values and the strict statistical lower-bound on power-law behavior for the power-law fits, a_{min}, following the methods of Clauset et al. (2009) and employing Aaron Clauset's power-law MATLAB toolbox. Please see our later responses.

First, often the intuitive "eye test" for power-law behavior is flawed, and alternative distributions fit better (as examined by Herman (2010)). In studies where this was adamantly required by certain unnamed co-authors, (e.g. Hwang, 2017), few of the obtained distributions actually had power law tails, or only could not reject a power law hypothesis over a small range of size scales. It is worth knowing this!

Thank you for your comment, which we address below. We note that Hwang et al. (2017) report that 77% of their data passed the goodness-of-fit test (p-value ≥ 0.1). Similarly, Stern et al. (2018b) find that "...only 9 out of 116 floe size data sets are rejected as not being power-law distributed in the 2013 MODIS imagery, and only 3 out of 140 floe size data sets are rejected in the 2014 MODIS imagery." They conclude that "in general the FSD in the Beaufort and Chukchi seas follows a power-law distribution for floes from 2 to 30 km in size." Neither of these groups elaborate on the statistically tested range of size scales over which a power law holds.

Second, power-law behavior is fundamentally about the tail of a distribution. Thus issues at the "artifact scale" are not important when performing such tests. This allows you to expand the range of floes into the less certain smaller sizes without actually losing information - if the fit doesn't extend there, it won't be included.

Thank you for the comment. In our case, it is clear a-priori that the floes at or below the "artifact scale" should not be considered in the least-squares fit. We discuss our computation of FSD slopes using MLE in the next response. For those fits, we do not exclude the artifact scale, but we do limit the data to floe sizes larger than the image resolution-limit scale (5 m^2).

Third, there is a mechanism for looking at binned (Virkar and Clauset, 2014) and un-binned data (Clauset et al, 2007,2009). The binning employed here seems to be done to facilitate fitting, but this is not necessary when using the Clauset-style tests. You have the full information on floe areas, so you can directly use that to obtain fits, goodness of fits, and p-values.

We now determine the MLE power-law fit, m_{MLE}, the strict statistical lower-bound on power-law behavior for the MLE fits, a_{min}, and goodness-of-fit for these MLE power-law models indicated by corresponding p-values, from our unbinned floe areas (for all images) following Clauset et al. (2009) and using Clauset's power-law MATLAB toolbox (https://aaronclauset.github.io/powerlaws/). We have added the following section 2.3.2 to the manuscript:

"The maximum likelihood estimator (MLE, see Clauset et al., 2009) can be preferable for the determination of FSD slopes as it does not rely on specifying bins or fitting ranges (Hwang et al., 2017; Stern et al., 2018a; Stern et al., 2018b). In addition to least-squares fitted slopes m and following Clauset et al. (2009), we compute FSD MLE slopes m_{MLE}, and conduct goodness-of-fit tests on these power-law fits, reporting corresponding p-values (where the p-value is the probability that the difference between the model fit and the observed FSD could be due to statistical fluctuations; see Clauset et al., 2009 for a detailed discussion), Table A1. The power-law fit is a plausible model for the FSD if the computed p-value is sufficiently large ($p \geq 0.1$); otherwise, the power-law model must be rejected.

"Clauset et al. (2009) argue that a strict statistical lower bound on power-law behavior must be computed for the observed distribution; we compute these values a_{min}, following their methodology (Table A1). Because m_{MLE} and a_{min} are determined directly from the unbinned floe areas for each image, we compute both over all floe areas (≥ 5 m^2), and do not exclude floes at or below the artifact scale."

We find that:

1. The resulting non-cumulative FSD slopes using Clauset's unbinned MLE method, m_{MLE}, are not significantly different than our least-squares fitted slopes, m, determined over our fitting range (50 m^2 to 5 km^2). For reference, the mean m_{MLE} over our 78 images is -1.77 $\pm$ 0.11, while mean m is -1.79 $\pm$ 0.08 (where uncertainty bounds represent the standard deviation); considering each image, m_{MLE} differs from m by about 3% on average. We now include slopes m_{MLE} in Table A1 and have added the following sentences to the manuscript in Sect. 3.1: "We find no significant difference between slopes m and m_{MLE} (Table A1). The mean m_{MLE} over all images is -1.77 $\pm$ 0.11. Considering each image, m_{MLE} differs from m by about 3% on average".

2. In addition, following Clauset et al. (2009), we find that for all 78 images, a_{min} spans 11 m^2 to 9,808 m^2. We report these in Table A1. We note that for our dataset, the median value of a_{min} is 361 m^2 while the largest floe areas are around 10 to 100 km^2. Even considering the maximum value of a_{min}, this is a large range of floe sizes over which the power-law models are plausible. We add the following sentences to Sect 3.1, "Finally, we find that the strict lower-bound to power-law behavior a_{min} varies considerably over the images (Table A1), spanning around 10 to 10,000 m^2 with a median value of a_{min} of 361 m^2. Considering that the largest floe areas in the images are around 10 to 100 km^2, the range of floe sizes over which the power-law fits apply is large. Values of a_{min} can vary significantly even across images acquired on the same day at the same location (see e.g. Table A1, images 14–15)."

3. We conduct goodness-of-fit tests using the K-S statistic (following Clauset et al., 2009) on the MLE-fitted power-law models (power-law fits with slopes m_{MLE} and lower-bound to the fits a_{min}) to our 78 FSDs, finding that 76% of them pass the test (p-value ≥ 0.1), indicating that the power-law fit is a plausible model to our floe size data. We have added the corresponding column of p-values for each image segmentation to Table A1 and the following sentence to Sect. 3.1: "We find that 76% of the fits pass the goodness-of-fit test with $p \geq 0.1$ (Table A1) meaning that the FSDs can plausibly be power-law distributed."

I would suggest that these tools be applied to the distributions obtained here, and p-values and tail beginnings (x_{min} in the power law toolbox) reported for power-law fits before making statements about the distributional fitting. Without them, it is hard to be convinced rigorously that this hypothesis is not "not true". Including these makes the reporting of PL behavior robust.

Please see our above response to your comments regarding the power-law fits, tail beginnings, and p-values.

Finally, it would be good for the community if your segmentation algorithm was posted publicly and DOId, through e.g. Github/Zenodo. Almost all existing algorithms used in FSD studies are similar in structure, but not able to be used by others. Thus each time a new group wants to compute FSDs they have to reinvent the wheel. Having an available FSD code would benefit many, and ensure your work was properly credited.

We appreciate your suggestion and will share the segmentation algorithm on such a repository. We will note in the final comment period where we have deposited the algorithm.

Specific Comments:

L26 - it is worth looking at the papers of Roach (2018,2019), Zhang et al (2017) who also discussed and theorized the impact of the FSD on both local and model-scale melt partitioning.

Thank you. We now cite these papers in Sect 1.1.

L95 - can you clarify how your method differs from the methodology used by other authors? Is it a similar concept?

Please see our response to Reviewer 1 in this regard.

L198 - I appreciate that you are using floe area to define the FSD - not sure why more authors don't do this. It makes for more accurate comparison to models (which use a fixed radius-area relationship) and is much easier to understand than the MCD. I would even add a statement here or in the discussion to make this point as it is quite helpful from the modeling side.

Thank you for your comment. See our response to Reviewer 1. We have added the following phrase to a sentence in Sect. 2.3: "In the present work, we use floe area because we obtain this directly in the segmentation (and it is directly relatable to floe models)..."

L180 - Have a look at the main comment here. I do not think it is a good idea to compute power law slopes from binned data, as you are sensitive to your bin choices. Still, there is a methodology for computing such slopes (Virkar and Clauset, 2014), which I would argue to employ here if you want to compute them using the binned data. However you have the raw area data, and so don't need to resort to binning. In that case you can use the methodology of Clauset et al (2009) to obtain slopes and p-values for fitting. You will find that the binned PL slopes and the unbanned PL slopes *will* differ.

Thank you. We have done this now. Please see our comments above.

L223 - Again, I would strongly caution to use the Clauset method to obtain slopes. This is discussed in the series of Clauset papers, and later by Stern et al (2018), that fitting straight lines to log-log plots can often fail you in unexpected ways.

Please see our comments above.

L279 - please clarify how you are testing for significance. You are highly sampled in the high SIC range, and weakly so in the low SIC range. What went into the choice of such a significance test, and why? For example, you bin floe areas and then fit them - what guides that choice, but not a similar choice for SIC? Hopefully that points a bit towards why it might be preferable to avoid binning.

Thank you for your comment. We note that the p-value on the linear fit between SIC and m is 3.18×10^{-8} (with the null hypothesis being that there is no relationship between the two), below the significance level 0.01, suggesting that the null hypothesis may be rejected, and the two variables share a statistically significant relationship. The r-squared value of our fit is 0.33, indicating that there is a relatively large degree of variability in m with SIC. However, the statistically significant relationship between the two remains. With regard to nonuniform sampling, we note that a residual plot of the linear fit between SIC and m indicates no clear bias or pattern to the residuals, which indicates that a linear model can be appropriate for the data here.

We have added the following sentence to Sect. 3.3: "Note that there are more sample points in the high SIC range than in the low range, and the linear fit can only explain 33% of the variation (r-squared is 0.33) in m with SIC. However, the linear relationship is statistically significant with a p-value of $O(10^{-8})$ (i.e., < 0.01)."

With regard to avoiding binning for our floe areas, please see our earlier response.

L360 - An extensive discussion of how multiple FSD size regimes can emerge was performed in Horvat and Tziperman (2017) which I will shamelessly plug. It is worth noting the difference in locations may impact the processes that give rise to PL behavior (i.e. the impact of waves leads to more small floes, etc).

Thank you for the useful reference. We have added the following sentence to Sect. 4: "For example, Horvat and Tziperman (2017) use a coupled ice-ocean model to show that increased lateral melt on specific floe sizes and transient oceanic forcing on the ice pack can perturb the FSD behavior from a single power-law at the relevant scale."

References.

Clauset, Young, and Gleditch. On the Frequency of Severe Terrorist Events. 2007

Clauset et al. Power-Law Distributions in Empirical Data. 2009.

Virkar and Clauset. Power-law distributions in binned empirical data. 2014.

Herman. Sea-ice floe-size distribution in the context of spontaneous scaling emergence in stochastic systems. 2010.

Horvat and Tziperman. The evolution of scaling laws in the sea ice floe size distribution. 2017.

Roach et al. An emergent sea ice floe size distribution in a global coupled ocean--sea ice model. 2018.

Roach et al. Advances in Modeling Interactions Between Sea Ice and Ocean Surface Waves. 2019.

Zhang et al. Sea ice floe size distribution in the marginal ice zone: Theory and numerical experiments. 2016.