Response to Referee #1

Dear Referee,

We would like to thank you for reviewing our manuscript and giving valuable comments and suggestions. Please find our responses to your comments in blue text and revised manuscript in red text below.

The major change in this revision is:

We give a clarification of the two metrics for model evaluation that we have already been used in our analysis. The definition of $p(r)dr$ is the perimeter of floes per unit ice area with radius between r and $r + dr$

$$P = \int_{r_{min}}^{r_{max}} p(r)dr$$

P is the perimeter of floes per unit area with r varying between minimum floe radius r_{min} and maximum floe radius r_{max} in the distribution.

The metrics evaluated in this study is not only P, a value of the integral of perimeter density per unit sea ice area over all floe sizes (e.g., the value given at Line 227-236 on Page 8) but also *p(r)*. P captures the most useful information about FSD shape from the perspective of modelling into a singular value, making it easier to present comparisons between observations and models. Whilst information about FSD shape is lost when calculating P, we also apply $p(r)$ to evaluate the model performance (e.g., Figures 2e-2k, 4e-4f). In our manuscript, $p(r)$ is shown as the perimeter density distribution for floe size categories (perimeter of floes in floe size category *i* per unit bin width per unit sea ice area). The text and figures have been revised when we show the results about $p(r)$ in the revised manuscript. It seems like the confusion between P and p(r) has resulted in a misunderstanding, so hopefully addressing this point should resolve most of your concerns around this point.

Best regards,

Yanan Wang and co-authors

This review is from Referee #1 of the original manuscript. The authors have made changes in the right direction, but there are still issues to address.

Each model produces a sea-ice floe size distribution (FSD) from which the sea-ice floe perimeter density (P) can be calculated (assuming a shape parameter, gamma). Each satellite image also yields a FSD and P. The authors compare P from models with P from satellite images, which is a perfectly proper and legitimate way to assess the skill of the models. But note that the relationship between P and FSD is a one-way street: given the FSD and shape parameter gamma, one can calculate P; but given P, one can say nothing about the FSD. The results of the analysis show that in general there is very poor agreement between modelled P and observed P, and that's fine -- it simply means that the models are not able to reproduce the observed P. But even if the agreement were good, it would say nothing about the modelled FSDs. That point should be acknowledged or conveyed somewhere.

Based on our clarification of $p(r)$ and P given above, $p(r)$ and $n(r)$ can be related by $p(r) = \frac{2\gamma r n(r)}{c_{ice}}$. Given $n(r)$, gamma and SIC, we can also calculate $p(r)$ and vice versa.

This definition has been given in the revised manuscript. Please see Lines 190–194 on Page 7 in the revised manuscript.

For the evaluation of FSD models, we consider the perimeter density distribution, $p(r)dr$ (units: km km^{-2}), which is defined here as the perimeter of floes per unit ice area with radius between r and $r + dr$,

$$P = \int_{r_{min}}^{r_{max}} p(r)dr, \quad (1)$$

P is the perimeter per unit area of floes between r_{min} and r_{max} in radius.

The perimeter density P is not a proxy for the FSD. A proxy is "a measurement of one physical quantity that is used as an indicator of the value of another" or equivalently "a measured variable used to infer the value of a variable of interest" (to quote two dictionary definitions). Now consider these examples:

1. All floes are circular with radius r0. The FSD is the Dirac delta function, f(r) = delta(r - r0), and P = 2/r0. Suppose r0 = 20 meters = 0.02 km. Then P = 100/km.
2. All floes are circular with uniform distribution on [0,L]. The FSD is f(r) = 1/L and P = 3/L. Suppose L = 30 meters = 0.03 km. Then P = 100/km.
3. All floes are circular with exponential distribution. The FSD is f(r) = (1/lambda)*exp(-r/lambda) and P = 1/lambda. Suppose lambda = 10 meters = 0.01 km. Then P = 100/km.

In all three examples, P = 100/km. If one is simply given that P = 100/km, it is not possible to "go backwards" and find the FSD. The FSD could be anything. Any distribution with a free parameter can be made to have P = 100/km by proper choice of the parameter. P is not a proxy for FSD.

We have clarified the terms P and p(r) in the updated, which we believe should address the above concerns. Based on the relationship between $p(r)$ and $n(r)$, $p(r) = \frac{2\gamma r n(r)}{c_{ice}}$, if the function of n(r) is different, p(r) is different. Yes, for these examples, the integral of p(r) over all floe sizes, i.e., P, are all 100/km, but the $p(r)$ is different.

Besides, we have also replaced the word 'proxy'. Please see Lines 195–198 on Page 7 in the revised manuscript.

P is often used as a way to capture useful information about the FSD using a singular value and the concept of an overall perimeter density to characterize the FSD has been *used in several previous observational studies (e.g., Perovich, 2002; Arntsen et al., 2015),*

The authors misinterpret the work of Perovich and Jones (2014). At lines 237-238 of the marked-up revised manuscript, the authors write, "P is a useful proxy for FSD and widely used in previous observational studies (e.g. Perovich, 2002; Perovich and Jones, 2014; Arntsen et al., 2015)." In Perovich and Jones (2014) they ASSUME that the FSD follows a power law, and then use the observed P to determine the power-law exponent. In other words, P is not used as a proxy for the FSD -- the FSD is already fixed as a power law. P is used for finding the exponent.

Thanks for pointing out this oversight. Yes, P is not used as a proxy for the FSD in Perovich and Jones (2014). We have deleted the citation of this paper in this sentence (Line 195–198 on Page 7).

P is often used as a way to capture useful information about the FSD using a singular value and the concept of an overall perimeter density to characterize the FSD has been used in several previous observational studies (e.g., Perovich, 2002; Arntsen et al., 2015),

This is what the authors wrote in their response to my comment about perimeter density (PD) and FSD in the original manuscript:

"1) The same PD corresponding to different number FSDs or the same FSD corresponding to different PDs does not mean there is no connection between PD and FSD. In previous studies, Rothrock and Thorndike (1984) first defined the FSD as both number density n(r) and area fraction f(r). These two definitions are related by f(r)=gamma * r^2 * n(r). Now, if we consider a set of circular floes and a set of elliptical floes with semi-major axis "a" and semi-minor axis "b" with the same number FSD n(r) in a given region. But assume that pi*a*b NE pi*r^2. In this situation, although these two groups of floes have the same n(r), their areal FSD f(r) is different. Similarly, if we consider 5 circular floes with the same radius of 10 m and 100 elliptical floes with semi-major axis "a=1 m" and semi-minor axis "b = 5 m" in a given region. Then we can get the same areal FSD, different number FSDs. Even the traditional FSD concepts n(r) and f(r) still shows the same situation that there is not always a unique 1:1 relationship between them. Similarly, it is not abundant evidence to prove that PD is not a metric related to FSD by showing the same PD corresponding to different number FSDs or the same FSD corresponding to different PDs."

Let's analyze some of the statements in the paragraph above.

1. The equation "f(r) = gamma * r^2 * n(r)" is presented but then later "Even the traditional FSD concepts n(r) and f(r) still shows the same situation that there is not

always a unique 1:1 relationship between them." But the equation given by the authors shows that once the constant gamma is chosen, there is indeed a 1:1 relationship between n(r) and f(r).

The reason that we mention "*there is not always a unique 1:1 relationship between them*" is because of that most examples in your previous comments are related to the variable shape of floe, i.e., γ is not the same for the two groups of circle and elliptical floes in these examples. However, we also responded that "*as pointing out by previous study, the ratios between floe size properties is generally constant and almost independent of floe size (e.g., Rothrock and Thorndike, 1984; Toyota et al., 2006). Besides, it is common in FSD studies to make assumptions about floe shape, e.g., FSD models assume a fixed shape parameter γ. This should not be a cause for concern.* "

Similarly, we have clarified the terms P and p(r) in the updated manuscript, which we believe should address the above concerns. Regarding the definitions P(r) and p(r) given in the revised manuscript, $p(r) = \frac{2\gamma r n(r)}{c_{ice}}$. Once the constant gamma and SIC are given, there is indeed a 1:1 relationship between n(r) and p(r) as well.

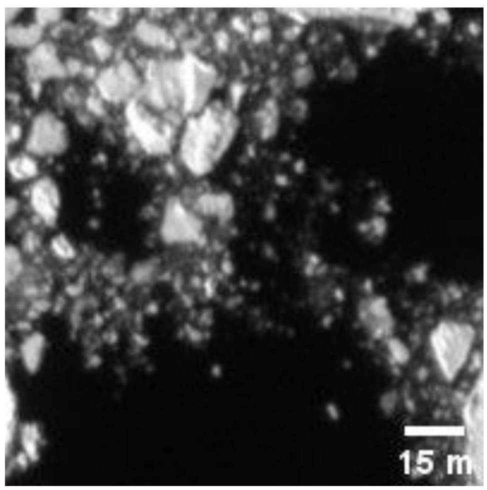
2. "if we consider 5 circular floes with the same radius of 10 m and 100 elliptical floes with semi-major axis a=1 m and semi-minor axis b = 5 m in a given region. Then we can get the same areal FSD, different number FSDs." In the first case, the 5 circular floes of radius 10 m have a total area of 500*pi. In the second case, the 100 elliptical floes also have a total area of 500*pi. Yes, the total area is the same in both cases. And yes, the number of floes in the first case (5) is different than the number of floes in the second case (100). But the authors are mistaken or confused in their use of "FSD" in these examples. Since all the circular floes have radius 10 m, the FSD is n(r) = delta(r-10). Since all the elliptical floes have a=1 m, the FSD is n(a) = delta(a-1). The FSDs are all delta functions. The authors don't seem to understand the difference between "area" and "areal FSD" nor between "number" and "number FSD". This is troubling.

As mentioned above, the metrics that we compared in this manuscript is not only P (e.g., the value given at Line 227-237 on Page 8) but also p(r) (e.g., Figures 2e-2k, 4e-4f). p(r) is perimeter density distribution rather than perimeter, which is also a function of floe size *r*.

Image Resolution.
The models overestimate P for floes with r < 15 m (Fig. 2). The observed P is based on images with pixel size 1 meter. The authors analyze 5 images with pixel size 0.5 meters, resulting in larger P for small floes (Fig. 4). The authors conclude that inadequate image resolution may be one of the causes of the discrepancy between modelled and observed P for small floes.
Line 23: "the image resolution is not sufficient to detect small floes"

Lines 586-587: "The causes of such differences include ... the limitations within the observations such as the image resolution"
Lines 632-633: "It requires much higher resolution images (e.g., aerial photographs) and further research in the future to properly investigate the effects of the image resolution"
Lines 662-663: "the ability to resolve the shape of the FSD for small floes remains constrained by the limited resolution of satellite images"

During late spring and summer, there is a large number of floes with radii smaller than 15 m. During processing the images, even in the 0.5 m-resolution images, we noticed there are several groups of crowded small floes that remain unresolved. Please see the following example (100 x100 m) from the WV image (δ = 0.5 m) on 5 June 2013 at 84.9°N, 0.1°E (Fram Strait). We do not know how much additional change we would see in P and p(r) if we had access to imagery at even higher resolutions. The limited size of the change from 1 m to 0.5 m compared to the size of the difference between models and observations suggests that image resolution cannot fully explain this difference. The image resolution could be one of the contributors but not the main contributor.



A few comments:
1. At lines 600-605, the authors admit that the increase in P in going from 1-m images to 0.5-m images "is still far too small to explain the difference between the observations and model outputs." Nevertheless, the next sentence says: "This suggests that the image resolution could be one of the contributors to the overestimation of modelled P for small floes, but it is still inconclusive whether the limited image resolution is the main contributor or other factors such as model parameterisations contribute to the difference." Really? Can't we conclude that image resolution is NOT the main contributor, given that halving the pixel size from 1 meter to 0.5 meters gives a change in P that is "far too small to explain the difference between observations and model outputs"?

Please see the example given in our response above. We do not know how much additional change we would see in P and p(r) if we had access to imagery at higher resolutions. What we would like to explain is that image resolution could be one of the contributors but not the main contributor. These sentences have been revised. Please see Lines 378–382 on Page 13 in the revised manuscript.

However, this difference is still too small to explain the difference between the observations and model outputs, varying between 20.42 km km^{-2} and 218.95 km km^{-2} in Figs 2e–2k (See Table S1). We do not know how much additional change we would see in P and $p(r)$ if we had access to imagery at even higher resolutions. This suggests that based on the recent satellite imagery, the image resolution could be one of the contributors but is not the main contributor to the overestimation of modelled p of small floes.

2. Judging by Figure 2, the models simulate floe sizes down to about r=3 meters. The authors suggest that the image resolution of 1 meter (or 0.5 meters) is not sufficient (see quotes above). Apparently centimeter-scale imagery is needed (e.g. from aerial photographs). Wouldn't that result in a mismatch in the scale of comparison, if images resolved floe sizes of r=0.1 meters but models only resolved r=3 meters?

As outlined in Roach et al. (2018a) and Bateson et al. (2022), 12 Gaussian-spaced floe size categories are applied in FSDv2-WAVE and CPOM-FSD simulations. In the smallest floe size category, floes with radii between 0.07-5.31m. In Figure 2, 2.69 metres is the midpoint of this bin rather than the smallest floe size simulated in models. The minimum floe size for the prognostic models is technically 0.13 m. Additionally, the image resolution of 1 m does not equal the smallest size of derived floe of 1 m. The distance between floes can also influence the separation of floes from images, especially when during late spring and summer. Please see the example of unresolved floes in 0.5m-resolution image given in our response above.

3. In order to explain or account for the difference in modelled P vs. observed P for small floes, the authors look for a potential shortcoming in the data (its resolution) rather than a potential shortcoming in the models. This seems like the wrong approach. In my opinion, 1-meter (or 0.5-meter) image resolution is good enough! I'm not convinced that we need centimeter-scale imagery to properly characterize the FSD and P.

Our results in Section 4.2 indicate that the minimum floe radius that could be retrieved accurately from the 1-m images was 15 m. However, during late spring and summer, there is a large number of floes with radii smaller than 15 m. Please see the example given in our response above. During processing the images, even the 0.5 m-resolution images, we noticed there are several groups of crowded small floes that cannot be separated. We do not know how much additional change we would see in P if we had access to imagery at even higher resolutions. But based on the comparison of 1-m to 0.5-m resolution imagery, we do not believe that this effect is the main contribution to the differences between models and observations. This discussion in the paper is not to assume that the problem is with observations, but to rule this out so that we can reach some conclusions on model performance

Lines 448-450. "floe welding rate is set to be proportional to the square of SIC in the two prognostic models ... In this section, we present the model-observation comparison results for SIC to validate floe welding for the prognostic models."
The model-observation comparison of SIC does not validate the floe welding parameterization. There are no observations of floe welding presented in this paper. The comparison of SIC validates SIC, not floe welding.

Yes, we agree that floe welding is not validated directly in this study. Roach et al. (2018b) provide observational evidence that floe welding correlates well with sea ice concentration, which is used to give this floe welding parameterization, e.g., the floe welding rate is set to be proportional to the square of SIC in the two prognostic models. This result just provides some possibilities of underestimated SIC leading to the decrease of P through floe welding in the two prognostic models. We have cited the observational evidence of the correlation between floe welding and SIC in the revised manuscript. Please see Lines 276–280 on Pages 9–10 in the revised manuscript.

The FSD models considered in this study include parameterisations with dependencies on SIC. For example, the floe welding rate is set to be proportional to the square of SIC in the two prognostic models, FSDv2-WAVE and CPOM-FSD based on observations that the welding rate correlates with SIC (Roach et al., 2018b). It is therefore useful to evaluate how well the models simulate the observed SIC and consider the extent to which errors in the simulated SIC could explain the differences between models and observations in simulating floe perimeter density.

Lines 479-481. "FSDv2-WAVE slightly underestimates SIC by 2-4% compared to the observations in the MIZ (SIC<80%). CPOM-FSD strongly underestimate the SIC by 13%-15% in the MIZ compared to the observations. This difference can be attributed to different atmospheric forcing that is used in the models (Schroder et al., 2019)." I don't see any evidence that the difference in SIC can be attributed to the different atmospheric forcing (NCEP-2 vs JRA55b) used in the models. This sentence simply makes a declaration without any further justification. I looked up the Schroder reference and it uses NCEP-2 but not JRA55b, so I don't see how it supports the claim being made here.

Schroder et al. (2019) didn't analyse the effects of specific atmospheric conditions on Arctic SIC. Schroder et al. (2019) suggested there is a strong influence of the spring and summer atmospheric conditions on the simulated summer sea ice extent in CICE, so the different atmospheric forcing used (NCEP-2 and JRA55b in our study) can be expected to be a significant factor in explaining the difference in SIC between FSDv2-WAVE and CPOM-FSD. This sentence has been revised at Lines 287-288 on Page 10.

This difference in SIC between FSDv2-WAVE and CPOM-FSD can be attributed to different atmospheric forcing that is used in the models (Schröder et al., 2019).

Lines 483-484. "The underestimated SIC from the two prognostic models may be a contributor to the underpredicted floe welding rate during spring and early summer." Since the floe welding rate is proportional to the square of SIC in the two prognostic models (line 448), underestimation of SIC translates directly into underpredicted floe welding rate in the models. It's not "may be a contributor to" but rather "is the cause of".

Thanks, we have revised this sentence. Please see Lines 289–290 on Page 10 in the revised manuscript.

The underestimated SIC from the two prognostic models will result in a too small floe welding rate during spring and early summer.

Incidentally, we don't know if the floe welding rate is, in reality, proportional to SIC-squared. That has not been validated in this paper.

As in our response above, Roach et al. (2018b) provide observational evidence that floe welding correlates well with sea ice concentration, which is used to give this floe welding parameterization. We have also included this information in our revised manuscript. Please see Lines 276–280 on Pages 9–10 in the revised manuscript.

The FSD models considered in this study include parameterisations with dependencies on sea ice concentration. For example, the floe welding rate is set to be proportional to the square of SIC in the two prognostic models, FSDv2-WAVE and CPOM-FSD based on observations that the welding rate correlates with SIC (Roach et al., 2018b). It is therefore useful to evaluate how well the models simulate the observed SIC and consider the extent to which errors in the simulated SIC could explain the differences between models and observations in simulating floe perimeter density.

Lines 484-485. "A negative bias in spring SIC shown in the prognostic models may partially explain the overestimation of P especially for small floes (Fig. 2)." First, the overestimation of P is ONLY for small floes (Fig. 2) so it's misleading to say "especially for small floes."

Based on our updated definition of P and p(r), we have revised these sentences as follows. Please see Lines 290–292 on Page 10 in the revised manuscript.

A negative bias in spring SIC shown in the prognostic models may partially explain the overestimation of P, and in particular the overestimation of $p(r)$ for small floes and the underestimation of $p(r)$ for large floes (Fig. 2).

Second, a negative bias in SIC could be due to too few small floes or too few large floes. Simply knowing that SIC is too low does not automatically imply an overestimation or underestimation of P.

Regarding your second comment, again, Roach et al. (2018b) provide observational evidence that floe welding correlates well with sea ice concentration, which is used to give this floe welding parameterization. Bateson (2021) has assessed the effects of floe welding on the p(r) in the CPOM-FSD model, suggesting that floe welding occurring in spring can influence the p(r) in summer. A low ice concentration reduces the floe welding during spring and consequently results in an initial over-fragmented state in early summer. Therefore, a negative bias in spring SIC shown in the prognostic models can partially explain the large value of p(r) for small floes and the small value of p(r) for larger floes in the two prognostic models (Fig. 2). We have cited the observational evidence of the correlation between floe welding and SIC (Lines 276–280 on Pages 9–10) and the effects of floe welding on p(r) for small and large floes in the revised manuscript (Lines 393–396 on Page 13).

The FSD models considered in this study include parameterisations with dependencies on sea ice concentration. For example, the floe welding rate is set to be proportional to the square of SIC in the two prognostic models, FSDv2-WAVE and CPOM-FSD based on observations that the welding rate correlates with SIC (Roach et al., 2018b). It is therefore useful to evaluate how well the models simulate the observed SIC and consider the extent to which errors in the simulated SIC could explain the differences between models and observations in simulating floe perimeter density.

A low ice concentration reduces the floe welding during spring and consequently results in an initial over-fragmented state in early summer. Therefore, a negative bias in spring SIC shown in the prognostic models can partially explain the overestimation of $\mathrm{p(r)}$ for small floes and the underestimation of $\mathrm{p(r)}$ for larger floes in the two prognostic models (Fig. 2).

Lines 652-654. "we examined the P in the northern regions where wave-induced breakup is negligible. In these regions, most modelled P match our observations better." There are no observations in the northern CS region.

The reasons of this comparison do not only depend on whether there are observations in the northern CS region or not. As we mentioned in the previous response, "*In the northern CS, the change in P arising from all FSD evolution processes is almost zero in the two prognostic models during our research period. For checking the P in these regions, we recognize the P value in the northern region is the value at the end of early spring or the beginning of our research period. This comparison could help us determine whether the differences between observations and models arise from FSD evolution processes in summer (lateral melt and wave fracture) or in spring.* "

Figure 6. The caption refers to changes in FSD but the figure shows changes in P.

Thanks. We have corrected the caption of Figure 6.

Figure 6: Monthly changes of ***P*** *simulated by the two prognostic models over the period May to July during 2000–2014. (a) Change of FSD arising from lateral melt for FSDv2-WAVE in the Chukchi Sea. (b) is same as (a) but for wave induced* ***P*** *change. (c) and (d) are same as (a)and (b) but in the Fram Strait. (e)–(h) is same as (a)–(d) but for CPOM-FSD. The blue and red box in (a) and (c) show the northern and southern region of the two study regions. Black dots indicate the location of observations in the study regions.*

Table 2. The caption refers to three models but the table lists only two models.

Thanks. We have corrected the caption of Table 2 in the revised manuscript.

Table 2. Statistical summary for the two prognostic FSD models against the NSIDC SIC and ASI SIC. NSIDC and ASI SIC data used for the comparison are between April and August for the analysis period of 2000–2014.

Response to Referee #2

Dear Dr Montiel,

We would like to thank you for reviewing our manuscript and giving valuable feedbacks. Please find our responses to your comments in blue text and revised manuscript in red text below.

Best regards,

Yanan Wang and co-authors

I have now gone through the revised manuscript and the authors' responses and am happy to recommend that the manuscript is accepted.

1. I only have one comment, which the authors and/or editor may want to take into consideration. From the viewpoint of the reader, I still think that having a discussion about model differences at the start of section 3.2 before the models are fully described does not flow well.

We have revised Section 3.2 in the revised manuscript (Lines 130–184 on Pages 5-6).

3.2 Sea ice models with floe size distribution

In this study, three FSD models are evaluated. An overview of the configuration of these three FSD models is given in Table 1. In Sects. 2.4.2–2.4.4, we will briefly introduce the major differences between the three models in simulating FSD related processes.

3.2.1 FSDv2-WAVE model

FSDv2-WAVE model is based on a sub-grid scale floe size and thickness distribution (FSTD) model by Horvat and Tziperman (2015, 2017).

……

3.2.3 WIPoFSD model

WIPoFSD is a diagnostic power law FSD model incorporated into the CPOM-CICE (Bateson et al., 2020, 2022) and has the same horizontal grid and run period as CPOM-FSD.

2. In addition, I suggest that the authors add the reference Mokus and Montiel (2022, The Cryosphere) in line 376, as that paper really complements the Montiel and Mokus (2022, Phil Trans A) paper and these should be taken in tandem.

We have added in the revised manuscript (Lines 385–387 on Page 13).

Other statistical models, e.g., log-normal distribution, is better to describe the FSDs rather than power laws (Montiel and Mokus, 2022; Mokus and Montiel, 2022).

3. Finally, the year of the reference to "Meylan et al (2019)" in line 150 should be 2021.

We have revised in the revised manuscript (Lines 142-143 on Page 5).

Attenuation of wave energy in the MIZ is modelled using multiple wave scattering theory developed by Meylan et al. (2021).