Anonymous Referee #1

General comments

The manuscript by Park et al presents a machine learning technique to provide operational sea ice charts. The use of machine learning is interesting from the operational side as it enables processing of large volumes of data in a consistent manner. Such sea ice charts are of significant use for the both scientific community as well as the general community.
The manuscript is well structured, though would benefit from a spell and grammar check, e.g. Sentinel is misspelled several times. The authors claim that the method is free from subjective judgements though it requires the input of manually derived open water vs sea ice charts.

First of all, we would like to thank the reviewer for the positive evaluation and providing important comments. We hope that all your concerns will be cleared after reading our responses and modifications made to the manuscript. Please find below our answers (in green) and modifications (deleted in red and added in blue) to your comments/suggestions/questions.

Specific comments

The authors claim that by using operationally provide sea ice charts they are avoiding subjective decisions in the sea ice classification training data. To my knowledge the sea ice charts provided by the US National Ice Center are based on exactly such subjective decisions and are made to the best ability of the excellent sea ice experts working there. Please clarify how the training and validation data used here are not subjected to such decisions. The manually derived open water vs sea ice maps used as input training data could also be viewed as subjectively derived data.

As the reviewer pointed, ice charting cannot be free from subjective decisions but it requires best knowledges from credible experts like national/international ice services rather than from anonymous individuals. The aim of this study is to address such an issue; however, we admit that the manual selection of training/validation dataset causes another subjectivity issue. Regarding the concern on subjective judgement in selecting training dataset, the open water vs sea ice charts are not derived manually; the boundary between open water and sea ice is extracted from the electronic ice chart. The manual part is to check if the SAR image show high image contrast at the position of the extracted ice-water boundary. Although the visual inspection along the ice-water boundary overlaid on SAR image is also not completely free from subjective judgements, it requires much less effort and expertise compared to drawing manual ice chart from the SAR image itself. We apologize for this confusion; we omitted an essential description. In the revised manuscript, we added detailed explanations about the procedure on selecting training dataset.

[Sections 2.2.6]

…

~~To automate image selection for training, a good ice/water classifier for SAR image is needed.~~ In order to automate image selection, the ice edges in SAR images needs to be identified first. Since ~~even such a simple binary~~ an ice/water classifier has not been well developed yet for Sentinel-1, the image selection procedure has to be done manually in the beginning. However once a classifier is generated with high accuracy, it can be used to automate the procedure, then the whole process in the proposed scheme will be fully automated. This is why the proposed algorithm is named “semi-” automated for now. Nevertheless, the manual selection is done by visual inspection of ice-water boundaries overlaid on SAR images. The ice-water boundary can be extracted easily from the reprojected ice chart by selecting the pixel borders of open water class. Then the SAR backscattering image contrasts across the ice-water boundaries are examined both in HH- and HV-polarization because the image contrast between ice-water is larger in HV but smooth level ice is better recognizable in HH.

It is unclear if only the sea ice parts of the images were incidence angle corrected using the sea ice estimated slope or if the whole image was corrected using the slopes derived for the sea ice part of the images. Please clarify. If the incidence angle slope derived for the sea ice were used over the entire image how would this affect the open water areas of the image? How are the slopes presented here derived?

The whole image was corrected using the slopes derived for the sea ice regardless of the surface type. Since this correction is prior to the classification, sea ice type-specific correction cannot be made in this stage. Nevertheless, the bulk slope correction has known to be effective in literature (Zakhvatkina et al., 2013; Zakhvatkina et al., 2017). We added the following explanations and references to Section 2.2.3:

[Sections 2.2.3]

…

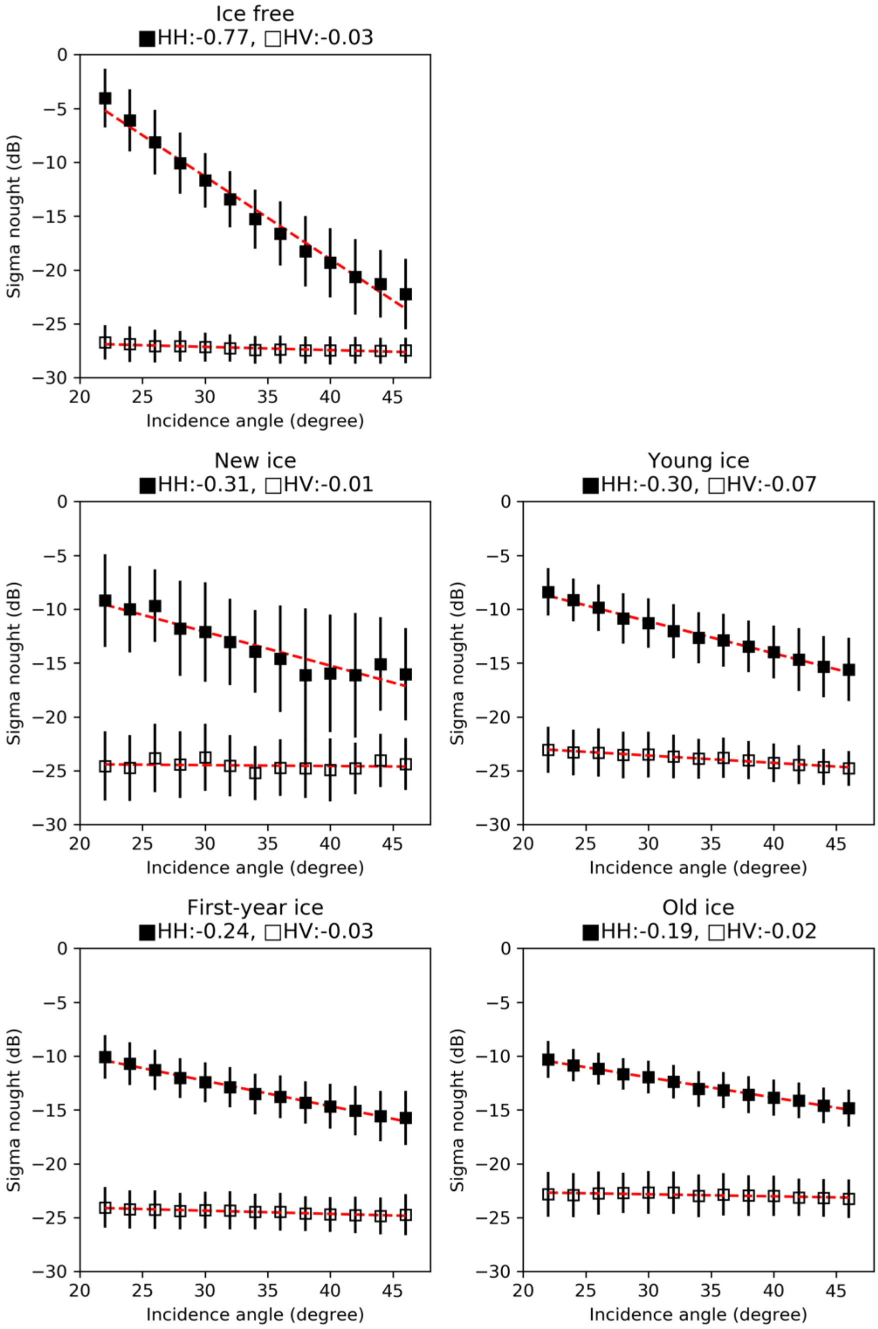
Figure 4. shows two-dimensional histograms of incidence angle versus sigma nought for sea ice pixels in HH and HV polarization channels from Sentinel-1 data collected over sea ice and open water in the study area in winter 2018. From the Sentinel-1 dataset described in Section 2.1, sea ice pixels were extracted by using daily global sea ice edge products available from the EUMETSAT Ocean and Sea Ice Satellite Application Facilities (OSISAF). For HH polarization, the estimated slope was -0.200 dB/degree, which is slightly different from the estimation of the first-year ice (-0.24 dB/degree) in Mäkynen and Karvonen (2017) and in between the estimations for first-year ice (-0.22 dB/degree) and multi-year ice (-0.16 dB/degree) in Mahmud et al. (2018). For HV polarization, the estimated slope was only -0.025 dB/degree, which is much lower than the estimation in Mäkynen and Karvonen (2017), however, it is in line with the estimations from RADARSAT-2 (Leigh et al., 2014; Liu et al., 2015). We compensate the incidence angle dependency using the estimated slopes, referencing to the nominal scene center angle of 34.5 degree. Although the incidence angle dependence changes with ice type and radar frequency (Mahmud et al., 2018), the compensation is done for all pixels in the image using a single value of mean slope because the ice types are not identified in this stage. Open water areas of the image are also affected; however, the correction is also beneficial since the incidence angle dependence for open water is stronger (-0.65 dB/degree for wind velocity of 5 m/s, computed from CMOD5 C-band geophysical model function in Hersbach et al., 2007), thus the corrected image has less incidence angle dependence.

Mahmud, M. S., Geldsetzer, T., Howell, S. E. L., Yackel, J. J., Nandan, V., and Scharien, R. K.: Incidence angle dependence of HH-polarized C- and L-band wintertime backscatter over Arctic sea ice, IEEE T. Geosci. Remote, 56(11), 6686-6698, doi:10.1109/TGRS.2018.2841343, 2018.

Hersbach, H., Stoffelen, A., and de Haan, S.: An improved C-band scatterometer ocean geophysical model function: CMOD5, J. Geophys. Res., 112, C03006, doi:10.1029/2006JC003743, 2007.

Mahmud et al 2018 showed that different sea ice types have different incidence angle dependent slopes. Have you considered using a sea ice type dependent slope factor? Moreover, how does the work by Mahmud et al 2018 fit in with the incidence angle dependencies presented here?

Typically, the slope for open water is higher than that for sea ice, thus the correction works in a way reducing the difference in sigma naught for open water as well. As the review pointed, different sea ice types have different incidence angle dependent slopes; however again, ice type-specific correction prior to ice type classification is controversial. Although estimating ice type dependent slope is not a part of this manuscript, we provide the values derived from the training/validation dataset for the review purpose only.



How do you define "good match" (P7 R3)? Temporal overlap? Spatial overlap? How was this manual selection of images performed? How were the open water vs sea ice charts, that are used as initial input into the classifier, derived?

Both temporal and spatial overlaps are important. Since the SAR image itself is potentially one of the sources for ice charting at the ice services, some images spatially match well with the shape of the polygons in ice chart. Temporal window length of 3 days from the publication date of ice chart was used for squeezing the number of images to make decisions of use/discard for training. We added a paragraph that explains processing details in the revised manuscript as follows:

[Sections 2.2.6]

…

To train an ice type classifier, a set of collocated SAR images and ice charts is required. After the preprocessing of ice chart including reprojection into the SAR image geometry, only the samples with spatially and temporally good match should be fed to the training phase. Image selection is trivial but not easy to automate. Since the weekly ice chart is made partly based on the SAR images acquired in the past 3 days from the date of publication, the ice edges in some images match well with those in the ice chart. ~~To automate image selection for training, a good ice/water classifier for SAR image is needed.~~ In order to automate image selection, the ice edges in SAR images needs to be identified first. Since ~~even such a simple binary~~ an ice/water classifier has not been well developed yet for Sentinel-1, the image selection procedure has to be done manually in the beginning. However once a classifier is generated with high accuracy, it can be used to automate the procedure, then the whole process in the proposed scheme will be fully automated. This is why the proposed algorithm is named "semi-" automated for now. Nevertheless, the manual selection is done by visual inspection of ice-water boundaries overlaid on SAR images. The ice-water boundary can be extracted easily from the reprojected ice chart by selecting the pixel borders of open water class. Then the SAR backscattering image contrasts across the ice-water boundaries are examined both in HH- and HV-polarization because the image contrast between ice-water is larger in HV but smooth level ice is better recognizable in HH.

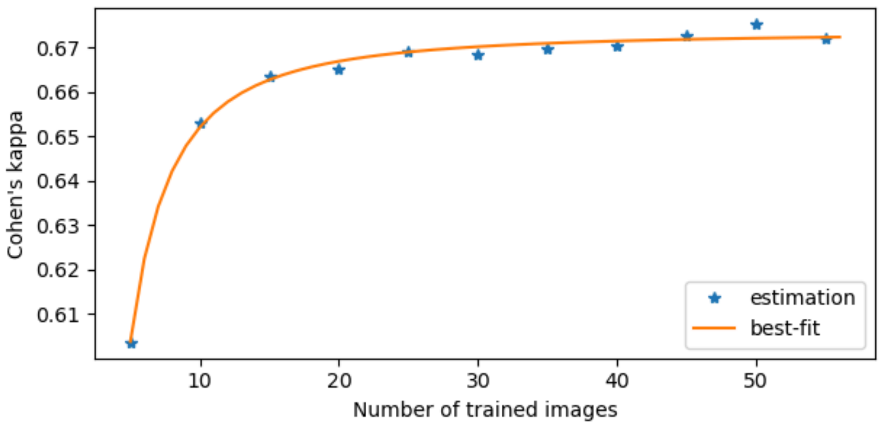
After the image selection, the samples in the selected images are split randomly into training and test dataset with a ratio of 7:3. For training dataset, further data selection is made by excluding the samples residing close to the polygon boundaries. This is to account for possible mismatch due to various reasons (e.g., ice drift, vector mapping error, image geocoding error, etc.). In this study, only the data from pixels more than 3 km away from the polygon boundaries was fed into the training process. Once the hyperparameter optimization is done, the RF classifier is trained for the training dataset. The trained classifier is then applied to the test dataset. We use confusion matrix for performance evaluation. The validation is done in the same way but using a completely independent dataset. The 2018 data was used to run the training phase. Among 958 images in total, we selected 57 images of which ice edges match well with the collocated ice chart. From the selected images, 6.4 million samples covering open water and sea ice were divided into training and test dataset.

Are the open water areas separated from the sea ice areas at this stage of the classification process? Or are they classified at the same time as the sea ice types?

At this stage, no classifier is introduced. The open water is classified at the same time as the sea ice types in the later stage.

Given that one possible reason for the low accuracy in 2019 were stated to be insufficient training data (P9 R9-11), have you tested how the classification improves/remains the same if additional training data is added? Such an assessment would add strength to the accuracy of the method presented here.

We tested the changes in classification accuracy with varying number of trained images. As the results shown below indicates, the accuracy increases rapidly until adding 15 images, but the improvement was not great after adding 20-25 images.



By following the test result, we revised the manuscript as below.

[Sections 2.2.6]

…

However, the accuracy decrease from 2018 data to 2019 data was in similar level to the case of five-class classification, and this could have been caused by ~~insufficient training of classifier and/or~~ inconsistent labeling in the reference ice chart.

Daily ice charts using Sentinel-1 data covering at least part of the study areas used here are provided by the Norway Ice Service. A comparison with their ice charts when they overlap, spatially and temporally, would have been beneficial and added strength to the accuracy assessment. Partially as it would have provided daily instead of weekly ice charts to compare to.

As far as we understand, the daily ice chart serviced by Met.no does not provide ice types; it provides ice concentration instead.

Would your method work also outside the winter season? Has it been tested for other seasons? A majority of the shipping industry is dependent on sea ice charts year-round and a consistent method employed year-round is therefore beneficial.

The method itself would work for other seasons if multiple classifiers are trained for each of the seasons. It is challenging to develop a single universal classifier that works for all seasons. Although we tried to make a classifier adapt to seasonal changes by including day of year as a feature (FC3), the result was not promising.

Day of year might not correspond to the same temperature, fluxes and weather regimes. Have you considered using a weather variant input parameter instead of day of year? Such a parameter might be more suitable to capture the seasonality within the scenes.

No, we didn't. As sea ice drifts continuously, weather variant information for each of the ice floes in the SAR image at the image acquisition time needs to be calculated and joined, and this may require complex and rigorous works. In order to simplify the problem, we used day of year only, but certainly other parameters need to be tested in the follow up paper. We are very interested in combining sea ice drift with ice type-specific texture changes in our future study. We added a discussion paragraph to the end of Section 3.

[Sections 3]

…

The proposed algorithm has several limitations as follows. First, the variations in radar backscattering and its corresponding image textures due to seasonal changes were not properly captured. Although the day of year was tested as a seasonality variable in the FC3 feature configuration, the result did not show any improvement. This is because day of year might not correspond to the same temperature, fluxes, and weather regimes.

How is the general accuracy derived? Is it a normalised average and does it account for the varying amounts of the different sea ice types? The overall accuracy of the sea ice classifier is only provided in the abstract and the conclusion. Please also provide it with the general results and discussion.

The overall accuracy in the abstract and conclusion is an average of the accuracies for each of the classes, thus the varying amounts of the different sea ice types are accounted for. In the revised manuscript, we adopted Cohen's Kappa to support the performance evaluation.

[Sections 2.2.6]

…

For performance evaluation, ~~W~~we use confusion matrix~~ for performance evaluation~~ and Cohen's kappa coefficient κ (Cohen, 1960), which measures the agreement between two raters (in this study, they are the trained classifier and the reference ice chart) with taking account of the possibility of the agreement occurring by chance for performance evaluation.

Is all level sea ice considered to have low backscatter? (P4 R7.)

No. They can have high backscatter in case frost flowers exist.

[Sections 2.2.2]

…

For surfaces with low backscattering such as calm ~~ocean~~ open water and level sea ice without presence of frost flowers on top, the effects from thermal noise contamination are visible not only in the backscattering image but also in some of the texture images (Park et al., 2019).

[Sections 3]

…

This might be because ~~there is no characteristic texture in the new ice patch; typically, they look just dark in SAR image~~ the new ice has different types of recently formed ice including nilas, which is smooth but rafting can make rough features, and frost flowers, which introduces high surface roughness and volume scattering (Isleifson et al., 2014), thus the new ice can appear either featureless dark or complex bright in SAR image (Dierking, 2010). The large range in backscatter values makes it hard to define characteristic texture in the new ice patch.

How does the spatial resolution of 1 km affect the results? Have you tested using different spatial resolution sizes?

The spatial resolution was set as 1 km to meet with two conflicting requirements. The size of subwindow for texture calculation must be larger than the spatial scales of the major sea ice structures, e.g., cracks, ridges, rafting patterns, narrow leads, which are up to a few hundred meters. The resolution needs to be comparable with modern high-resolution sea ice products from SAR. The regional ice concentration map published by Met.no has also 1 km resolution.

How was the value of 3 km set? (P7 R14)

There is no specific reason but we assumed that positional disagreements can be up to this amount in case the same SAR image was also used in NIC ice charting.

Sea ice does not form bergy bits, the terminology used in Figure 3 of bergy water may therefore be misleading. Please change to a more appropriate term.

You are right. The “bergy water” label was removed from Figure 3.

Visual inspection of Figure 7 seems to indicate that the method struggles with SAR image edges and that the same sea ice on different side of an image edge is classified differently, e.g. top right corner where one size of the edge is young ice and on the other side there is first-year ice. Please comment. Is the First-year ice observed in Figure 8 to the east of the old ice an artefact of a beam problem for the method?

As the reviewer pointed, the proposed method partly fails in resulting consistent result when the same sea ice on different side of an image edge is classified. Regarding the artifacts in Figure 8, it is most likely due to the wind roughened surface. When the sea surface is roughened by wind, it creates texture and easily misclassified as ice.
https://www.star.nesdis.noaa.gov/sod/mecb/sar/AKDEMO_products/APL_winds/wind_images2/2019-02/S1A_ESA_2019_02_08_07_21_08_0602925668_003.77E_79.54N_HH_C5_GFS05CDF_wind.png
Regarding these limitations, we added a discussion paragraph to the end of Section 3.

[Sections 3]

…

The proposed algorithm has several limitations as follows. First, the variations in radar backscattering and its corresponding image textures due to seasonal changes were not properly captured. Although the day of year was tested as a seasonality variable in the FC3 feature configuration, the result did not show any improvement. This is because day of year might not correspond to the same temperature, fluxes, and weather regimes. Second, the proposed method struggles with SAR image edges that the same sea ice on different side of an image edge was classified differently. This impose that the incidence angle dependence could not be normalized perfectly. An example of such a failure can be seen along the image boundaries at 80N, 35E and 82.5N, 60E, approximately. Third, some artifacts are observed under an extreme sea state. In the classified results in the bottom right panel of Figure 8, there is a misclassified FYI patch (yellow) in the open water area. According to the NOAA SAR wind image service, ANSWRS 2.0, the wind speed ranged from 17 to 21 m/s at the time of image acquisition, thus the water surface might be heavily roughened by strong wind.

What is your definition of New ice? For the ice types used here are you using the WMO definitions?

Yes, we follow the WMO definitions as the NIC ice chart does. In the WMO Sea Ice Nomenclature (WMO No. 259, volume 1 – Terminology and Codes), the definition of New ice is, "A general term for recently formed ice which includes *frazil ice*, *grease ice*, *slush* and *shuga*. These types of ice are composed of ice crystals which are only weakly frozen together (if at all) and have a definite form only while they are afloat."

In Figure 9 the sea ice type FYI thin is include in the sea ice classification results. Please clarify what thin FYI means. Why is this class not used throughout?

The label was wrong. It was just FYI without "thin". As the bottom panels (SAR results) are irrelevant in the context of the corresponding description, they were removed in the revised manuscript.

Consider adding something indicating the semi-automatic aspect of the manuscript in the title. And also indicate that the Sentinel-1 scenes used here only reflect the winter season.

We changed the title as "Classification of Winter Sea Ice Types in Sentinel-1 SAR images". The word "semi-automatic" and "semi-automated" appears several times in the revised manuscript including abstract.

Technical comments

In many places references to the appropriate work is missing, e.g. P1, R27, P4 R15-16 and P6 R9-10. Please carefully revise the manuscript to include references to earlier work.
Corrected.

The method is claimed to be semi-automatic in the body of the manuscript though the word semi- is left out of the abstract of the manuscript. Please correct.
Corrected.

P1 R22-24. Unclear sentence, consider rewriting.
Corrected.

P.1, R.14-15. Unclear sentence, consider rewriting.
Corrected.

Minor grammatical errors are present throughout the manuscript, e.g. P2 R5 : : : particularly in the cross-polarization: : : P2 R7 : : : considering the relative: : : p2 R31 : : :to train the classifier: : : P3 R15 To take advantage of the objective: : : Please carefully revise the English language throughout and pay particular attention to "the"
Corrected.

P2 R24 Region of interest -> region of study
Corrected.

P2 R25 coexit -> are found
Corrected.

P3 R2 National Ice Center (NIC) -> US National Ice Center
Corrected.

P3 R29. What does the precision of decimals mean here? That e.g. that the sea ice concentration can be 10.1%?
Reworded.

> Note that ice concentration label in the SIGRID-3 format ~~has precision of decimals~~ is assigned in increments of 10%.

P4 R5. What is strong noise?
Image contaminations as in Park et al., 2018

Consider using the commonly used term "open water" instead of "ocean" throughout. The same goes for ice free (IF) please use the much more commonly used open water (OW). MFYI could easily be confused with Multi-first-year ice. If MFYI is meant to indicate Mixed First Year Ice please change this or at least include this information in the header of the tables 4 and 5. In Table 4 and 5 are the MFYI meant to be used also on the predicted header?
Corrected.

P4 R12. Senitnel -> Sentinel
Corrected.

P5 R9 "Furthermore, further: : :" consider changing this.
Corrected.

P6 R10-12 you argue that when training dataset are prepared manually the sample size is usually less then 20 images. Please provide several references to support this statement.

> When training dataset is prepared by manual work (i.e., manual classification by human expert), the number of images is not large, usually less than 20 (e.g., 12 scenes in Zakhvatkina et al., 2013; 20 scenes in Leigh et al., 2013; 2 scenes in Liu et al., 2015; 4 scenes in Ressel et al., 2015).

P7 R6. Unclear sentence please revise.
Corrected.

P7 R19-21 appears to be a description of the method please move them to the methods section.
Following the reviewer's suggestion, they were moved to the end of the Section 2.

P8 R13-14. Substantial work studying young ice type has been carried out by e.g. Dierking 2010. Consider referencing such works. The young sea ice has a large range in backscatter values from very low to high and are also often subjected to frost flowers, see e.g. excellent work by Isleifson. Stating that young ice "typically just look dark in the SAR images" may therefore not be strictly true.
Revised.

> This might be because ~~there is no characteristic texture in the new ice patch; typically, they look just dark in SAR image~~ the new ice has different types of recently formed ice including nilas, which is smooth but rafting can make rough features, and frost flowers, which introduces high surface roughness and volume scattering (Isleifson et al., 2014), thus the new ice can appear either featureless dark or complex bright in SAR image (Dierking, 2010). The large range in backscatter values makes it hard to define characteristic texture in the new ice patch.

P9 R13-15. Unclear sentence, please revise.

Corrected.

P9 R16. What does “charted in” mean?
Reworded.

> Considering the ice edges in ice charts match well with those in the SAR backscattering images, thus the ices in the inner parts are also expected to be ~~charted in~~ mapped at the same time, …

P9 R17. What does “included wrong samples” mean? What makes a sample wrong?
Reworded.

> … it should be noted that training with ice chart might have included ~~wrong samples~~ mislabeled small features even if the image selection based on ice edge matching was successful.

P9 R19. When mentioning previous studies please provide references to these studies.
Added numbers from the references.

> Therefore, the lower classification accuracies compared to those in the previous studies (80% in Zakhvatkina et al., 2013; 91.7% in Liu et al., 2015; 87.2% in Aldenhoff et al., 2018), which used manually classified ice maps as training and validation reference, are expected.

P9 R31-32. Daily ice charts building on Sentinel-1 are provided by the Ice service at the Meteorological office in Norway, so it is certainly true that this can be done.
The ice chart from Met.no does not provide ice type but ice concentration.

Figure 4. What does the colorbar represent?
We accidentally omitted the title for the colorbar. It represents the number of subimages used for computing the histogram.

Figure 8. In Figure 3 and 7 the ice chart from 3 days later is used yet in Figure 8 the image from the same day is used. Please be consistent in which time interval is used for these weekly ice charts.
Since the reference ice chart is published weekly, the same NIC ice chart both in Figure 7 and 8 is supposed to be valid for both dates of Figure 7 and 8, but it is not as shown. The common time interval of 3 days in Figure 3 and 7 is just a coincidence. It can be one day or two days depending on the date of SAR image that used as source materials for ice charting at ice services. If the satellite images acquired two days prior to the publication of weekly ice chart, then the overall distribution of ice would represent the status of that time. Note that among the 57 images used for training, 42 and 35 percent of the images were acquired 2 and 3 days prior to the publication date of the corresponding weekly ice chart, respectively.