Point by Point Reply for the manuscript titled: "Bedfast and Floating Ice Dynamics of Thermokarst Lakes Using a Temporal Deep Learning Mapping Approach: Case Study of the Old Crow Flats, Yukon, Canada."

Maria Shaposhnikova, Claude Duguay, and Pascale Roy-Léveillée

We would like to thank the referees and the Editor for valuable comments which have substantially helped improve the clarity and quality of the manuscript and stimulated interesting and constructive discussion. The major changes to the manuscript are as follows:

1. *A brief comparison to a state-of-the art thresholding algorithm with the TempCNN has been carried out and added to the manuscript.*
2. *The advantages of the two approaches (threshold-based and TempCNN) have been emphasized and clearly documented in both the comparison subsection of the Results and Discussion, as well as the Conclusion section.*
3. *Most responses from the previous point-by-point response have been incorporated into the manuscript.*
4. *The manuscript has been revised and edited and all of the referees' questions/comments have been addressed below.*

Referee #1

Dear representatives of the TC editorial board and authors of the revised manuscript egusphere-2022-388,

Thank You for Your response to the reviewer comments. Now many of the reviewer comments have been responded, but the manuscript has not been changed much. Therefore, I still propose a major revision. Including most of the information and references given in the responses to reviewers will already improve the manuscript significantly. A typical reader will very likely not check the responses to reviewer comments, even though they were available, but wants to see all the necessary information in the paper.

Thank you for a valuable comment.

The manuscript has been revised to incorporate the majority of the information provided in the previous point-by-point response. Please, refer to the marked-up version of the manuscript which identifies the modified portions.

Major comments:

1) There are many explanations, additional information and references in the responses to the reviewers. However, the manuscript has not been changed much. I recommend the authors to include most of the useful information and references given in the responses in the manuscript.

This would significantly improve the manuscript and this is easy to implement as the text has already been written.

Thank you for a valuable comment.

The manuscript has been revised to incorporate the majority of the information provided in the previous point-by-point response. Please, refer to the marked-up version of the manuscript which identifies the modified portions.

2) Even though the authors emphasize that this is a proof of concept study, it is necessary to make some kind of comparison to some existing method or at least clearly indicate the advantages of the proposed method including reasoning to the advantages. A comparison to another method would be very informative, the comparison could also include comparison of the properties of the algorithms in addition

to the classification accuracy. It does not make sense if the method does not provide any improvement compared to the earlier methods. Even in a proof of concept paper some kind of reference is needed even though the comparison would not be very thorough.

Thank you for a valuable comment.

Based on the request of both referees a comparison to a state-of-the-art thresholding algorithm designed by Duguay and Wang (2019b) has been carried out. Consequently, the Data and Methods, Results and Discussion, and Conclusion sections of the manuscript have been modified to include the comparison. Please, see the details below (lines of the updated manuscript: 317-337; 487-510; 592-603):

"3. Data and Methods

3.6 Comparison to thresholding

In order to benchmark the proposed method against commonly used techniques of lake ice regime classification, it was compared to one of the most recent variations of the thresholding approach designed by Duguay and Wang (2019b) and applicable to S1 data acquired at HH and VV polarization. This thresholding algorithm defines the backscatter threshold between floating and bedfast ice as a linear function of the local incidence angle. As such, lake ice regime of each lake pixel is determined in a two-step process: 1) a threshold value is calculated using the following equation: $f(\theta)=-0.257*\theta-6.933$; 2) if the backscatter value (VV or HH) of a specific lake ice pixel is greater than or equal to the threshold it is classified as floating, if the value is less than the threshold it is classified as bedfast. Due to the fact that this approach is suitable only for lake ice pixels it is necessary to apply a lake mask to the SAR scene prior to the classification, as is also the case for other previously proposed thresholding approaches (see Section 1).

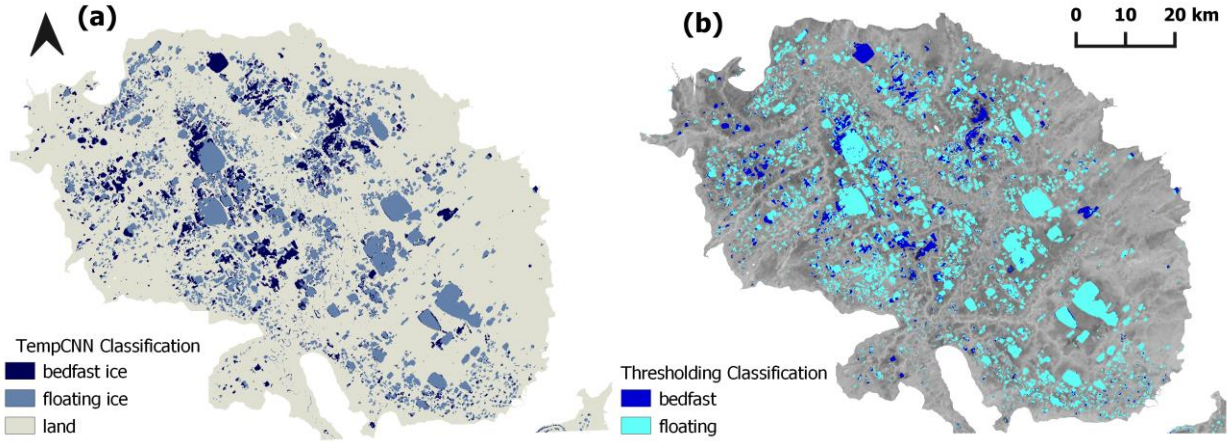
For the purpose of comparison between the thresholding approach and the temporal deep learning approach (TempCNN), lake ice regime maps were created for the four years of S1 data using thresholding: 2021 (March 15), 2020 (March 13), 2019 (March 14), 2018 (March 14). The thresholding algorithm linear function was applied to the local incidence angle layer obtained as part of the RTC level 2 products from ASF. The resulting threshold layer, where each pixel corresponded to the calculated threshold, was applied to classify the VV backscatter layer into either bedfast or floating ice based on whether the backscatter value was above or below the threshold for a given pixel. Next, it was necessary to apply a lake mask. Extraction of lakes is challenging in a wetland environment. As such, for simplicity, a single lake mask was created using an October 3, 2020, scene and a threshold of -16.5 dB identified experimentally by changing the threshold value in increments of 0.5dB, until lake boundaries where accurately captured. The four resulting lake ice regime maps were evaluated in terms of overall accuracy by utilizing the labelled dataset created as part of this work as ground truth. Results were then compared to those obtained from the TempCNN model for the same set of lakes.

4. Results and Discussion

4.3 Comparison to the state-of-the art thresholding approach

To benchmark the proposed temporal deep learning approach against the state-of-the-art methods of lake ice regime classification from SAR, a brief comparison to the thresholding algorithm proposed by Duguay and Wang (2019b) was carried out. The overall accuracy for each year was found to be as follows: 2018 – 87.8%; 2019 – 99.4%; 2020 – 98.8%; 2021 – 99.3%, with a mean accuracy of 96.3%. The mean overall accuracy of the TempCNN model with a 20/80% testing and training split which was used to create lake ice maps further employed for lake ice dynamics analysis of the OCF was 99.6 % (Table 2). It should be noted that the overall accuracy for TempCNN was carried out using the 18 years of data, while thresholding algorithm was evaluated using only 4 years of S1 data showing accuracies ranging from 87.8% to 99.4%. Figure 10 contains a side-by-side comparison of the lake ice maps created

by the TempCNN (Fig. 10a) and the thresholding algorithm (Fig. 10b) for March 2021. Visual analysis of the results of both methods are rather similar. Nonetheless, let us summarize the benefits and shortcomings of both methods. The thresholding algorithm: (1) produces highly accurate results overall for the four years examined; (2) is simple in implementation; however, (3) requires a lake mask; due to the dynamic nature of the wetlands a new mask would be needed for each year for the best results; (4) a local incidence angle layer is necessary; and (5) this algorithm has been designed to work with S1 data (VV and HH polarizations), while its suitability for other SAR platforms is yet to be explored. Temporal deep learning (TempCNN): (1) is more complex in implementation due to the requirement for time-series of scenes, rather than one scene; nonetheless, (2) produces highly accurate results; (3) does not require a lake mask due to its ability to classify VV and HH backscatter into three classes: floating ice, bedfast ice, and land, which is critical for dynamic thermokarst landscapes; (4) does not require incidence angle information; (5) based on visual comparison of the lake ice maps, TempCNN is better at classifying lake ice in deeper portions of larger/deeper lakes than the thresholding algorithm (in Fig. 10b, some misclassifications of floating ice as bedfast ice are observed in large lakes), and (6) the approach is applicable to multiple SAR platforms (S1 - VV polarization, ERS1/2 - VV polarization, and R1 - HH polarization) as shown in this study.



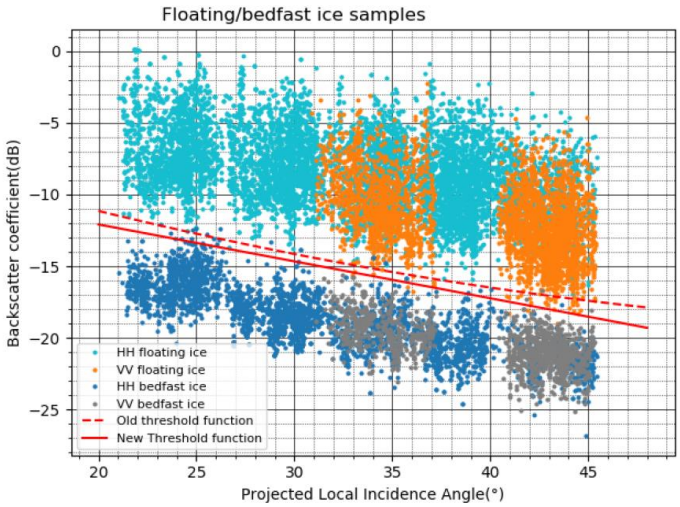
5. Conclusion

Comparison of the proposed approach to a state-of-the-art thresholding algorithm (Duguay and Wang, 2019b) has shown the lake ice classification results to be very similar with each algorithm offering its advantages. While both methods allow to produce accurate lake ice maps, thresholding is simpler in implementation as it requires a single SAR scene. TempCNN, on the other hand, is complexed by heavier data requirements, but does not require a lake mask or incidence angle information. In addition, due to extensive training on deeper portions of larger lakes, TempCNN is better at avoiding misclassifications of floating ice as bedfast ice in deeper portions of larger lakes. Although both methods are applicable to VV and HH polarizations, the thresholding algorithm used for comparison was designed to work with S1 HH and VV imagery and its applicability to other SAR platforms remains unexplored; although Engram et al. (2018) obtained an overall accuracy of 93% using a threshold-based algorithm with ERS1/2, RADARSAT-2, Envisat, and S1 SAR imagery evaluated over seven lake-rich regions in Arctic Alaska. The present study has shown TempCNN to also be applicable to S1 (VV), ERS1/2 (VV), and R1 (HH), achieving a mean overall accuracy of 95% in the classification of bedfast ice, floating ice and land, and 99.6% in the classification of bedfast ice and floating ice (using the TempCNN model trained on 80% of the labelled dataset).”

3) More evidence of the performance for different instruments and polarizations would be needed. Would it be possible to provide classification accuracies for different instruments (polarizations) separately and compare the results with each other and possible performance measures published earlier. Now it is just mentioned that "the authors believe..." in the response to reviewers. I don't think this is very scientific, a more concrete evidence would be needed. If there are some earlier studies with classification accuracies for similar classification problems then the accuracies could be compared directly to give even some idea of the performance. Of course it would be better to apply some of already published algorithms to the same data but if this will require too much resources, at least some kind of comparison would improve the manuscript significantly.

Thank you for a valuable comment. A comparison to one of the state-of-the-art thresholding algorithms (Duguay and Wang, 2019b) has been carried out and is described in response to the previous comment, as well as included in the manuscript.

At this point, it is not possible to provide separate accuracies for different instruments and polarization due to the way that the dataset has been split and used for training and testing. However, using different instruments and polarizations within the same classification algorithms has been done by multiple other researchers. For instance, Duguay and Wang (2019b) have developed a thresholding algorithm for Sentinel-1 and have demonstrated comparability of VV and HH polarized C-band SAR imagery for the purpose of classifying lake ice regimes (Please, see the graph below where New Threshold function shown as solid red line corresponds to the equation used for the comparison). A study by Engram et al., (2018) proposed an interactive threshold classification method to analyze floating and bedfast lake ice regimes across Arctic Alaska using 25-year time-series (1992-2016) of C-band SAR images from different platforms with both HH and VV polarizations. Engram et al. (2018) obtained an overall accuracy of 93% using an interactive threshold-based algorithm with ERS1/2, RADARSAT-2, Envisat, and S1 SAR imagery (including HH and VV polarizations) evaluated over seven lake-rich regions in Arctic Alaska.



Duguay, C.R. and J. Wang, 2019b. Arctic-wide ground-fast lake ice mapping with Sentinel-1. ESA Living Planet Symposium, Milan, Italy, 13-17 May.

Engram, M., Arp, C. D., Jones, B. M., Ajadi, O. A., and Meyer, F. J.: Analyzing floating and bedfast lake ice regimes across Arctic Alaska using 25 years of space-borne SAR imagery, Remote Sens. Environ., 209, 660–676, 2018.

Responses to my comments:

Lee filter size: it has not been reasoned why 7x7 filter size was used for S-1, I guess ASF has some kind of reasoning for the size (or number of looks) used in the filtering? How is it better than e.g. 5x5 or 9x9? For the other instruments the filter sizes have been selected to cover approximately the same area, how many looks these correspond?

Also include all this information in the manuscript. There is also a problem with the even size of the filter (4x4 and 16x16) because their center falls between pixels, for this reason an odd number as a size of a filter is recommended, as it has a center pixel.

Thank you for a valuable comment. ASF does not provide any reasoning for the speckle filter size. However, it does appear that a 7x7 kernel size effectively removes the speckle without loosing valuable information for a given pixel size. The reviewer is absolutely right about the odd filter size. After looking into the issue, we have realized that in fact the filter sizes used during the processing were as follows:

S1: 7x7 (area of 44,100 m^2)

ERS1/2: 17x17 (area of 45,156 m^2)

R1: 5x5 (area of 62,500 m^2)

The manuscript has been updated as follows (lines of the updated manuscript:170-173):

"Therefore, to match the RTC S1 products filtered using a 7x7 Lee Filter (the filter kernel covers approximately 44,100 m^2) with a dampening factor of 1 and 180 looks, ERS1/2 and R1 were speckle filtered using a 17x17 (45,156 m^2) and a 5x5 (62,500 m^2) Lee Filter, respectively. Adjusting the filter size allowed to account for the pixel size differences."

The experimentally defined threshold of -16.5 dB: I think it would be better to say "experimentally" than by "trial and error" in the manuscript. Also include the explanation of the method in the manuscript, it is now only in the response to my comments. Also indicate which data were used to experimentally define the threshold (training data set?).

Thank you for a valuable comment. The manuscript has been updated as follows (lines of the updated manuscript:332-335):

"Next, it was necessary to apply a lake mask. Extraction of lakes is challenging in a wetland environment. As such, for simplicity, a single lake mask was created using an October 3, 2020 scene and a threshold of -16.5 dB identified experimentally by changing the threshold value in increments of 0.5dB, until lake boundaries where accurately captured."

Linear interpolation: You say that more complex interpolation has a little influence on classification. How about even more simple interpolation i.e. nearest neighbor interpolation then? There is no evidence that nature is linear, often changes in nature are quite sudden and fast. Please, include the text of Your response and the reference in the manuscript.

Thank you for a valuable comment. The temporal deep learning classification method proposed in this work for lake ice regime classification from SAR closely follows the method described in Pelletier et al., 2019, which used linear interpolation to fill the temporal gaps. In future work, we will definitely consider exploring other methods of interpolation, such as nearest neighbor as has been suggested by the referee.

The manuscript has been updated as follows (lines of the updated manuscript:204-214):

"Resampling to a daily frequency and linear interpolation were applied to compensate for the temporal irregularity of the data ensuring that each of the backscatter time-series for each year of data had the same length (161 values) and gearing it for the deep learning classification (Pelletier et al., 2019; Valero et al., 2016). Although the lake ice lifecycle is non-linear, previous studies have shown that more complex

interpolation methods have little influence on classification accuracy (Pelletier et al., 2019, Valero et al., 2016). Linear interpolation was performed utilizing python programming language and the tools of pandas module. Interpolation was performed individually on every time-series (backscatter value of each pixel traced through time). As a result, we obtained SAR image stacks consisting of 161 full coverage scenes, which were subsequently input into the TempCNN to perform classification. In addition to the proper SAR processing and speckle filtering, further quality control was implemented by filling any missing or Not a Number (NaN) values, especially common for ERS1/2 and scene fringes, as part of the temporal interpolation process. The final labeled time-series consisted of 161 time steps (i.e., one time step per day) covering the time period between October 4 and March 13."

Check all the selected parameters and threshold and give reasoning to their values, also indicate on which data set the parameter selections are based or give references to publications where similar selections have been used.

Thank you for a valuable comment.

The selection of parameters closely follows Pelletier et al., 2019, as has been indicated in the manuscript. The number of convolutional units was selected through a cross-validation procedure using the labelled dataset created as part of this work with the inclusion of 2020/2021 season imagery, reserved for final testing.

Referee #2

The manuscript has been improved significantly since the first submission, some of the technical issues have been addressed and a number of ambiguities have been clarified. I think that the method presented here has the potential to provide a reference for the lake ice research.

However, I still worry about the novelty of this manuscript. Deep learning method has been widely used in the Earth Science and usually performs better than previous methods. In the original and current versions of manuscript, the comparison to the previous methods is still missing. I understand that the main purpose of this study is just to present a new Deeping learning method for lake ice research. However, it is difficulty to determinate whether the presented method significantly outperform the existing methods and therefore does not provide a substantial added value. The authors admitted that the thresholding approach is indeed very useful and definitely wins over the proposed approach due to its simplicity. Therefore, was this method performed because it could be or because it should be? The latter needs to have demonstrated scientific value. The authors should provide more quantified evidence rather than theory.

Thank you for a valuable comment. Based on the request of both reviewers a comparison to the state-of-the-art thresholding algorithm designed by Duguay and Wang (2019b) has been carried out. Consequently, the Data and Methods, Results and Discussion, and Conclusion sections of the manuscript have been modified to include the comparison. Please, see the details below (lines of the updated manuscript: 317-337; 487-510; 592-603):

"3. Data and Methods

3.6 Comparison to thresholding

In order to benchmark the proposed method against commonly used techniques of lake ice regime classification, it was compared to one of the most recent variations of the thresholding approach designed by Duguay and Wang (2019b) and applicable to S1 data acquired at HH and VV polarization. This thresholding algorithm defines the backscatter threshold between floating and bedfast ice as a linear function of the local incidence angle. As such, lake ice regime of each lake pixel is determined in a two-step process: 1) a threshold value is calculated using the following equation: $f(\theta)=-0.257*\theta-6.933$; 2) if

the backscatter value (VV or HH) of a specific lake ice pixel is greater than or equal to the threshold it is classified as floating, if the value is less than the threshold it is classified as bedfast. Due to the fact that this approach is suitable only for lake ice pixels it is necessary to apply a lake mask to the SAR scene prior to the classification, as is also the case for other previously proposed thresholding approaches (see Section 1).

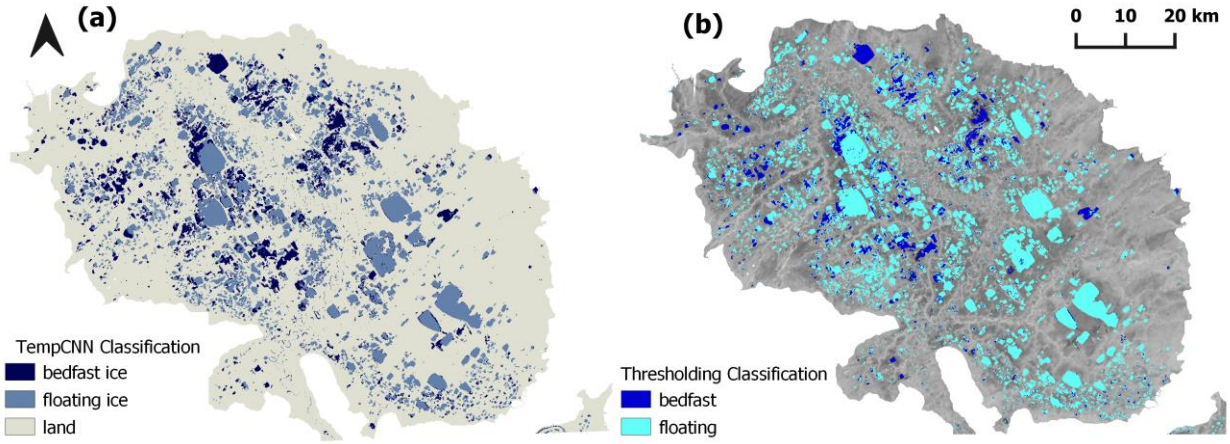
For the purpose of comparison between the thresholding approach and the temporal deep learning approach (TempCNN), lake ice regime maps were created for the four years of S1 data using thresholding: 2021 (March 15), 2020 (March 13), 2019 (March 14), 2018 (March 14). The thresholding algorithm linear function was applied to the local incidence angle layer obtained as part of the RTC level 2 products from ASF. The resulting threshold layer, where each pixel corresponded to the calculated threshold, was applied to classify the VV backscatter layer into either bedfast or floating ice based on whether the backscatter value was above or below the threshold for a given pixel. Next, it was necessary to apply a lake mask. Extraction of lakes is challenging in a wetland environment. As such, for simplicity, a single lake mask was created using an October 3, 2020, scene and a threshold of -16.5 dB identified experimentally by changing the threshold value in increments of 0.5dB, until lake boundaries where accurately captured. The four resulting lake ice regime maps were evaluated in terms of overall accuracy by utilizing the labelled dataset created as part of this work as ground truth. Results were then compared to those obtained from the TempCNN model for the same set of lakes.

4. Results and Discussion

4.3 Comparison to the state-of-the art thresholding approach

To benchmark the proposed temporal deep learning approach against the state-of-the-art methods of lake ice regime classification from SAR, a brief comparison to the thresholding algorithm proposed by Duguay and Wang (2019) was carried out. The overall accuracy for each year was found to be as follows: 2018 – 87.8%; 2019 – 99.4%; 2020 – 98.8%; 2021 – 99.3%, with a mean accuracy of 96.3%. The mean overall accuracy of the TempCNN model with a 20/80% testing and training split which was used to create lake ice maps further employed for lake ice dynamics analysis of the OCF was 99.6 % (Table 2). It should be noted that the overall accuracy for TempCNN was carried out using the 18 years of data, while thresholding algorithm was evaluated using only 4 years of S1 data showing accuracies ranging from 87.8% to 99.4%. Figure 10 contains a side-by-side comparison of the lake ice maps created by the TempCNN (Fig. 10a) and the thresholding algorithm (Fig. 10b) for March 2021. Visual analysis of the results of both methods are rather similar. Nonetheless, let us summarize the benefits and shortcomings of both methods. The thresholding algorithm: (1) produces highly accurate results overall for the four years examined; (2) is simple in implementation; however, (3) requires a lake mask; due to the dynamic nature of the wetlands a new mask would be needed for each year for the best results; (4) a local incidence angle layer is necessary; and (5) this algorithm has been designed to work with S1 data (VV and HH polarizations), while its suitability for other SAR platforms is yet to be explored. Temporal deep learning (TempCNN): (1) is more complex in implementation due to the requirement for time-series of scenes, rather than one scene; nonetheless, (2) produces highly accurate results; (3) does not require a lake mask due to its ability to classify VV and HH backscatter into three classes: floating ice, bedfast ice, and land, which is critical for dynamic thermokarst landscapes; (4) does not require incidence angle information; (5) based on visual comparison of the lake ice maps, TempCNN is better at classifying lake ice in deeper portions of larger/deeper lakes than the thresholding algorithm (in Fig. 10b, some misclassifications of floating ice as bedfast ice are observed in large lakes), and (6) the approach is applicable to multiple SAR platforms (S1 - VV polarization, ERS1/2 - VV polarization, and R1 - HH

polarization) as shown in this study.



5. Conclusion

Comparison of the proposed approach to a state-of-the-art thresholding algorithm (Duguay and Wang, 2019b) has shown the lake ice classification results to be very similar with each algorithm offering its advantages. While both methods allow to produce accurate lake ice maps, thresholding is simpler in implementation as it requires a single SAR scene. TempCNN, on the other hand, is complexed by heavier data requirements, but does not require a lake mask or incidence angle information. In addition, due to extensive training on deeper portions of larger lakes, TempCNN is better at avoiding misclassifications of floating ice as bedfast ice in deeper portions of larger lakes. Although both methods are applicable to VV and HH polarizations, the thresholding algorithm used for comparison was designed to work with S1 HH and VV imagery and its applicability to other SAR platforms remains unexplored; although Engram et al. (2018) obtained an overall accuracy of 93% using a threshold-based algorithm with ERS1/2, RADARSAT-2, Envisat, and S1 SAR imagery evaluated over seven lake-rich regions in Arctic Alaska. The present study has shown TempCNN to also be applicable to S1 (VV), ERS1/2 (VV), and R1 (HH), achieving a mean overall accuracy of 95% in the classification of bedfast ice, floating ice and land, and 99.6% in the classification of bedfast ice and floating ice (using the TempCNN model trained on 80% of the labelled dataset).”

Some specific comments:

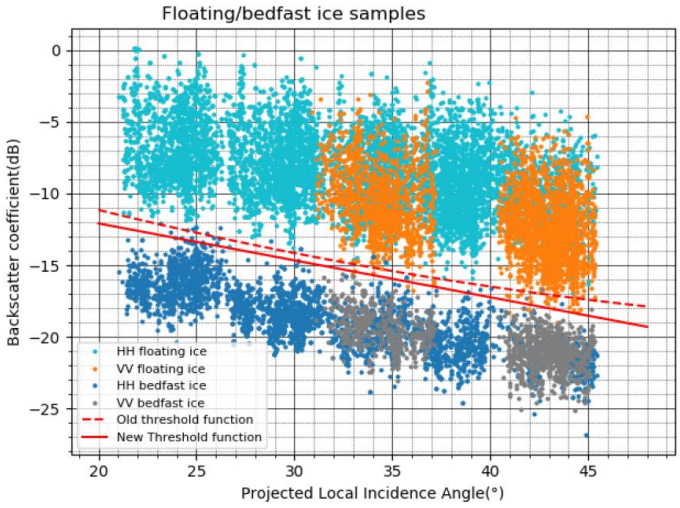
1) As also pointed out by another reviewer, the input data come from different instruments and they even have different polarizations, the effect of this should be analyzed. Usually, it may be more difficulty to construct a Deeping learning method by using multiple sources of satellite data. However, the accuracy of classification result is still quite high, this may imply that the lake classification by using SAR backscatter is not very complexed. In author’s response, the authors provided some counterarguments for this, they thought ‘One of the strengths of a deep learning approach is the ability of the network to be trained on different types of data, learn all of its different aspects, and then be able to recognize and classify them correctly.’. However, the obtention of the accurate classification result is not the only aim, the potential mechanism is also worth to explore.

Thank you for a valuable comment. The reviewer is right, the classification problem is indeed not very complex. However, the proposed approach does resolve multiple problems not tackled by the existing state-of-the-art lake ice regime mapping approaches. The proposed temporal deep learning approach is able to recognize 3 classes: bedfast ice, floating ice, and land, a property not offered by any other method. SAR signatures of land and floating ice become very similar towards end of the season and are not easily separated without applying a temporal approached, proposed in this work. The ability to work

with 3 classes, alleviates the need for a lake mask, which is a significant advantage in an environment where lake boundaries are constantly changing. In addition, a method that analysis temporal evolution rather than a single scene is more robust as instead of relying on a single value, it makes a classification decision based on a seasonal backscatter evolution.

Moreover, the contribution of the manuscript is not only in proposing a new method of lake ice regime classification, but also in creating and analyzing a time series of lake ice regime maps for Old Crow Flats, Yukon, Canada, - a wetland of international significance - which has not been done by other researchers and has the potential to benefit numerous researchers carrying out other research activities for this study area.

In terms of polarization differences, it has bee shown by other approaches that VV and HH polarized C-band imagery is comparable for the purposes of lake ice mapping. For instance, the thresholding approach proposed by Duguay and Wang, 2019b and used by us for comparison (please, see above), is suitable for both HH and VV polarised Sentinel-1 as is demonstrated in the graph below.



Duguay, C.R. and J. Wang, 2019. Arctic-wide ground-fast lake ice mapping with Sentinel-1. ESA Living Planet Symposium, Milan, Italy, 13-17 May.

In addition, Engram et al. (2018) obtained an overall accuracy of 93% using an interactive threshold-based algorithm with ERS1/2, RADARSAT-2, Envisat, and S1 SAR imagery (including HH and VV polarizations) evaluated over seven lake-rich regions in Arctic Alaska.

2) The authors used ice thickness data to evaluate the lake ice classification results: 'where ice thickness is equal to the lake depth, lake ice regime is bedfast, whereas in areas where ice thickness is less than the lake depth, lake ice regime is floating.'. What are the precisions of these two datasets, is the ice thickness exactly equal to the lake depth?

Thank you for a valuable comment. The precision of the field lake depth/ice depth measurements is 1- 2 cm. The ice thickness was determined to be equal to lake depth where ice was determined to have frozen to bed as indicated in our previous sentence.