Review of Sivaraj et al.

This study presents an improved methodology for classification of Sentinel-2 imagery for melt pond fraction estimates. Specifically, the combination of random forest models and morphological algorithms allows for the reduction of misclassifications from nilas, submerged ice, and brash ice.

The goal of this study is compelling, and the new methodology presented shows promise. However, the methods section requires greater detail and additional figures to enhance clarity and support the results. I've noted in the specific comments where a supporting figure would be useful. Furthermore, some key details need to be addressed, such as determining the minimum detectable melt pond size using Sentinel-2. Once this is established, a follow-up analysis should be conducted to assess the implications of this threshold on the study's findings.

Specific:

Line 15: how are you citing a 2002 paper for a trend 1979 to 2020. Either reference a dataset, or a more recent paper

Line 29: either a more specific location and time of that study or a less specific number. 50% should be a range or should include "on average", or a specific location.

Line 56: A little misleading to reference Webster 2015 here- because this study utilizes NTM satellite imagery- which has other limitations (coverage, consistency, accuracy) but is not aerial photographs

Line 61: delete "similarly"

Line 62: fragmented sentence starting with "considering that.." is hard to follow

Line 63: long sentence with a lot of thoughts, consider separating into two sentences

Figure 1: consider putting this map in polar stereographic

Figure 1: include a bounding box for the area noted on line 77.

Line 98: does the atmospheric correction mask out areas of thick clouds? What do you mean accounts for them?

Section 3.2: who created the training data, a single sea ice expert, several people, non-experts?

Line 113: what does exceptionally thin mean? I'm sure the WMO nomenclature has specifications for nila definitions.

Line 117: it would be great to have a figure representation of the morphological dilation and reconstruction to eliminate brash ice and nila misclassifications

123: Was this a pixel-based classification then? Figure 5 makes it look like the melt ponds and misclassifications are objects- groups of pixels. Do you eliminate the whole object or just the overlapping pixels? Does that leave some misclassifications in the middle of an area that is misclassified? An example of this would make a great figure

Line 134: how do you create the open water mask that is "only open water pixels"

Line 141: did you assess the performance on all images?

Line 155: why are the two sites such different sizes if you are using a worldview image for comparison at both?

Line 161: it would be interesting to know the statistics of the WV melt ponds that are less than the S2 image pixel size. How many are there? what MPF does this make up?

Line 164: "a higher pixel resolution of less than 2 m" is confusing phrasing

Section 3.5: it would be great to see how the Sentinel-2 and Worldview images look next to one another. Also please note the delta time for acquisition between S2 and WV

Line 170: how much does the MPF change between the dilation/reconstruction methods and the manual reduction? In other words, Figure 3 could include three points indicating the RF output, the morphological reduction, and the morphological +manual reduction.

Line 170-174: how does these stages compare with the Eicken stages of melt pond evolution?

Eicken, H., H. R. Krouse, D. Kadko, and D. K. Perovich, Tracer studies of pathways and rates of meltwater transport through Arctic summer sea ice, *J. Geophys. Res.*, 107(C10), 8046, doi:10.1029/2000JC000583, 2002.

Line 179: show this- show misclassification or ambiguous surfaces listed here- the melt ponds, melted through melt ponds, disintegrated ice.

Line 183: what about other types of misclassifications- melt ponds identified as open water, melt ponds that are not identified (classified as ice), what are other misclassifications that could occur, what is the magnitude of these misclassifications, and in what direction would each of these misclassifications bias the MPF.

Figure 3: include more grid lines.

Line 196: how are you determining melt pond onset? From you MPF estimate? Consider looking at melt onset passive microwave product to compare with your results.

Line 197: if you are going to compare with other studies, make sure you indicate the time and location of the other studies and how you might expect those factors to contribute to the same or different results you are seeing

Line 203: the large difference could be also due to lack of data and not capturing the maximum MPF date

Line 210: populated by mix of ice types in which year?

Line 210: show the image during the seasonal maximum of 2108MPF and see if they are in fact MYI floes.

Line 211: how are you defining "spatial variability"

Line 211: and what is the reason for "spatial variability of melt ponds"- does Perovich 2002 give an answer?

Line 215: is the Neihaus data public? Could you look just specifically at their region to examine the differences?

Line 215: if they only provide Arctic wide estimates of MPF it does not make sense to compare your results with theirs at all- of course the melt timing is different.

Line 225: by "smaller ponds < 200 m^2" doesn't that mean just one single Sentinel-2 pixel? Do you have a limit on the number of pixels required to call the object a pond? This is common of other studies- see Buckley et al 2023 and Perovich et al 2002 papers. If you eliminate small ponds which may be erroneous, you will probably have a totally different distribution to discuss.

Perovich, D. K., W. B. Tucker III, and K. A. Ligett, Aerial observations of the evolution of ice surface conditions during summer, *J. Geophys. Res.*, 107(C10), 8048, doi:10.1029/2000JC000449, 2002.

Buckley, E. M., Farrell, S. L., Herzfeld, U. C., Webster, M. A., Trantow, T., Baney, O. N., ... & Lawson, M. (2023). Observing the evolution of summer melt on multiyear sea ice with ICESat-2 and Sentinel-2. *The Cryosphere*, *17*(9), 3695-3719.

Line 232: see also Buckley et al., 2023 for melt pond evolution for comparison especially because it also uses Sentinel-2 data for MPF.

Figure 4: I think this could be better represented with month and day on the y axis and then the data plotted across the season with each year as a different color- that would be easier to visualize the patterns and how they line up year to year.

Line 239: was the performance the lowest or the presence of errors and misclassifications lowest? How can you know? Are these percentages a percentage of the misclassifications being removed? This is a confusing discussion, please clarify

Line 260: It would be beneficial to establish a lower limit for the detectable melt pond size. Given that a pond in the 100 m^2 to 200 m^2 size range corresponds to just a single Sentinel-2 pixel, the discussion around melt pond size distribution within this category lacks robustness. I would be particularly interested in seeing the distribution of ponds starting with a minimum size equivalent to 10 Sentinel-2 pixels. Additionally, it would be valuable to include a dedicated section discussing the limitations of the 10 m resolution, specifically addressing how many ponds and the total pond area may be missed compared to higher-resolution imagery, such as that from WV.

Line 281: There is the stat I was looking for! Maybe include this earlier where noted.

Line 299: The conclusions of the study feel overly optimistic and would benefit from a more balanced discussion of the methodology's limitations. Specifically, the following points should be addressed:

- Limitations of Sentinel-2 coverage: Sentinel-2 imagery is restricted to areas within a certain distance from the coastline, which limits its applicability in more remote regions. Additionally, temporal limitations should be discussed—how frequently are these specific locations sampled, and does this sampling frequency align with the temporal dynamics of melt pond evolution?
- Impact of cloud contamination: A detailed assessment of the extent to which cloud cover renders data unusable would provide valuable context. Quantifying the proportion of affected data could help clarify the reliability of the results.
- Manual aspects of the algorithm: The methodology currently involves manual corrections and decisions, such as selecting the optimal morphological operation. If this approach were scaled up, manual intervention would not be feasible, necessitating a discussion on how these steps could be automated or standardized.