The Cryosphere Discussions

The Cryosphere Discuss., author comment AC2
https://doi.org/10.5194/tc-2022-86-AC2, 2022

Reply on RC2

Wenkai Guo et al.

Author comment on "Sea ice classification of TerraSAR-X ScanSAR images for the MOSAiC expedition incorporating per-class incidence angle dependency of image texture" by Wenkai Guo et al., The Cryosphere Discuss., https://doi.org/10.5194/tc-2022-86-AC2, 2022

Review to submitted manuscript; "Sea ice classification of TerraSAR-X ScanSAR images for the MOSAiC expedition incorporating per-class incidence angle dependency of image texture". The manuscript investigates per-class sea ice incidence angle dependencies in TerraSAR-X ScanSAR images and GLCM textures and trains a Bayesian classifier to classify sea ice surrounding the MOSAiC expedition.

Thank you for a well written manuscript with strong English and interesting results covering a high-profile scientific campaign. To summarize the main critique points: the paper is too long, convoluted to read at times, and it is difficult to keep track of discussed subjects. In addition, parts of the methodology needs to be further clarified.

I agree with the other reviewer that the manuscript would be better suited split into two, and resubmitting them with major revisions. One manuscript could focus on the IA dependency of the TSX SC intensity and the GLCM textures, while the other could examine the GIA and the time-series of the MOSAiC campaign.

Major Comments

Disclaimer. I have limited practical experience with bayesian classifiers but extensive knowledge of deep learning with emphasis on sea ice segmentation using convolutional neural networks. Reviewing the methodology regarding the bayesian classifier raises the following concerns for me, which I would like you to consider and address:

• Limited testing (validation in your words) examples rectangles □ 10 reference 3x3 pixels for each class is selected for each reference scene (13 scenes in total). This is a total of 1,170 pixels for each class. Considering the abundance of data at your disposal (>1,000 x >1,000 pixels in each image?), I would refrain from needle picking select small areas. Labelling data is a time consuming task but there are tools available, which could assist, e.g. https://github.com/ESA-PhiLab/iris. At least I would require a justification for the approach.

• Small size of testing rectangles

□ Why are 3 x 3 pixel rectangles selected? Could they be larger? Why not? Do the pixels have to be separate or could you label an area with multiple classes?

We aim to standardize the training/testing pixels across different classes. 3x3 pixel rectangles are selected considering typical widths of linear or small features, mainly leads, young ice areas, deformation features, and small, homogeneous ice floes. We also try to keep a relatively even distribution of polygons in each scene, thus adjacent polygons are far away from each other (roughly larger than 50 pixels), an example of which can be seen on fig.3. This approach has been used by one of our recent studies, but is not elaborated in the text of this manuscript. The above information is now added to the text to improve clarity.

• Spatial and temporal biased training and testing

□ Generally training and testing should be carried out on areas without spatial or temporal correlation, i.e. on different scenes to avoid biases spilling over from the training to the testing phase. As the data is randomly split in training and test, I fear that some pixels may lie very close together, and could artificially improve the model performance but without carryover to generalization of the classifier (i.e. may not be as reliable on non-testing data).

As now mentioned in the text, reference polygons are selected to be far from each other to make for a relatively even distribution over the scene. These polygons, not the pixels inside, are randomly split into training and testing. Therefore, the resulting training and testing polygons/pixels still keep a reasonable distance from each other.

More information on how the classifier is trained should be included. How is it optimized?

Assuming linear IA dependence, the class-specific IA slopes and intercepts for each feature (TSX HH intensities and textures) are estimated using values from the training pixels. These IA dependence parameters are then used to calculate linearly variable mean vectors and subsequently covariance matrices which characterize the class distributions, thus fitting the classifier to the training data.

These technical details can all be found in the paper introducing the GIA classifier and is

thus omitted from this manuscript given its length. A note is added to clarify this point: 'HH intensities and textures calculated accordingly for the training pixels are used to train the GIA classifier **(for details of the training process see Lohse et al., 2021)**.'

Data

The data selection should be more clearly explained or alternatively visualized using the acquisition dates. 53 scenes are used in this study, 50 during the MOSAiC campaign, 3 afterwards with low IAs to complete the spectrum. 40 of these scenes are not used for training the classifier (as I understand it). 13 of the 53 scenes have 10 3x3 rectangles labelled and used for training and testing.

All scenes are within the MOSAiC period. This paragraph is edited for better clarity and avoid misunderstanding:

'This study analyzes 53 scenes (2019.11.01 to 2020.04.11, IA: 17.18◦to 59.56◦) during the MOSAiC winter, with an average of 3 scenes per week. All scenes are used for the examination of IA dependencies of HH intensities. Among these, 50 scenes (2019.11.01 to 2020.03.28, IA: 31.90◦to 59.56◦) are further used for the demonstration of IA dependencies of image textures and sea ice classification. The remaining 3 scenes (2020.03.31, 2020.04.03 and 2020.04.11, IA: 17.18◦to 36.70◦) are only used to complete the coverage of the full IA range of TSX SC data in examining the IA dependency of HH intensities. All scenes are radiometrically corrected and calibrated to σ0 and converted to dB. '

Generally, when optimizing models, data is typically split into training, validation and testing and if supervised methodologies are applied, each split will have raw data (X) and a reference (Y), i.e. the "ground truth". Typically, a validation subset should be utilized for decision making during the optimization process, i.e. should we stop (early stopping), should we tweak the learning or regularization parameters? And finally the model performance is evaluated on the test data, which no optimization changes have been made upon. As I understand the GIA training process, you are using a test subset, and should call it as such. In regards to the segmentation tools applied, personally, I would have chosen to apply convolutional neural networks. At least it should be mentioned as a potential area of future work.

The terms 'training and testing' are now used across the text. This study has a specific aim to demonstrate class-specific IA dependencies, and thus chose a classifier which specifically incorporates this phenomenon. Yes CNN is surely a powerful tool for image classification, and a sentence is added to the end of this section to mention its potential use in the future.

Minor Comments

L54: Why does the TSX SC data only come in the HH polarization?

A decision was made for all TSX SC scenes for MOSAiC to be acquired in the HH polarization for consistency and to enable comparison with C-band SAR which typically come in HH+HV.

L128: 10 reference rectangles of 3 x 3 pixels sounds very small. That is only 90 pixels per class per scene, i.e. 1170 pixels. Is every class represented in every scene? And how certain are you of your qualitative selection?

See reply to the first major comment. Yes every class is represented in every scene in an equal amount. The qualitative selection is aided by co-authors who participated in the MOSAiC campaign and have extensive knowledge of the ice conditions along the expedition. A comparison between the classified results and a manual sea ice categorization map is shown in fig.8. The lack of continuous in-situ observation of sea ice type through the time series, which is the only definite 'ground truth,' is mentioned in the section 'Limitations and future steps.'

L129: Improving consistency between training scenes using a 40 km x 40 km area is unclear to me. How does this work?

The sentence lacks a bit of explanation. We meant to choose training polygons in roughly the 'same ice' through the time series, through visual examination of how the ice evolved/moved and tracing the ice in the polygons. Given variable positioning of the CO relative to TSX image borders, a 40km x 40km square around the CO is roughly the maximum area which are captured by all images throughout the time series. Thus, within this area, the task of placing training on 'the same ice' can be achieved.

These information is now added to the text for better clarity.

L180: Only textures of HH intensities have a consistent relationship with IA.. HH intensities as opposed to what? Or is it referring to the scaling of the image, i.e. dB.

This is now clarified as: 'In an initial examination of GLCM textures, we found that only textures of HH intensities in the logarithmic (dB) domain have a consistent linear

relationship with IA, given properly constrained IA range (more details below), while textures of HH intensities in the linear domain does not.'

L337: 'On the contrary..' This sentence is quite difficult to read. I think it should be split up into two sentences.

Edited.

In addition, there is a pdf document attached with grammatical suggests.

These have been integrated into the text.