Reviewer 1

The manuscript “Sea ice classification of TerraSAR-X ScanSAR images for theMOSAiC expedition incorporating per-class incidence angle dependency of image texture” presents methodology and results of sea ice type classification of TerraSAR-X imagery obtained during the MOSAiC expedition. Despite very interesting findings, due to large diversity of methods, results and analysis and a large size of the manuscript, it is recommended to split the manuscript in two parts, improve the order of the presentation and resubmit the manuscript(s) after a major revision.

Major comments

1. Although only two objectives are formulated in the introduction, the impression is that the manuscript attempts to fulfil at least four: 1. Investigate per-class AI dependence; 2. Optimize parameters of texture features; 3. Train and evaluate classifier; 4. Analyse time series. In my opinion such variety of objectives does not allow to focus well. That makes the manuscript too long to read and the story too difficult to follow. I would suggest to completely remove section 3.3 and correspondingly reduce section 3.4. I’m confident that results shown in these sections deserve a separate paper. I will therefore focus my review on the first, methodological part.

We think it's reasonable to maintain the current manuscript structure, while streamlining the text for better clarity and readability, sharpening the aim of the manuscript to 2 points: introducing a classification product, and demonstrating TSX IA dependency in support of the classification.

As suggested by the title, this paper mainly serves as an introduction to a classified time series which can be a useful dataset for MOSAiC-related research. Quantitative evaluation of the classification, the comparison to ice roughness transects, and the comparison to ice opening records in other studies have shown the reliability of the classified time series.

Method development was not central, having used an established classifier and commonly used GLCM textures to aid the classification. The parameter optimization process is implementation, rather than development. The demonstration of IA dependencies of TSX intensities/textures, which is another major finding of the paper, secondary to the main objective, is also based on the time series itself.

The manuscript has been re-aimed around these 2 points in all sections, with more emphasis placed on demonstrating the IA dependencies & relating the classification results to the MOSAiC mission. The text is more concise and the length of the methodology part has been reduced.

2. What is GIA classifier? Authors refer to that term in many places, but it is never defined or explained. I guess, that’s one of the central blocks in the classification algorithm: apparently, the backscatter, the texture features, the IA are passed into the mysterious “GIA classifier” for doing the actual classification. But how?! I’m very curious to know. GIA classifier needs to be clearly explained.

As stated in the text, the GIA classifier was developed and published in Lohse et al., 2020. The investigation of IA dependencies of sea ice types on TSX SC is central to this study, hence the choice of the GIA classifier which specifically incorporates this phenomenon well. This is explained by the most part of the 3rd paragraph in the Introduction. We think that this length is suitable for the manuscript which is currently already lengthy.

3. Order of presentation needs to be revised in order to correspond to the selected logic (Intro, Data and methods, Results and Discussion): Lines 133 – 144 and Fig. 4 should come in Section 2.1 Data; Lines 151 – 164 with Table 1 and lines 271 – 275 with Table 2 belong to Introduction as they describe state-of-the-art; Section 2.3.2 belongs to Results as it describes WHAT is achieved and not HOW it is achieved.

Edited.

Exceptions:

a. Table 1 is specific to this study. Therefore, we think it's best to keep it in its current position.

b. Section 2.3.2 demonstrates IA dependencies, which we think is secondary to the main aim of introducing the classification product, and belongs to the 'Materials' category, introducing one of the characteristics of the dataset. We wish to focus the Results & Discussion part to evaluating the product and relating it to other MOSAiC products/studies. Therefore, we think it's best to keep Section 2.3.2 in the M&M section.

4. Analysis of IA dependence for various ice types need to be increased as it is an important result of this work. What is error-bars of the slopes (it can be computed, e.g. by bootstrapping) and what is significance? What is the reason for large positive bias of the slopes – speculation on stronger volume scattering needs to be expanded. What is physical reasoning behind positive slope for leads? The suggested method and parametrisations seem to be difficult to use in other conditions (C-band, other IA, other ice types, summer). Although it is mentioned as a limitation in the end, I believe it is important to also underline in the Introduction – the goal is to study a specific TSX SC timeseries and for analysis of another dataset a similar full-scale analysis needs to be performed.

In the current fig.4, IA slope values in bold fonts indicate statical significance of the linear regression model, while regular fonts indicate otherwise. For HH intensities, all slope values are significant except for the leads class, which is expected as all leads pixels are under the noise floor of the sensor, leading to a wide distribution of pixel values that does not exhibit a significant linear dependency to IA (mentioned in the text). The positive slope is therefore not significant and has no physical meaning (visually, a slight negative slope can be observed). Errors can be computed for the linear models but is of less interest to our study, but can be included in the appendix if desired.

Yes, an explanation is added to the introduction to clarify the limited setting of this study: 'In summary, the objectives of this study are: 1. to investigate and demonstrate per-class IA dependencies of TSX SC HH intensity and GLCM textures **specifically for the above mentioned study area and period**….'

5. Image size and number of texture features are undoubtfully important hyperparameters of the Haralick algorithm. However, neglecting quantisation level and distance to neighbour pixels can lead to significantly worse results. Sensitivity to these two parameters should also be studied, for example in this respect: how does despeckling boxcar filter impact the GLCM? In theory, if a 3x3filter is applied and then GLCM is computed with 2 pix distance, there should be almost no elements in GLCM off the main diagonal. On another note, Haralick (1973) suggested using adjacent pixels (d=1) so the choice of authors d=2 should be tested and explained better.

The number of quantization levels directly impacts the precision of the converted integer values used for GLCM calculation in representing the actual pixel values. Therefore, within reasonable computational loads, more levels are desirable. A level of 64 is thus chosen, and the reasoning is included in the text.

The speckle filter is no longer used to further preserve spatial details.

The displacement value is now directly added to the parameter optimization process (section 2.3.3), i.e., an optimal set of displacement size, window size and texture combination is selected together.

Minor comments

L7. Phrases in parenthesis make the sentence very unclear. Please split into two sentences.

All instances in manuscript are adjusted accordingly, except for very short clarifications and acronyms.

L12. Unfortunately the GIA classifier and class probabilities are never explained in the manuscript.

See reply to major comment 2.

L24. Please provide reference to prove the “largest expedition in history”?

Reference added.

L71. Objective 1 is actually two objectives: 1 . to investigate and demonstrate per-class IA dependencies of TSX SC HH intensity and GLCM textures; 2. to determine the feasibility and optimal parameterization of including texture measures as input features to the GIA classifier.

This paragraph has been re-edited to more clearly show our main aims – introducing the classification product, while demonstrating IA dependencies.

Figure 1 shall be removed as it does not explain anything.
Edited.

L96. “and shown in details in ” -> “, dates shown in”
Edited.

L107. Why were these ice categories chosen? It should be written that other categories were not present in the dataset and the method cannot be extrapolated.

‘Based on the ice conditions in the study area and period’ is added to show that the choice is made considering the specific conditions of this study.

L129. Polygons == rectangles? This is unclear.
The first instance has been edited to ‘polygons in rectangles,’ and ‘polygons’ are used subsequently.

L133. Maybe “evolution of young ice” ?
Edited.

L135. Please rewrite “wide-spread lead openings of open water or nilas” as “wide opening of leads with water or nilas”
This is meant to emphasize the spatial abundance of leads within the area, instead of the physical widths of the leads. It has been edited to ‘ wide-spread opening of leads with water or nilas.’

Figure 3. The smallest sub-images seem to be very blurred. Is it the effect of the despeckling filter or just visualisation?
The smallest sub-images are zoomed in to a level where individual pixels are visible, in order to give a (example) visual impression of textures of each class at this scale. This visual effect is natural at this zoom scale, and also it can be seen that different classes have different ‘blurriness’ which are related to how their textures differ from each other. In the current revision, we no longer apply a speckle filter on the images, and visually these subsets are now less ‘blurred,’ as expected.

L208. Cannot agree here. Other authors also studied distance and number of grey levels (e.g. Clausi 2002). Sensitivity to these two parameters need also to be studied (see major comments).
See reply to the corresponding major comment.

L221. “...and thus is a relatively…”
Edited.

L236 and 237. Is that already results of parameter optimization? Then it is better to move to the Results section.
Edited.

L268. Whay volume scattering is presumed to be stronger?
In this sentence, ‘stronger volume scattering’ refers to MYI, and ‘added randomness in backscatter caused by deformation features’ refers to deformed FYI. ‘Respectively’ is now added at the end of the sentence to avoid confusion.

Figure 6. Is positive slope for leads even physical? How the strong positive bias of the slopes can be explained?
All of these pixels are under the noise floor of the sensor, resulting in unreliable IA dependencies. Explanation is added to the text.

L274. “This is given that” can be removed.
Edited.

L276 – 280. This seems to be logical after the results, in the Conclusions section.
As this relates to the limitations of this study, these sentences have been moved to the ‘Limitations and future steps’ section, which also avoids lengthy conclusions.

L321. A reference to unpublished work just supports my concern that it is too early to include this section in the manuscript.
This sentence only refers to the method of classifying the sea ice roughness transects that happens to be also used in another study. These roughness transects are analyzed specifically for this study.

Figure 10. It is impossible to see shades of blue on the roughness transects.
The symbology is adjusted for better visualization with thicker transects.

Figure 11 and Lines 389 - 392. Why the 10% sudden drop of the polygon area on ~15 December is not reflected in a similar change of MY or young ice? Why does the lead ice increases ~3 times on 1 March and this is not reflected in the polygon area? Where are the plots of “other mosaic studies” that are easy to compare with fractions of different ice types? I’m afraid it is too early to write “that the classified time series is valuable as indicator of ice openings” as I cannot see a proof of that. Instead, the variations of ice fractions seem to be rather spontaneous and connected to uncertainty of the algorithm.
As mentioned in the text, the polygon formed by the buoys is a small, variable area around the ship, which is much smaller than the parts of TSX scenes used for classification. Therefore, not all variations are synchronized between these time series. We’ll consider putting results from other papers directly within the plots for easier comparison.

L396. “The leads class are mostly fully represented in the classification map” is it really a limitation? Can be removed from that section.
This is added in comparison to the sentence before (incomplete representation of thin young ice areas). ‘Comparatively’ is now added to clarify.

L425. Convolutional neural networks also deserve being mentioned as a potential tool.
Deep learning-related methods are definitely important tools for sea ice classification, but this sentence only talks about potential utilization of different forms of image texture. A sentence is added to the end of this section to mention the future use of CNN in the classification.

Reviewer 2

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: 'Sea ice classification of the time series is conducted using the GIA classifier trained with HH intensities and textures with optimal parameterization. **Details of the training process can be found in 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: please see lines 93 – 107 in the revised manuscript.

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?
This is indeed confusing to the reader, and non-essential to the work. The use of this extra 40km x 40km square cut of the scenes is now removed.

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 do 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.