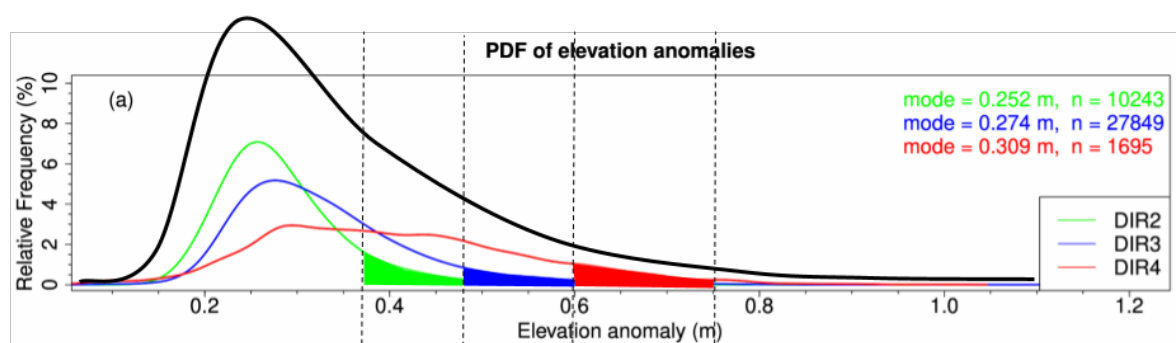


General comments

The manuscript has been greatly improved. Comparison with SAR imagery increases confidence in the value of IS2 measurements. Analysis of uncertainties and error sources is detailed. Nevertheless, feasibility of using FIS DIR for training the classification algorithm is still questionable.

I'll try to explain my concern from a different angle. In Section 2.4 you mention that clear separation of elevation anomalies (EA) into three groups appears only when you use P90 and you even use only 5% of anomalies with the highest values for computing the thresholds. But it means that only 5% of the classified IS2 DIR correctly correspond to FIS DIR. The rest, 95% of EA cannot be classified in accordance FIS DIR by definition.

To illustrate I use your plot with histograms of EA separated by DIR level. I draw a sum of histograms as a thick black line – that's a distribution of *a priori* unknown EA. It is not accurate as I draw it by hands, but it gives a realistic representation. I add dashed lines at location of your thresholds. I mark with green, blue, red patches the sections of EA histograms that are correctly classified. If you compare the area of these patches with the rest of the histogram, you can also see that it constitutes approximately 5%. The rest of EA (area under the thick black curve) are classified incorrectly or not classified.



The fact that less than 5% IS2 DIR correspond to FIS DIR must be clearly presented in the manuscript.

In my opinion we can confidently conclude that the available FIS DIR cannot be used for training or testing of DIR classification from IS2 data. Therefore, either the title should be changed as proposed earlier to "Comparison of IS2 elevation anomalies with FIS DIR...", or a clear statement that more data (e.g. more FIS DIR observations or quantitative SAR data) is needed for developing a reliable DIR level classification algorithm should be added to conclusions.

Detailed comments

Line 180. Higher reliability of the classification is not proven. Only the classification of the 5% of points is more reliable. Classification of the rest of the data (extrapolation ability) may become less reliable when you decrease the amount of data. Either remove "reliable" or prove that it is more reliable by splitting the training dataset in two random parts, training the classifier on one part of data and applying it to another part of data.

Figure 6.

The intention behind the figure is very good but it needs to be improved.

SAR images are too small for visual analysis, profile plots are too busy to distinguish anything. Please enlarge SAR images and reduce size of profiles and remove excessive data from profiles (e.g. present only one beam on a plot) and decrease the range of Y-axis down to [-0.5, 1] in order to stretch the profiles vertically.

Titles on planes f and h should be corrected to 27 March 2019.

Line 470. The argument that “IS2 seems to carry similar information as the S1 images, and the FIS DIR are based partly on icebreaker observations and partly on SAR imagery” is obviously weak and cannot be used to justify usage of FIS DIRs to train or validate algorithms. As mentioned above, only 5% of the classified anomalies correspond to FIS DIR, making it unusable.