Comment on tc-2022-57
Anonymous Referee #2

The paper compares methods to detect days with rain on snow. Two separate comparisons are made for two separate periods. The first comparison is between estimates of ROS from WRF and estimates of ROS from a gridded meteorological product, seNorge. The second comparison is between estimates of ROS from a method applied to Sentinel-1 SAR data, and estimates from the seNorge product.

Although evaluations of different methods for mapping ROS events are needed, the present paper has a number of shortcomings that the authors need to address before it can be published.

1) All of the methods presented use a proxy for detecting rain on snow. There is no attempt to validate these methods. Instead, the authors use a thresholding of daily station observations of near-surface air temperature, precipitation and snow water equivalent as "truth". This thresholding approach is one way to estimate precipitation phase, especially if no other data is available, but it is still only an estimate. A number of the stations used in the study have precipitation type as an observation, which could provide a way to evaluate the thresholding approach, giving some confidence in the comparisons of the other methods. These observations are not without problems as they are often based on automated detection systems. So care needs to be taken but they are likely the best observations available, with the exception of documented events. Validating estimates of rain on snow using precipitation type would improve the paper.

2) The SAR product used in the paper detects wet snow (and is used by
Nagler and Rott to map melting snow cover) not rain on snow events. I don't think maps of melting snow can be used to identify rain on snow on their own because there is no way to know if wet snow results from liquid precipitation or from melting. A rain on snow event has to be conditional on precipitation occurring. The authors find that many more rain on snow events occur in the SAR-derived product than in the seNorge product. This doesn't seem a surprising result. I don't think they are comparing the same phenomena: on the one hand some proxy for liquid precipitation over snow cover derived from temperature, precipitation and snow cover, and on the other hand liquid water in the upper layers of the snow pack, which may or may not be because the snow is melting. So I don't know what is learned from the analysis.

There is also a sampling problem. The SAR maps are an instantaneous mapping of surface characteristics about every 2 to 4 days (based on Fig A1). The authors then appear to add up the days wet snow is detected to get the number of ROS days, assigning a limit of 20 days. This seems to be over-counting ROS days. A single liquid precipitation event, under warm conditions, could result in a wet snow pack for several days but I don't think this should be counted as several days of rain on snow events. Surely, only the day(s) with liquid precipitation should be counted as rain on snow. Certainly, none of the station data suggests that there are 20 consecutive days of liquid precipitation but wet snow mapping for the Grotli station comparison suggests there is about a month of days with rain on snow in December. This seems an unlikely long sequence of rain on snow days, especially as there does not appear to be precipitation on these days. Further evidence needs to be provided to show that this occurred.

My suggestion is that the SAR ROS mapping algorithm needs to be modified to include precipitation.

3) The same thresholding algorithm is applied to the WRF, seNorge, and station datasets. So the differences in the results come not from the method but the different inputs to the algorithm. I would like to see some attempt to diagnose the cause of differences between the ROS frequency maps by looking at the input data.

4) The authors use accuracy as performance metric. This is not a good metric for a skewed dataset like rain on snow; rain on snow days are not as common as non-rain-on-snow days. In such cases, Accuracy can be spuriously high. Statistics such as F1-scores are more robust and should be used.