I appreciate the authors changes to the manuscripts and believe the motivation for this work and for using an RF is much improved in the text. I have a few more things that need clarification before I am happy to accept this manuscript.

Why did the metrics change? For example, in the abstract the MAE changed from 11.6 to 10.5. What caused this change?

L109 – change ‘masked’ to ‘missed’

L227 – delete “with”

L227 – I’m confused that you say SnowModel is forced with both ERA5 and MERRA. Do you take the average of output from model runs with each of these forcings?

L240 – What do you mean by “TDD… is returned as a cumulative percent”. Is this the percent of time that the temperature was above freezing?

Section 4.1.2 – for your training/testing split – were there any spatial/temporal considerations that were taken into account? Is it possible that for one site, some of the timeseries could be included in the training data set and some in the testing?

Section 5.3 – The % of pixels categorized into each QC category changed fairly substantially between versions of the manuscript. I don’t think I fully understand how this categorization is made and I’m, curious what led to this change. Maybe some more details about this categorization would be helpful. Also in Figure 6 (previously 5) the areas that are categorized as ‘best’ change substantially, particularly for snowoff.

L519 – Typo – ‘gennerally’

Section 5.4.2 – I believe there are still errors in this section. See this comment from before:

> “L421 –“snowmelt onset occurred ~9 days earlier than average, and snowoff 6 days earlier, lengthening snowmelt duration by 4 days…” Maybe this is a rounding error, but should this not be 3 days? Same with the following sentence. Maybe including decimals would help.”

The same is true for the next sentence: “Conversely, in 2013, a cooler year, snowmelt onset was 12 days later and snowoff 7 days later, shortening the snowmelt duration by 4 days.”

Figure A6 – Why does this figure only go until 2020, not 2023?