Response to Reviewer 1:

We again thank Dr. Dunmire for taking the time to read our manuscript and provide a constructive and thorough review. We did our best to address each comment and believe the revised manuscript will be improved. Our responses to each comment are pink.

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?

The updated MAE is the result of result of re-running the RF model after incorporating two new predictor variables (daily total column water vapor from MERRA2 and a topographic complexity layer. These additions were made in response to Reviewer 1's suggestions. The inclusion of these features led to slight improvements in MAE reported in abstract and paper.

L109 – change 'masked' to 'missed'
Done
L227 – delete "with"
Done
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?

Changed: 'The SnowModel is driven by with MERRA-2 and ERA5 meteorological reanalysis data and provides daily estimates…'

Now reads as: 'The SnowModel dataset was developed using meteorological data derived from both MERRA-2 and ERA5 reanalysis data. These reanalyses were used to create bias-corrected inputs for SnowModel, resulting in daily estimates…'

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?

Changed: 'TDD was created by summing daily mean temperature above 0 degrees Celsius for each pixel and is returned as a cumulative percent.'

Now reads as: 'TDD was created by summing daily mean temperatures above 0 degrees Celsius for each pixel during the melt season. To facilitate comparison across regions and years, TDD values were normalized to a cumulative percent scale (0-100%), where 0% represents the onset of above-freezing temperatures and 100% represents the total seasonal accumulation.'

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?

Added the following paragraph for clarity: 'To ensure independence between training and testing data, we implemented our 80/20 split at the timeseries level. Such that, each timeseries represents a single pixel across a full snow season and is treated as an induvial unit. This precaution avoided any overlap in time or space between the training and testing sets. Consequently, no individual pixel's timeseries contributed data to both training and testing subsets in each iteration, ensuring that model evaluation reflected only out-of-sample predictions.'

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 snowo^.

Thank you for this comment. In the previous revision, we used Jenks natural breaks to classify predicted model error into four QAQC categories. At that stage, we had already incorporated additional predictor variables (e.g., total column water vapor and topographic complexity), which may have influenced the distribution of predicted errors. Because Jenks classification determines category thresholds based on the distribution of values, these updates led to shifts in both the percentage of pixels in each category and their spatial distribution. In the current revision, we changed the QAQC classification method to a quantile-based approach, which divides predicted error into four equal-sized groups. This change was made to improve interpretability and allow for more consistent comparisons between snowmelt onset and snowoff QAQC maps. While the overall spatial patterns remain broadly similar, the quantile-based classification results in more intuitive and balanced category assignments.

L519 – Typo – 'gennerally'
Corrected

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."

Thank you for noting this again. We updated these to 3 and 5 days, respectively.

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

A6 goes to 2020 because we are using 1991-2020 as the 30-year climatological mean. This 30 year span is the current 30-year normal defined by NOAA.

Updated line 355 to : 'To analyze snow phenology trends over time, we developed snow phenology climatology for the period 1991–2020, which corresponds to the current 30-year climate normal (Palecki et al., 2021).'