A random forest derived 35-year snow phenology record reveals climate trends in the Yukon River Basin
Pan et al (2025)

This manuscript uses a random forest to estimate snowmelt and snowoff timing in the Yukon River Basin, Alaska. First, I apologize to the authors for my delay in submitting these comments. The manuscript is generally well written, but would benefit from additional details provided in the methodology. In general, I am not convinced that this new methodology is a substantial improvement compared to already existing methods. As such, I believe that major revisions are necessary before publication.

Major comments

My major concern with this work is that it seems that the RF approach isn't really necessary. In fact, for snowmelt onset, the previous thresholding approach appears to do better (MAE of 11 days compared with 11.6 days for RF), and the IMS dataset does better for snowoff timing (MAE of 15.87 days compared to 16.03 days for random forest). Why can a combination of the thresholding approach for snowmelt onset, and IMS for snowoff not be used then?

Further, In the conclusion, it is stated that this study delivers "new insights into the timing and variability of snow melt onset and snowoff" but I actually fail to see these new insights. A lot of this phenology was already presented in Pan et al 2020 (i.e. section 3.2). Is there actually any new insights that are gained from the RF approach, that are absent when using the thresholding method? The motivation for this work needs to be substantially improved. Given the previous work from 2020, which already presents a 30-year snow phenology record of the YRB with many of the same insights, what does this work add? Why is current knowledge of snow-melt and snow-off conditions in YRB insufficient?

Second, some details are lacking about the data/methodology. Also, in general, a schematic of the workflow would be incredibly helpful. Please see the below comments, organized by section:

Section 3.2.1
L165 – How frequent are missing temporal observations in this data?

Section 3.2.2
Isn't there also a 1km IMS product that exists? Why was this not used? I also find details about SnowModel to be lacking here. What is the model forced with? Is it consistent with IMS when the two products overlap?

Section 3.2.3
How was the cumulative thaw degree days calculated? Cumulative from what point in time?

Section 4.1

Did you try any other ML models such as xgboost or neural networks?

Section 4.1.1
How many combinations of parameters did you try in the grid search? What were the parameters that were changed and what was the range of values tested? Please also reference Table A2 in this section.

Section 4.1.2
More details in this section would be helpful. For example, how was the split between training and testing done? How was the feature importance computed? Why was the QC metric split into 4 categoric categories, instead of just kept as a number between 0 and 1? Is there information that is lost in doing this?

Section 4.2.1
The classification into 'earlier' and 'later' snow events is not clear. What does 'snow event' mean specifically? Why was this done?

Section 4.2.2
Please specify what this data was used for.

Minor/technical comments:

L 66: "While higher-frequency K-band…" Perhaps this line should go to the next paragraph? It seems a bit out of place in this paragraph which primarily covers C-band SAR and Sentinel-1.

Figure 1 – Please change the triangle markers to dots.

Section 3.1 - This study is for the YRB, but many of the sites used are located outside the YRB, particularly a large number along the southern coast. I understand this is likely done to enhance the limited in-situ data available to train the model, but the impact of this this should be mentioned somewhere.

L140 – What is meant by 'analysts'?

L144 – Why is there a different number of observations available for training snow melt onset and snowoff?

L149 – please specify that the snow cover is binary.

L146 – 155 – Please mention somewhere in these paragraphs that these variables will be described in more detail in the following sections.

L153 – "GTOPO" in parenthesis is not needed.

L172 – 174 – please change "reduce" to "reduced" in these sections to be consistent with the tense throughout the methods section. Check other areas in the methods as well.

L176, L178 – The equations are numbered in the text, but not in-line. If possible, please include a number next to the equations for easy reference.

L209 – Can the 'TC' abbreviation be changed to 'FTC' to be consistent with 'FW' ("fractional water")?

L226 – I do not fully understand the sentence: "Although RF does not inherently model temporal sequences like some other algorithms, temporal dimensions were incorporated by structuring each day as a sample within a sequential framework." Specifically, the 'within a sequential framework' is throwing me off. RFs don't have a sequential framework, and each sample is treated distinctly, not as part of a sequence. Can you please clarify this here?

L230 – what were the spatial and temporal resolution of the other similar snow record? Is it really correct to say that this study is a 'significant temporal enhancement'?

L239 – "By labeling the timeseries accordingly..." It is not clear to me how labeling the timeseries this way enables any of the following numbered points.

Figure 2 – In the bottom panel, it would be more clear to set the y-axis as merely "dry" for 0, and "wet" for 1.

L302/306 – Are these the scores of the training dataset? testing?

L315 – Is there any bias in the snowmelt/snowoff timing?

L320 – What is "the full YRB dataset"?

L322-323 – It is not clear to my how these MAE values differ from the ones presented in L315-317.

Section 5.1.2 – What are you using as the ground truth in this section to compute the MAE?

L335 – the 'FW' feature was not grouped into 4 natural breaks (not a variable in Fig. 3)

L339 – "higher FW is associated with much higher MAE values for snowmelt onset." I don't see this in Figure 4. FW of 10 and 13% have by far the highest MAE and then it decreases for higher values. Maybe I'm misunderstanding this figure?

Figure 3 – I think the division into 'four natural breaks' is not super helpful in this figure. I guess we can see that MAE for snowmelt is higher at higher elevation but what are the

elevations? Also, it doesn't make sense for the aspect feature as aspect describes a direction that has a cyclical nature, not really a standard increasing variable.

Figure 4 – The x-axis for this plot is confusing to me. What is it binned so seemingly randomly? For example, the x-axis jumps from 10 to 11 to 13 to 22.

L341 – Why is it anticipated that error increases with higher surface water cover?

L352 – "These results support inclusion of the static variables as additional RF predictors, despite their relatively low importance". I'm not totally convinced by this. All the ANOVA results are saying is that the MAE is statistically significantly different for different land cover types, not that the land cover type is helping improve the model MAE, right? This analysis could indicate that perhaps your model is missing processes for certain land cover types, thus influencing the MAE for these land cover types, but I don't think it indicates that the static variables are helping improve the model. Did you test models with these static variables omitted to see how the MAE changes? This would give a better indication that the static variables are useful.

Section 5.4.2 – Are these anomalies better captured by the RF or are they already seen in other records?

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.

Section 5.5.2 – Can a figure be included for the trends in seasonal temperature? Also, what does the 'tau' value represent?

A figure comparing the performace of the different methods (IMS, SnowModel, thresholding, RF) for snowmelt and snowoff would be helpful.