

Volumetric Quantifications and Dynamics of Areas Undergoing Retrogressive Thaw Slumping in the Northern Hemisphere

Corresponding Author: Dr Chunli Dai

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The manuscript by Dai et al. is a highly ambitious piece of work that presents novel information on the dynamics of RTSs at the Pan-Arctic scale. By using large amounts of ArcticDEM data, the analysis covers the entire Pan-Arctic at a very high spatio-temporal resolution. The approach shares aspects (using ArcticDEM at Pan-Arctic scale with deep learning approaches) with the work by Huang et al. (2023 ISPRS Journal of Photogrammetry and Remote Sensing, <https://doi.org/10.1016/j.isprsjprs.2023.10.008>). However, the present manuscript goes beyond identifying RTSs by presenting spatially and temporally detailed volumetric estimates of soil loss and associated carbon release. The found magnitude of carbon release from RTSs provides an important new viewpoint to the globally relevant issue of permafrost degradation and associated carbon budgets. The manuscript also makes an interesting point by showing how volumetric RTS activity demonstrates latitudinal and regional differences in their responses to climate, but also that ecosystem properties markedly complicate this connection.

A comment that needs to be made is regarding the relatively weak classification precision of the presented deep-learning method. High number of false positives is typical for approaches using automated recognition of landforms, and I appreciate the authors for addressing this challenge and pursuing to improve the automated mapping by combining solutions (change detection algorithm and deep learning) and using additional training data with data augmentation methods. Yet, also after this model precision remains low and coming up with reliable results requires undoubtedly laborious visual inspection. On the other hand, the outcomes of the visual inspection are also affected by which locations are selected as candidate RTSs with the initial model and augmented models. Please see my specific comments regarding this aspect below.

The results are interesting and novel and should be of broad interest to the permafrost research community. The manuscript has an easy-to-follow structure even though the methodology presented involves several processing steps. The figures are clear and interpretable, and the language seems to me (as a non-native speaker) to be in very good order. Below I raise a couple issues and add minor suggestions for corrections and improvements. I recommend publication of the article after my raised issues have been addressed.

Specific comments:

rr166-183: Here, the authors describe how the mapped RTS distributions are consistent with previous studies. I would also like to see a discussion of whether there are any regions or environments where RTS activity could not be mapped with a similar precision, or if the results contradict previous findings. Now, the discussion of uncertainties centers on the large number of false positives but the potentially missed RTS activity is less addressed. Although, the limitations in detecting some established activity at megaslumps due to narrow horizontal scale of activity was satisfactorily explained, which indicates that the approach may not have high recall in all conditions. Do we know if model evaluation statistics (from prediction and evaluation with Victoria-Peel Plateau data) hold for the Pan-Arctic area or can there be regional/environmental variation?

r517: The split ratio of 9:1 sounds reasonable. Were other ratios tried in model training to examine their effect on overall accuracy?

rr563-564: Were the samples of 1660 training and 189 validation images also here split randomly based on the manually selected and inspected 1660+189 images? Please elaborate because models should be affected by how well the variation in the modelling data is represented.

rr576-578: I find it quite surprising how the combination of five models (original + augmented) increase precision so drastically. Model ensembles often do outperform single models, but the improvement is surprisingly large. Maybe single models identify different FPs and when put together, there are not so many common FPs, and thus the precision is higher? Could this be elaborated?

r745: I could not access the produced data via the (supposedly) DOI link (Mendeley?). Are the data available already?

Fig. S7-8. The final detected RTSs are seemingly concentrated in the training data (both 1 and 2) locations (as should be expected) while relatively little RTSs are detected in some areas known to have abundant occurrences. For example, only 13 RTSs have been detected in Alaska. I was wondering why there were no training images in Step 3b selected from Alaska at all? Could this be because the initial model from Step 3a could not predict potential RTSs there due to different environmental conditions in those areas, and consequently that also Step 3b results in very little RTSs in the western part of Brooks Range, for example? Or are these possibly “missing” RTSs simply filtered out in the classification in Step 2 for being too small or not active during the study period?

Technical touches:

r99. Replace eg with e.g.

Fig. 1 a, b. Imagery source not visible, please enhance.

r223: 3gh -> 3g,h

Table 1: The pink and grey colors are hardly separable. Consider adding contrast.

Fig. 4: The legend for the correlation coefficients has different ranges to those presented in the figure caption and elsewhere in the text. Which are correct?

rr355-360: This sentence could be more clearly formulated. Is there a comma missing? Consider revising.

r643: Please elaborate the idea behind selecting the growth rate of double the long-term rate in Segal et al. 2016?

r652: “...for a pan-Arctic total of $\pm 12,514$ ”. Please check if this is correct. Are confidence intervals/standard deviation for the value missing here or how should one read the value?

r940: Please check the text font.

Figs. S7: This is a bit rough figure. Consider using the same longitude of origin for the polar projection that was used in the other maps for better comparability. Please, also acknowledge the imagery source and consider adding a legend in which training and validation images are preferably presented with unique colors.

(Remarks on code availability)

Unfortunately, I only found the Github link now as I am submitting my assessment. Anyway, I use different coding languages in my work, so I would not be able to test the reproducibility of the code or provide a sufficient assessment.

Reviewer #2

(Remarks to the Author)

The manuscript by Dai et al quantify retrogressive thaw slumps (RTS) in the northern permafrost area in terms of location, time of development and volume. The data is generated by comparing multitemporal DEMs based on ArcticDEM and identifying RTS based on a machine-learning algorithm. The data set not only identified polygons of RTS, but also the volume of the landform. This in turn could be used to estimate possible carbon release due to permafrost thaw and erosion. This combination makes the manuscript and its data base highly valuable for the research community, and this study is as far as we know unique and absolute worth publishing, even if the time series of 10 years is too short to evaluate real climate trends.

We are not experts on the machine-learning algorithms and let this part be evaluated by other reviewers. However, the following major comments are given and may be considered:

- Title: Be precise, you deal (at least according to your writing) almost exclusively with retrogressive thaw slumps (RTS), why write then “large permafrost-thaw driven mass wasting events”? See also comment below.
- Type of RTS: We are not always sure what landforms you have detected. Of course, the big RTS in Arctic lowlands are captured and exemplified in images in the paper. However, in the Arctic islands, e.g. on Svalbard, you mentioned large RTS developed (by thermal erosion?) in large recent (< 250 years old) ice-cored moraine complexes in front of recent glacier fronts. Those those would have a quite different development history than ground-ice induced thaw slumps, even in areas of buried old Pleistocene glacier ice (there is a large literature base on this thermal erosion in ice-cored moraines on e.g. Svalbard). Maybe, you should show different types of identified RTS from the various regions? We wanted to have a look in your data for e.g. Svalbard, but the link provided failed.
- Correlation of events to temperature and precipitation: First of all, the results of the correlation analysis, which is interesting, is in the Discussion (chapter 3.1.), but it should be a Result. Consider moving the first part of that chapter (l. 252. onward). You then write that 60% of the RTS are neither correlated with trends in ERA temperature or precipitation. I am not sure how

you made this correlation, but an event one year is not necessarily correlated with climate the same year, but with the weather development the previous year. E.g. an extreme summer one year may lead to weakening of the structure and failure in one of the subsequent years. You should consider this in your discussion (see pre-conditioning papers in high-mountain permafrost slopes, e.g. here <https://iopscience.iop.org/article/10.1088/1748-9326/ad3c55>). This pre-conditioning may be more important further south than in the northernmost parts?

- Soil organics release: These are nice data, again, why do you not present these as a result, and then discuss them later. Line 341 ff: The reference Tanski et al. (2017) does not support the statement of less SOC release. Here, the numbers are just small because SOC is buried again. However, in this study, the deposition of SOC is not taken into account, so that your numbers should be compared only to the released SOC.
- Discussion in general: The Discussion in general could benefit from a more detailed elaboration of the results from Fig. 3, which is nice, but not discussed further in the paper (just presented as a result). E.g., why do you observe such areas, volumes, depths, rates and accelerations? Is air temperature and precipitation the only explanation as you focus on this in the discussion? What about ground ice contents, topography, geomorphology etc.?
- Comment on the so-called "Rainbow pattern": What about RTS where activity started or ended during the observation period and thus show no or only a weak rainbow pattern? Are these excluded? If that is the case it should be discussed in the uncertainties.
- Uncertainties: A major result of the discussion is the relationship between RTS activity and climate, and the influence of RTS on carbon release. This discussion is based on ERA5 data and SOC map by Hugelius et al. Here, we miss a more in-depth description of the uncertainties in ERA5 and the SOC map, which of course influence the conclusions. soil organics carbon map.
- Figures: Some of the figures are a bit messy and should be revised. A list of suggestions are given below in the minor comments.

Minor comments are as follows:

Figures and tables:

Fig. 1: Scales are lacking in a, c and d. Lat/long are missing in at least a and b.

Is c and d the same RTS? If yes, which one of the RTS shown in b?

Fig. 2: Scale is missing in a, no lat/long grid. D show temperature, given again in Figure 4b? Can be avoided here.

Figs. 2d and Fig. 4b: Gives the same information, only once for a specific RTS and once for all RTS. Maybe Fig. 2d is redundant (only mentioned in the discussion as an example where one could refer also to Fig. 4b)? At least have the same axis labels and legend.

Fig. 3: Nice figure, but not discussed, see comment above

Fig. 4: The figure is a bit messy, lots of different information, do you need the big pie diagrams? They more or less are similar. D is very messy, what do you want to show? High loss rates for arctic lowlands are never discussed really.

Fig. 5: This figure may be combined with Fig. 4b because of redundancies. The trend in 5b is really $r > 0.5$? To be honest, for us the distribution of the red dots is quite random.

Fig. S2: We find the explanation of the figure is not clear: We assume that the red area around P1 is the deposition area? Why is the deposition earlier (2011) than the retreat of the headwall (2012-2015, indicated by Fig. S2g)? This is also corroborated by Fig. S2f, where deposition was in 2011 but head wall retreat until 2016? How can this be explained? Is it because only the largest elevation difference is analysed (see Line 1281 ff)? Then maybe mention it briefly here and not only in the supplement. Finally in Fig. S2b and e: Why are there so many elevations in 2015 available and only one for the other years?

Fig. S11a: The blue dots in a, are those all RTS? 2747? This is not possible.

Table 1: Maybe add a column with loss rate. The loss rate is shown in Fig. 2a but it would be nice to have a summary for each region.

Other comments:

Line 37: Add the exact time period 2012-2022

Line 98: You state that the existing literature focus on one or two-dimensional detection and measurement of RTSs, but give only an example for 2D detection. Please add another one for 1D detection.

Line 162: Information in the brackets is unclear: Why give it in ha/ 1000 km²? This is not really intuitive. And what do the percentages describe? Many of such numbers in the text decrease readability, better to use a table to summarise numbers.

Line 177ff: RTS and ice-cored moraines. We think there is quite a difference between recent ice-cored moraines like on Svalbard and old buried glaciers ice like on Banks Island (see major comment above). You identified 11 such RTS on Svalbard, this is quite few, considering the widespread distribution of ice-cored moraine complexes in Svalbard and certainly other Arctic islands.

Line 222 ff: The rate is not really described in the results, this paragraphs mentions it but then goes straight to the acceleration. However, the rate could be interesting, see Fig. 4d: Why do the Arctic Lowlands have such a higher rate?

Line 222 ff: How much sense does the analysis of mean acceleration make if the pattern show such pronounced peaks (see Fig. 4b)

Lines 321 and 336: the carbon loss is first given for decade, later again annually, confusing as only the exponents change when converting decadal to annual values. Suggest deleting the number in line 336.

Line 378: What about uncertainties regarding the ERA5 data? Uncertainties regarding SOC as only upper 3m are used? Uncertainties in SOC database in general? (see also major comment above)

Line 418: Is the mentioned uncertainty of 0.3-0.5m from literature or own validation?

Line 420: Concerning the removed DEMs, did this have an impact on the overall analysis and results?

Line 435: Why "e.g. larger than 3 times the RMSE"? Are there other significance criterions?

Line 438: How is the uncertainty calculated?

Line 444: What is b? Explanation not clear if reader is not familiar with Dai et al (2022).

Line 463 ff: How were the numbers set for the classification of the confidence level?

Line 478: Why criteria 3, is it necessary?

Line 602: June 2012?

Line 613 ff: Can the area-frequency distribution just be extrapolated to smaller RTS? Uncertainties? You can see that in Fig. S11a, the linear fit does not work for large areas. Why should it work for small ones?

Line 745: Link for RTS shapefiles not found (see comment above)

Line 1292: Do you mean "exclude winter scenes"?

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

See also uploaded doc file, with same content.

General comment

I am not at all a specialist of neither remote sensing nor deep learning methods. Therefore my comments will concentrate on general considerations and formal issues.

The paper offers an overview and pan arctic quantification of volume loss and soil organic carbon loss by retrogressive thaw slumps obtained through deep learning on a comprehensive DEM dataset.

Compared to the already very rich literature on the subject, partly produced by the author team members themselves, the novelty of the contribution can be summarized as follows:

- First evaluation of the phenomenon on a pan-arctic scale (previous works were done on regional levels)
- First quantification of soil loss volumes on a decadal and annual time scales (previous works focussed on areas)
- First estimation of soil organic carbon released by these slumps
- The systematic use of the ArcticDEM data set.

The methodological approach seems to be robust. It uses a high resolution dataset, robust and recognized co-registration and extraction algorithms, several field evidence datasets for calibration and validation. The discussion on uncertainties, error sources and limitations is honestly done (text and mainly supplementary information) and gives a realistic idea of the reliability of the results.

The methodological developments (online methods) shows however that the approach needs a whole set of adjustments and model runs, and still gives approximative results. The results still need a huge work of visual/manual correction, which was done by the team. From this point of view, the paper should possibly be regarded rather as a methodological development paper than for its quantitative results.

Some specific comments.

Title

The title doesn't really reflect the content, and is written in too general terms. I suggested a more precise wording, like:

Pan-arctic quantification of total decadal and annual volume loss and soil organic carbon release by large retrogressive thaw-slumps during the decade 2012-2022.

In fact, the work doesn't quantify the total mass loss, but only the retreat during the considered decade. And it is explicitly restricted to retrogressive thaw slumps, excluding other mass wasting processes. This should be reflected in the title.

Area estimates – lines 161-164

Significance of percentages given in brackets is unclear: ... in areas with high ground ice content (1.37 ha 1000 km²; Fig. S10; > 20 %), ... etc.

20 % of what ? ice-content ? of the data set ? ... ?

Discussion – RTS Activity in response to Climate Trends – lines 252-280 + fig 4 and caption + fig 5 caption + fig S13

Text gives thresholds at 0.5, 0.3, -0.3, -0.5 for both temperature and precipitation. Same in figure 4 caption, and figure 5 caption

But in figure legends of figure 4, figure 5 and figure S13, the thresholds are given at 0.6, 0.2, -0.2, -0.6.

Correct either the figures or the text and caption, according to the thresholds used for the pie charts.

Soil organic carbon release

The discussion on possible much lower SOC release than previously modelled is interesting. However, after reading the methodological part, and regarding the large limitations and uncertainties, it seems that your quantification rather underestimates the total volumes (missing landforms, limited considered depth, ...) and should therefore be considered as a minimum (but I may have misunderstood the final result). May be this should be made more clear and less affirmative ?

Sources of error and analysis limitations

This sub-title is a bit peculiar at this place in the manuscript ! The following text is neither a conclusion nor a real uncertainty analysis.

This part should rather be placed right after the results, and before the discussion. It should best integrate the whole text S1, which is important for understanding.

Figures

Most figures are OK and informative.

Figure 4

The figure 4a is small and contains too much information. Remove single dots and leave only the pie diagrams ?

Figure 5

Figure 5b: is the linear fit really significant ? Visually there is no fit at all ...

Supplementary material

Good and useful for understanding.

The part on "Sources of error ..." should be more explicitly referred to in the text, since it is essential for understanding the limitations of the study.

References and reference list

There is insufficient check of citations and reference list:

- Many references cited in the text are missing in the reference list (highlighted red in the edited PDF).
- For several references there is a wrong date, or an ambiguous date (e.g. missing precision for 2022a, 2022b) (year highlighted red in PDF).
- Many references in the list are not cited. They have to be removed from the list (highlighted in red in the list).
- Formatting of references in the list: many journal titles are not in italic (highlighted green in the PDF).
- Some publications years do not correspond between text and reference list (year highlighted purple in PDF).
- Some references near the end of the list are not in alphabetical order (highlighted with purple line number in PDF).

Philippe Schoeneich

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I wish to thank the authors for their careful consideration of my and other reviewers' comments. I see that the manuscript is now improved in its methodological clarity and overall structure of the presentation. As a minor comment, I think that it could be considered if the text on the contribution of small RTS disturbance (<10,000 m²) to carbon release in the discussion should be already presented in the results section ("Soil organic carbon release"). Moreover, around lines 418-427 in the discussion there is a lot of repeating information from the Results, section "RTS Activity in Response to Climate Trends". Apart from these notions, I recommend publication.

(Remarks on code availability)

I am not sufficiently qualified to assess the code in the provided format and language.

Reviewer #2

(Remarks to the Author)

Revision Review "Volumetric quantifications and dynamics"

The manuscript by Dai et al quantify retrogressive thaw slumps (RTS) in the northern permafrost area in terms of location, time of development and volume. The data is generated by comparing multitemporal DEMs based on ArcticDEM and identifying RTS based on a machine-learning algorithm. The data set not only identified polygons of RTS, but also the volume of the landform. This in turn could be used to estimate possible carbon release due to permafrost thaw and erosion. This combination makes the manuscript and its data base highly valuable for the research community, and this study is as far as we know unique and absolute worth publishing.

We now have received a revised version of the manuscript, and we are pleased to see that many of our suggestions from the initial review has been followed, and that most of our comments are incorporated. The paper appears now much better, with a new title, better division of results and discussion, a improved discussion of landforms, incorporation of a statistical analysis of previous years temperatures and major improvements of many figures. However, some issues remained:

1. Download of the shape files: It is now possible that the individual RTS can be loaded into a GIS, but they are simply numbered, so you have to download each one individually to find the one for a certain region. Maybe the authors could divide it into subfolders with the individual regions. In addition, we couldn't load the shapefiles with the coordinates for all slides. Could possibly be due to the fact that e.g. the PRJ file could not be downloaded (opens only in a separate table).
2. The discussion is lengthy, and readability would be improved if divided into subsections with own headings, e.g. "Limitations and uncertainties", "Volumetric yield estimates", "Relation to climate trends".
3. Concerning climate relationships, avoid the term "climatic trend" for a 10-year time series. Determining real trends in climate demands at least 30 years, and we are still critical to the significance of figure 5b, even if the figure has improved. It is also normal to give R²-values (not only simple r-values). R² would give below 25% of explanation, which is not much. It does not mean that climate is not influencing, but you cannot show it statistically significantly on a certain level in your study.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

The main issues raised by my review were correctly addressed in the new manuscript version.

This can thus be published.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)

My comments focus on machine learning section only. Overall, the methodology presented in the manuscript is rigorous, reliable, and represents a well-justified technical approach.

1. In this manuscript, deep learning primarily serves as a tool to reduce manual workload, utilizing existing adapted methods. The main novelty lies in integrating change detection and classification into a unified framework, with prior knowledge incorporated via three selected feature bands.
2. For machine learning-based land surface information extraction, training sample distribution is critical. As noted in the manuscript, performance varies across land cover types, indicating limited spatial transferability of the method. Thus, training samples should comprehensively located in the study area.
3. The two-stage iterative training aligns with active learning principles, which is meaningful. While model ensemble yields marginal improvement, the method (e.g., voting or other techniques) should be clarified. What is the size of sample blocks?
4. Using 10% of samples for accuracy validation may slightly overestimate performance compared to ground truth, but this does not undermine the conclusions.
5. The core identification relies on predefined rules (e.g., elevation change, area thresholds). Machine learning here functions as a labor-saving tool rather than a solution driver, so high methodological novelty is unnecessary.

(Remarks on code availability)

The authors provided a README document and a tutorial for reproducibility. The workflow primarily relies on cited third-party tools, with minimal custom-developed components. Consequently, I did not conduct in-depth testing or evaluation of the code.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reviewer #1 (Remarks to the Author):

The manuscript by Dai et al. is a highly ambitious piece of work that presents novel information on the dynamics of RTSs at the Pan-Arctic scale. By using large amounts of ArcticDEM data, the analysis covers the entire Pan-Arctic at a very high spatio-temporal resolution. The approach shares aspects (using ArcticDEM at Pan-Arctic scale with deep learning approaches) with the work by Huang et al. (2023 ISPRS Journal of Photogrammetry and Remote Sensing, <https://doi.org/10.1016/j.isprsjprs.2023.10.008>). However, the present manuscript goes beyond identifying RTSs by presenting spatially and temporally detailed volumetric estimates of soil loss and associated carbon release. The found magnitude of carbon release from RTSs provides an important new viewpoint to the globally relevant issue of permafrost degradation and associated carbon budgets. The manuscript also makes an interesting point by showing how volumetric RTS activity demonstrates latitudinal and regional differences in their responses to climate, but also that ecosystem properties markedly complicate this connection.

A comment that needs to be made is regarding the relatively weak classification precision of the presented deep-learning method. High number of false positives is typical for approaches using automated recognition of landforms, and I appreciate the authors for addressing this challenge and pursuing to improve the automated mapping by combining solutions (change detection algorithm and deep learning) and using additional training data with data augmentation methods. Yet, also after this model precision remains low and coming up with reliable results requires undoubtedly laborious visual inspection.

Response: we carried out the evaluation in another test area in an Arctic desert environment with little vegetation, which gives a better precision (65%). Now we add the results to Table S3 and a paragraph to discuss this in the Methods section (Step 3b). We further note that the inventory was fully quality-checked by permafrost experts so that the samples included in the inventory are of the highest confidence.

On the other hand, the outcomes of the visual inspection are also affected by which locations are selected as candidate RTSs with the initial model and augmented models. Please see my specific comments regarding this aspect below.

Response: We now add: “Furthermore, the impact of the RTS missed in the deep learning detection is likely around 20% (recall is high, 77% to 80% - Table S3).”

The results are interesting and novel and should be of broad interest to the permafrost research community. The manuscript has an easy-to-follow structure even though the methodology presented involves several processing steps. The figures are clear and interpretable, and the language seems to me (as a non-native speaker) to be in very good order. Below I raise a couple issues and add minor suggestions for corrections and improvements. I recommend publication of the article after my raised issues have been addressed.

Specific comments:

rr166-183: Here, the authors describe how the mapped RTS distributions are consistent with previous studies. I would also like to see a discussion of whether there are any regions or

environments where RTS activity could not be mapped with a similar precision, or if the results contradict previous findings. Now, the discussion of uncertainties centers on the large number of false positives but the potentially missed RTS activity is less addressed. Although, the limitations in detecting some established activity at megaslumps due to narrow horizontal scale of activity was satisfactorily explained, which indicates that the approach may not have high recall in all conditions. Do we know if model evaluation statistics (from prediction and evaluation with Victoria-Peel Plateau data) hold for the Pan-Arctic area or can there be regional/environmental variation?

Response: we carried out additional evaluation in the Arctic desert environment, which gives a good precision of 65% and high recall of 77%.

We now add: “The prediction performance shows regional variations due to different land cover and environmental settings such as vegetation cover, snow and ice cover, and water bodies. Further evaluation is carried out in Eureka, Canada (Fig. S5c,⁵⁶), which represents the Arctic desert environment with little vegetation. There are 45 RTS as ground truth in this test area manually examined by subject experts. Due to minimal impact from vegetation changes, the ensemble model achieves a considerably higher precision of 65% and a recall of 77% (Table S3).”

To address potentially missed RTS we add: “the impact of the RTS missed in the deep learning detection is likely around 20% (recall is high, 77% to 80% - Table S3).”.

We also discuss additional visual examples in the Discussion with locations from the literature of regions and environments where RTS activity could not be mapped because of our minimum size and elevation change threshold.

rr517: The split ratio of 9:1 sounds reasonable. Were other ratios tried in model training to examine their effect on overall accuracy?

Response: The split ratio 9:1 seems to be the widely used ratio. No other ratios were tested.

rr563-564: Were the samples of 1660 training and 189 validation images also here split randomly based on the manually selected and inspected 1660+189 images? Please elaborate because models should be affected by how well the variation in the modelling data is represented.

Response: Yes, they are split randomly.

Now we have: “The dataset is randomly split into training and validation sets at a 9:1 ratio, resulting in 1660 training images and 189 validation images (Fig. S7).”

rr576-578: I find it quite surprising how the combination of five models (original + augmented) increase precision so drastically. Model ensembles often do outperform single models, but the improvement is surprisingly large. Maybe single models identify different FPs and when put together, there are not so many common FPs, and thus the precision is higher? Could this be elaborated?

Response: We now add Fig. S6b to illustrate the improvement from ensemble modeling. The improvement (false positives from 357 to 210) is not very drastic, now we soften the tone in our sentences.

Now we have: “since individual models may detect different false positives, when combined, the overlap of common false positives reduces, hence enhancing the precision (Fig. S6b). For the test area in Peel Plateau, the ensemble combination of five models (Table S3) reduces the number of false positives from 357 to 210.”.

r745: I could not access the produced data via the (supposedly) DOI link (Mendeley?). Are the data available already?

Response: The data link was not ready. It should be ready now for download in the link provided in the data availability section. The associated metadata is still under review at Arctic Data Center.

Fig. S7-8. The final detected RTSs are seemingly concentrated in the training data (both 1 and 2) locations (as should be expected) while relatively little RTSs are detected in some areas known to have abundant occurrences. For example, only 13 RTSs have been detected in Alaska. I was wondering why there were no training images in Step 3b selected from Alaska at all? Could this be because the initial model from Step 3a could not predict potential RTSs there due to different environmental conditions in those areas, and consequently that also Step 3b results in very little RTSs in the western part of Brooks Range, for example? Or are these possibly “missing” RTSs simply filtered out in the classification in Step 2 for being too small or not active during the study period?

Response: No training images in Alaska were selected because the number of candidates in Alaska is too high, which is laborious for visual inspection. As shown in table S4, there are 19,154 candidates from step 3 (multimodel ensemble) for North Alaska. The number of candidates from step 3a (model 1) is normally higher, hence they are not missed in Step 3a.

Indeed, a lot of slumps in the western part of Brooks Range are filtered out in Step 2. We checked and now visually address in the manuscript the published slumps in (Swanson and Nolan, 2018; Swanson, 2021). Two of those are detected in our study. The rest is either too small or filtered out in Step 2 because no pixels within the scarp area exhibited a significant elevation change exceeding 2 m, thereby failing to meet Criterion 2.

We added Fig. S14h to illustrate these cases. And we also added further clarification at the beginning of the results section and elsewhere throughout the document about our minimum sampling unit (minimum >10,000 m² and minimum elevation change of 2 m) and make further direct statements such as “ Due to the minimum size or elevation change threshold adopted in this work, our dataset also excludes many small or shallow slumps characterized by smaller elevation changes (e.g., Fig. S15h,^{26,70}).”

Technical touches:

r99. Replace eg with e.g.

Response: corrected.

Fig. 1 a, b. Imagery source not visible, please enhance.

Response: We have added the following text at the end of the figure description to clarify the image sources: ESRI basemap imagery was used for (a), (b) and (d) accessed via ArcGIS Pro v2.7, Sources: Esri, DigitalGlobe, GeoEye, i-cubed, USDA FSA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

r223: 3gh -> 3g,h

Response: Here Fig. 3h is more relevant. We change it to Fig. 3h.

Table 1: The pink and grey colors are hardly separable. Consider adding contrast.

Response: Changed as suggested.

Fig. 4: The legend for the correlation coefficients has different ranges to those presented in the figure caption and elsewhere in the text. Which are correct?

Response: The legend in the figure is now updated to be consistent with the caption.

rr355-360: This sentence could be more clearly formulated. Is there a comma missing? Consider revising.

Response: We decide to retain this sentence.

r643: Please elaborate the idea behind selecting the growth rate of double the long-term rate in Segal et al. 2016?

Response: Segal et al determined the mean standardized/normalized growth rate of 0.67% by considering changes between 1950-1960 imagery and 2004-2008. This growth rate is too low to reflect 21st century conditions as many studies from various areas have reported a recent intensification of thaw slumping. These intensifications are often reported in the literature as exponential, but to remain conservative in our assumptions we simply doubled Segal's standardized growth rate to 1.3%.

We clarified the sentence by providing more details: "We multiplied the scar area estimates with a mean standardized growth rate of 1.34% (i.e., twice the long-term rate observed between 1950-1960 and 2004-2008 in ⁵⁵ to account for RTS intensifications in the last two decades) to obtain an estimate of annual area change."

r652: "...for a pan-Arctic total of $\pm 12,514$ ". Please check if this is correct. Are confidence intervals/standard deviation for the value missing here or how should one read the value?

Response: these should be read as one value. The plus/minus was removed in the text (and was supposed to indicate an approximation, not a confidence interval).

r940: Please check the text font.

Response: Font corrected.

Figs. S7: This is a bit rough figure. Consider using the same longitude of origin for the polar projection that was used in the other maps for better comparability. Please, also acknowledge the imagery source and consider adding a legend in which training and validation images are preferably presented with unique colors.

Response: Figure revised in polar stereographic projection. Google Earth image is acknowledged.

Reviewer #1 (Remarks on code availability):

Unfortunately, I only found the Github link now as I am submitting my assessment. Anyway, I use different coding languages in my work, so I would not be able to test the reproducibility of the code or provide a sufficient assessment.

[Response: Now we add a code availability section and instructions.](#)

Reviewer #2 (Remarks to the Author):

The manuscript by Dai et al quantify retrogressive thaw slumps (RTS) in the northern permafrost area in terms of location, time of development and volume. The data is generated by comparing multitemporal DEMs based on ArcticDEM and identifying RTS based on a machine-learning algorithm. The data set not only identified polygons of RTS, but also the volume of the landform. This in turn could be used to estimate possible carbon release due to permafrost thaw and erosion. This combination makes the manuscript and its data base highly valuable for the research community, and this study is as far as we know unique and absolute worth publishing, even if the time series of 10 years is too short to evaluate real climate trends.

We are not experts on the machine-learning algorithms and let this part be evaluated by other reviewers. However, the following major comments are given and may be considered:

- Title: Be precise, you deal (at least according to your writing) almost exclusively with retrogressive thaw slumps (RTS), why write then “large permafrost-thaw driven mass wasting events”? See also comment below.

[Response: we now change the title as “Volumetric Quantifications and Dynamics of Areas Undergoing Retrogressive Thaw Slumping in the Northern Hemisphere”.](#)

- Type of RTS: We are not always sure what landforms you have detected. Of course, the big RTS in Arctic lowlands are captured and exemplified in images in the paper. However, in the Arctic islands, e.g. on Svalbard, you mentioned large RTS developed (by thermal erosion?) in large recent (< 250 years old) ice-cored moraine complexes in front of recent glacier fronts. Those those would have a quite different development history than ground-ice induced thaw slumps, even in areas of buried old Pleistocene glacier ice (there is a large literature base on this thermal erosion in ice-cored moraines on e.g. Svalbard). Maybe, you should show different types of identified RTS from the various regions? We wanted to have a look in your data for e.g. Svalbard, but the link provided failed.

[Response: We now add a Fig. S13 to show examples of different RTSs such as thermoterraces and glacier dead-ice back-slumping in Svalbard. The data link should work now.](#)

[We also made further clarifications with respect to the scope of the study and the types of thaw-driven slumping we included.](#)

[“Other areas undergoing RTS were detected along current glacier margins in northern Greenland, Svalbard, near Klutlan Glacier, Canada; and in Novaya Zemlya, Russia, where glacier dead-ice backwasting occurs \(Fig. S13f\) \(e.g., ⁶⁰\). There are considerable geomorphological differences between dead-ice backwasting in recent ice-cored moraine complexes and RTS, yet the small number of large dead-ice backwasting detections \(> 10,000](#)

m²) were not manually excluded from the inventory as both mass wasting phenomena occur on ice-cored slopes, are thaw related, and exhibit similar elevation change “rainbow” patterns due to their retrogressive, rotational sliding failure mechanisms^{39–41}. Furthermore, the strong linkage between RTS detections with modern-day glacier margins is compatible with RTS activity across northwestern Canada where terrain encompassing the former glacial margins of the Laurentide Ice Sheet contains relict glacier ice preserved by permafrost⁶¹.”

- Correlation of events to temperature and precipitation: First of all, the results of the correlation analysis, which is interesting, is in the Discussion (chapter 3.1.), but it should be a Result. Consider moving the first part of that chapter (l. 252. onward).

Response: The first part Section (RTS activity in Response to Climate Trends) is now moved to Results.

You then write that 60% of the RTS are neither correlated with trends in ERA temperature or precipitation. I am not sure how you made this correlation, but an event one year is not necessarily correlated with climate the same year, but with the weather development the previous year. E.g. an extreme summer one year may lead to weakening of the structure and failure in one of the subsequent years. You should consider this in your discussion (see pre-conditioning papers in high-mountain permafrost slopes, e.g. here <https://iopscience.iop.org/article/10.1088/1748-9326/ad3c55>). This pre-conditioning may be more important further south than in the northernmost parts?

Response: We thank the reviewer for pointing this out, and in turn we carried out additional analyses concerning the correlation with the climate (temperature and precipitation) from the previous year. Indeed, the preconditioning effect is more prominent in the south than in the north regions. We added the results to a new Fig. 5a.

We also fully updated the results and discussion text considering same-year and previous year preconditioning effects.

- Soil organics release: These are nice data, again, why do you not present these as a result, and then discuss them later.

Response: The section (Soil organic carbon release) is now moved to Results.

Line 341 ff: The reference Tanski et al. (2017) does not support the statement of less SOC release. Here, the numbers are just small because SOC is buried again. However, in this study, the deposition of SOC is not taken into account, so that your numbers should be compared only to the released SOC.

Response: We have removed this reference from the manuscript. In places where field studies are used to support our argument that modelling studies have overestimated carbon release amounts we clarified our language used by specifying carbon released and net carbon amounts, modified text reads: “While our study shows less overall carbon release from RTS process, field studies^{19,20} on the Peel Plateau have shown that released carbon is sequestered within redistributing and accumulating sediments (e.g. debris tongues, Figure 1c) by becoming chemically bound through interactions with mineral soil. While variations in net carbon release vary regionally⁷⁷, likely due to variation in carbon content of near-surface and deeper permafrost¹⁹, overall, our study along with these recent field studies^{19,20,77} point to net carbon

released by RTS to be of lower magnitude than initially suggested by modelling¹⁸ and merits future work to better constraint values.”

- Discussion in general: The Discussion in general could benefit from a more detailed elaboration of the results from Fig. 3, which is nice, but not discussed further in the paper (just presented as a result). E.g., why do you observe such areas, volumes, depths, rates and accelerations? Is air temperature and precipitation the only explanation as you focus on this in the discussion? What about ground ice contents, topography, geomorphology etc.?

Response: we added a new paragraph in the Discussion to elaborate on Fig 3 and what it means in the larger context of RTS studies. We also clarified our scope of this work as detection and description of patterns, but acknowledge that the drivers/reasons behind the areas/volumes/depths/rates/accelerations is an area of future work.

Specifically about the scope we added: “ It was beyond the scope of this study to determine the drivers behind the regional differences in volumes, depths, yields, scaling coefficients, or momentum, which should be considered in future work. Drivers governing these differences are most likely related to regional topographic slope configurations, subsurface conditions, landscape genesis (e.g., glacial legacy), and the most recent disturbance or thaw history (e.g., presence of old or ancient RTS disturbances, wildfire) ^(31,42,61,64,67). ”

- Comment on the so-called “Rainbow pattern”: What about RTS where activity started or ended during the observation period and thus show no or only a weak rainbow pattern? Are these excluded? If that is the case it should be discussed in the uncertainties.

Response: We added a new paragraph in the Discussion to describe the uncertainties in more details, including weak rainbow patterns, RTS below the minimum size threshold or below the minimum elevation threshold. In the carbon release discussion we further added: “Nevertheless, the carbon release is likely underestimated because some RTS are missed due to a less prominent time transient change pattern if only a single year of retrogression occurred during the observation period or if the disturbance area was affected by surrounding noise.”

- Uncertainties: A major result of the discussion is the relationship between RTS activity and climate, and the influence of RTS on carbon release. This discussion is based on ERA5 data and SOC map by Hugelius et al. Here, we miss a more in-depth description of the uncertainties in ERA5 and the SOC map, which of course influence the conclusions. soil organics carbon map.

Response: For the SOC results, we have added the following text to the discussion to clarify uncertainties:

“Our calculated value of approximately 1.95×10^{-3} Pg carbon per year within our time period, which is the first attempt at this spatial scale with observed volumetrics, is much less (~0.2%) compared to what was estimated (using modeling) to be released annually from gradual thawing permafrost (~1 Pg carbon per year)¹¹. It could be argued that a large number of small disturbance areas undergoing RTS (area < 10,000 m², small elevation changes) not considered in this study could lead to underestimations of carbon release. To estimate this effect, we used area-frequency distribution modeling and comparisons with detailed manual RTS inventories to estimate the total RTS population size and volumetrics of small areas undergoing RTS (Figs. S11a, S12),

resulting in an additional 2,244 to 12,514 areas or 18.0 to $63.8 \times 10^6 \text{ m}^3$ (see Methods). The inclusion of small RTS increases the total Northern Hemisphere volume estimate from $317.0 \times 10^6 \text{ m}^3$ to a range of 335 to $381 \times 10^6 \text{ m}^3$, and shows that the effect of undetected smaller RTS on carbon release is small, comprising only 5-17% of the estimated total volume. Furthermore, the impact of the RTS missed in the deep learning detection is likely around 20% (recall is high, 77% to 80% - Table S3). Nevertheless, the carbon release is likely underestimated because some RTS are missed due to a less prominent time transient change pattern if only a single year of retrogression occurred during the observation period or if the disturbance area was affected by surrounding noise. The mean depth of elevation loss is 3 m, so our estimates represent at least average conditions within the study domain.⁵⁸ found that the depths exceeding 3 m contributed to 15 to 20 % volume change and 5 to 10 % of total soil carbon mobilization. In our dataset, excluding carbon at depths greater than 3 m represents 30 % of volume with no carbon estimate. Ground ice content was not included in our carbon release estimate because the uppermost layers are comprised of the active layer and ice-poor permafrost above massive ice¹⁹. We believe the underestimation of not including the deeper ground layers offsets this potential overestimation for RTS with ground ice within the upper 3 m of soil.”

For uncertainties of ERA5, we have: “Limitations also arise from the heterogeneity in the accuracy and representativeness of ERA5 trends due to the limited number of Arctic weather stations and the short amount of yearly measurements ($n \sim 10$, Fig. 2d) used in this study. ”

- Figures: Some of the figures are a bit messy and should be revised. A list of suggestions are given below in the minor comments.

Minor comments are as follows:

Figures and tables:

Fig. 1: Scales are lacking in a, c and d. Lat/long are missing in at least a and b.

Response: Scales are added to a, b and d. Lat/Long information has been added in the figure description: Examples of slump landforms in the Eureka Sound Lowlands ($\sim 79^\circ \text{N}$) that encompass south-eastern Axel Heiberg Island and west-central Ellesmere Island, Canada....The example in (b) can be found at 79.49°N ; 86.84°W and (d) at 79.61°N ; 86.23°W .

Is c and d the same RTS? If yes, which one of the RTS shown in b?

Response: No the field example is not the same as the remote sensing examples but they occur in the same area, this clarification has been added to the figure description: (c) shows a field example of a slump (different from those shown in (b) and (d), but from the same area) taken from a helicopter in 2016.

Fig. 2: Scale is missing in a, no lat/long grid. D show temperature, given again in Figure 4b? Can be avoided here.

Figs. 2d and Fig. 4b: Gives the same information, only once for a specific RTS and once for all RTS. Maybe Fig. 2d is redundant (only mentioned in the discussion as an example where one could refer also to Fig. 4b)? At least have the same axis labels and legend.

Response: Scale and lat/long are now added.

We prefer to keep Fig. 2d, as it shows the correlation for individual RTS, as a nice transition to Fig. 4a.

Fig. 3: Nice figure, but not discussed, see comment above

Response: Fig. 3 is now discussed in the Discussion.

Fig. 4: The figure is a bit messy, lots of different information, do you need the big pie diagrams? They more or less are similar. D is very messy, what do you want to show? High loss rates for arctic lowlands are never discussed really.

Response: We deleted the pie diagrams from the figure. In the Results and Discussion we expanded on the loss rate information. For example, we provide explicit reasons as to why the Arctic Lowlands exhibited a higher volume loss rate (due to its large RTS population).

Fig. 5: This figure may be combined with Fig. 4b because of redundancies.

Response: We now add the correlation with the previous year temperature and precipitation to Fig. 5a, which shows the unique latitudinal variations in their responses to climate, showing different information from Fig. 4b.

The trend in 5b is really $r > 0.5$? To be honest, for us the distribution of the red dots is quite random.

Response: We clarified in the discussion that the direct correlation is $r = 0.23$, and stated “a positive yet weak correlation was observed...”. We further added an error bar for the fit in Fig. 5b. “

Fig. S2: We find the explanation of the figure is not clear: We assume that the red area around P1 is the deposition area? Why is the deposition earlier (2011) than the retreat of the headwall (2012-2015, indicated by Fig. S2g)? This is also corroborated by Fig. S2f, where deposition was in 2011 but head wall retreat until 2016? How can this be explained? Is it because only the largest elevation difference is analysed (see Line 1281 ff)? Then maybe mention it briefly here and not only in the supplement.

Response: The earliest retreat in Fig. S2g started in 2011 (dark blue areas). Yes, the largest deposition seems to occur in 2011. The deposition from later years may flow and spread out downstream.

In the caption, we now add: “The deposition at this point (P1) appears to have occurred in 2011, while the deposition from later years may have flown and spread out downstream.”

“The estimated time identifies when the most substantial elevation change occurred within the time series.”. We also added additional narrative about retaining only the largest elevation difference in the Discussion (second paragraph).

Finally in Fig. S2b and e: Why are there so many elevations in 2015 available and only one for the other years?

Response: For local areas, our dataset (ArcticDEM) may have more data in some years and less data in other years. But generally, it has data coverage in most years. This is discussed in the text, e.g., we have: “an overall limitation of ArcticDEM is that the temporal resolution varies throughout its geographic coverage depending on the availability of stereographic pairs of images used to create the DEMs. Therefore, the temporal and geographic coverage of DEMs is not uniform across the study region.”.

Fig. S11a: The blue dots in a, are those all RTS? 2747? This is not possible.

Response: Yes, blue dots represent the histogram of 2747 RTS. Each dot is not an RTS, rather represents the probability density for RTS in each area bin (bin width 1000m²).

Now we add: “ $p = c/(Nw)$, where c is the number of RTS within an area bin, w is the bin width (1000 m²), N is the total number of RTS (2,747)”

Table 1: Maybe add a column with loss rate. The loss rate is shown in Fig. 2a but it would be nice to have a summary for each region.

Response: A column of total volume loss rate is now added to Table 1.

Other comments:

Line 37: Add the exact time period 2012-2022

Response: Changed as suggested.

Line 98: You state that the existing literature focus on one or two-dimensional detection and measurement of RTSs, but give only an example for 2D detection. Please add another one for 1D detection.

Response: we’ve provided additional references to support this statement.

Line 162: Information in the brackets is unclear: Why give it in ha/ 1000 km²? This is not really intuitive. And what do the percentages describe? Many of such numbers in the text decrease readability, better to use a table to summarise numbers.

Response: we clarified what the percentages describe (ice content). There are different ways of providing density information, with RTS areas usually given in hectares and study areas in squared km. Also, because the total RTS area across the Arctic is small (and the study area is large), we felt that “1.37 ha 1000 km⁻²” was more intuitive than using the same area units, e.g., 0.0137 km² 1000 km⁻².

Line 177ff: RTS and ice-cored moraines. We think there is quite a difference between recent ice-cored moraines like on Svalbard and old buried glaciers ice like on Banks Island (see major comment above). You identified 11 such RTS on Svalbard, this is quite few, considering the widespread distribution of ice-cored moraine complexes in Svalbard and certainly other Arctic islands.

Response: Our identified number is small probably due to the area threshold (>10,000 m²).

We now add more notes in the Table S4 to clarify this: “ This table only includes medium- to large RTSs with areas larger than 10,000 m².” and further expanded on the Results section:

“Other areas undergoing RTS were detected along current glacier margins in northern Greenland, Svalbard, near Klutlan Glacier, Canada; and in Novaya Zemlya, Russia, where glacier dead-ice backwasting occurs (Fig. S13f) (e.g., ⁶⁰). There are considerable geomorphological differences between dead-ice backwasting in recent ice-cored moraine complexes and RTS, yet the small number of large dead-ice backwasting detections (> 10,000 m²) were not manually excluded from the inventory as both mass wasting phenomena occur on ice-cored slopes, are thaw related, and exhibit similar elevation change “rainbow” patterns due to their retrogressive, rotational sliding failure mechanisms^{39–41}. Furthermore, the strong linkage

between RTS detections with modern-day glacier margins is compatible with RTS activity across northwestern Canada where terrain encompassing the former glacial margins of the Laurentide Ice Sheet contains relict glacier ice preserved by permafrost⁶¹.”

Line 222 ff: The rate is not really described in the results, this paragraph mentions it but then goes straight to the acceleration. However, the rate could be interesting, see Fig. 4d: Why do the Arctic Lowlands have such a higher rate?

Response: We expanded the section and clarified specifically why the Arctic Lowlands exhibited higher volume loss rates.

Line 222 ff: How much sense does the analysis of mean acceleration make if the pattern shows such pronounced peaks (see Fig. 4b)

Response: We acknowledge the comment and have revised the statement with further clarification and linkage to the next section (we do want to note that we used momentum, not mean momentum in this analysis).

“Temporal metrics provided insights into the annual dynamics of RTS. The majority (75%) of areas undergoing RTS yielded $<10,000 \text{ m}^3$ annually ($<20,000 \text{ m}^3$ 89.6%), with median yields of $5,013 \text{ m}^3$ (Fig. 3g). Summarizing annual yields across regions (named after Provinces and Divisions as defined by ⁶³), RTS located in the Arctic Lowlands (e.g., Banks Island), Mackenzie Mountains, and Brooks Range (incl. Peel Plateau) exhibited the largest yields of 10.0, 4.1, and 4.1 million m^3/year , respectively, due to their large population of areas undergoing RTS (Table 1). At a combined 18.2 million m^3/year , these three regions contributed up to 56.8% of the observed annual yields across the Northern Hemisphere. The majority of RTS exhibited a relatively stable momentum in their annual yields (i.e., near $0 \text{ m}^3/\text{year}^2$), meaning that annual activity is neither accelerating or decelerating. However, as a whole Northern Hemisphere population, momentum was slightly positive with a mean of $179 \text{ m}^3/\text{year}^2$ (Fig. 3h) indicating a slight acceleration in annual volume loss. These momentum estimates are affected by pronounced annual peaks and lows of activity, but provide useful insights into the net long-term trends of RTS in the Northern Hemisphere as well as sub-regions.”

Lines 321 and 336: the carbon loss is first given for decade, later again annually, confusing as only the exponents change when converting decadal to annual values. Suggest deleting the number in line 336.

Response: We delete the number. And we now only use Pg carbon per year to be consistent.

Line 378: What about uncertainties regarding the ERA5 data? Uncertainties regarding SOC as only upper 3m are used? Uncertainties in SOC database in general? (see also major comment above)

Response: SOC uncertainties and ERA5 uncertainties are addressed in the previous response.

Line 418: Is the mentioned uncertainty of 0.3-0.5m from literature or own validation?

Response: Mainly from literature and our group's other related work.

Line 420: Concerning the removed DEMs, did this have an impact on the overall analysis and results?

Response: Deleting bad quality DEMs has little impact on the overall analysis since bad DEMs only occur occasionally.

Line 435: Why “e.g. larger than 3 times the RMSE”? Are there other significance criterions?

Response: We listed five conditions for identifying confidence levels (significance) in the last paragraph of Step 1.

Line 438: How is the uncertainty calculated?

Response: We now add: “Its uncertainty is estimated based on the posteriori variance of unit weight and error propagation (Dai et al., 2022).”

Line 444: What is b? Explanation not clear if reader is not familiar with Dai et al (2022).

Response: Now we add: “a+b represents the surface elevation after the event.”

Line 463 ff: How were the numbers set for the classification of the confidence level?

Response: The thresholds are empirical.

Line 478: Why criteria 3, is it necessary?

Response: The criterion 3 is chosen to exclude cases where there is one single event within the scarp area. This criterion will help select slumps with multiple thawing events which exhibit the unique rainbow pattern. We now add: “Criterion 3 is designed to identify slumps with multiple thawing events that display the distinctive rainbow pattern.”

Line 602: June 2012?

Response: It’s now changed to: “summer defined as the period from June 20th to September 22nd each year”

Line 613 ff: Can the area-frequency distribution just be extrapolated to smaller RTS?

Uncertainties? You can see that in Fig. S11a, the linear fit does not work for large areas. Why should it work for small ones?

Response: The area-frequency distribution function is commonly used in landslide studies to estimate the number of small, undetected landslides, so indeed it can be extrapolated for this first-order estimate. This is one estimation method we used, which yields the volume of small RTS is about 5-8% of the entire population. This estimation generally agrees with the second method in the next paragraph, which shows that the volume of small RTS is about 17% of the entire population.

Line 745: Link for RTS shapefiles not found (see comment above)

Response: New link is provided.

Line 1292: Do you mean “exclude winter scenes”?

Response: We include winter scenes. To clarify this, we now add: “we had to include winter scenes to ensure adequate temporal data coverage.”

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #4 (Remarks to the Author):

See also uploaded doc file, with same content.

General comment

I am not at all a specialist of neither remote sensing nor deep learning methods. Therefore my comments will concentrate on general considerations and formal issues.

The paper offers an overview and pan arctic quantification of volume loss and soil organic carbon loss by retrogressive thaw slumps obtained through deep learning on a comprehensive DEM dataset.

Compared to the already very rich literature on the subject, partly produced by the author team members themselves, the novelty of the contribution can be summarized as follows:

- First evaluation of the phenomenon on a pan-arctic scale (previous works were done on regional levels)
- First quantification of soil loss volumes on a decadal and annual time scales (previous works focussed on areas)
- First estimation of soil organic carbon released by these slumps
- The systematic use of the ArcticDEM data set.

The methodological approach seems to be robust. It uses a high resolution dataset, robust and recognized co-registration and extraction algorithms, several field evidence datasets for calibration and validation. The discussion on uncertainties, error sources and limitations is honestly done (text and mainly supplementary information) and gives a realistic idea of the reliability of the results.

The methodological developments (online methods) shows however that the approach needs a whole set of adjustments and model runs, and still gives approximative results. The results still need a huge work of visual/manual correction, which was done by the team. From this point of view, the paper should possibly be regarded rather as a methodological development paper than for its quantitative results.

Response: We understand the reviewers' point-of-view but disagree with regarding this manuscript to be solely a methodological development paper. Most deep learning papers have results typically limited to describing accuracy and runtime. In this manuscript we document for the first time a representative quantification of volumetric change from large areas undergoing retrogressive thaw slumping in the Northern Hemisphere, whereby we contribute new insights into the magnitude and manifestation of hillslope thermokarst by means of volumetric sediment budgets, non-linear scaling, and regional patterns We provide context and interpretations based

on transparent estimates and implemented considerable due diligence beyond what a typical remote sensing/methodological paper would provide.

Some specific comments.

Title

The title doesn't really reflect the content, and is written in too general terms. I suggested a more precise wording, like:

Pan-arctic quantification of total decadal and annual volume loss and soil organic carbon release by large retrogressive thaw-slumps during the decade 2012-2022.

In fact, the work doesn't quantify the total mass loss, but only the retreat during the considered decade. And it is explicitly restricted to retrogressive thaw slumps, excluding other mass wasting processes. This should be reflected in the title.

[Response: The title is now changed to: "Volumetric Quantifications and Dynamics of Areas Undergoing Retrogressive Thaw Slumping in the Northern Hemisphere"](#)

Area estimates – lines 161-164

Significance of percentages given in brackets is unclear: ... in areas with high ground ice content (1.37 ha 1000 km⁻²; Fig. S10; > 20 %), ... etc.

20 % of what ? ice-content ? of the data set ? ... ?

[Response: Now it is changed to: ice content > 20%; ice content between 10% and 20 %; ice content < 10%.](#)

Discussion – RTS Activity in response to Climate Trends – lines 252-280 + fig 4 and caption + fig 5 caption + fig S13

Text gives thresholds at 0.5, 0.3, -0.3, -0.5 for both temperature and precipitation. Same in figure 4 caption, and figure 5 caption

But in figure legends of figure 4, figure 5 and figure S13, the thresholds are given at 0.6, 0.2, -0.2, -0.6.

Correct either the figures or the text and caption, according to the thresholds used for the pie charts.

[Response: Legends in Figures 4 and S13 are now updated to be consistent with the text and caption.](#)

Soil organic carbon release

The discussion on possible much lower SOC release than previously modelled is interesting.

However, after reading the methodological part, and regarding the large limitations and uncertainties, it seems that your quantification rather underestimates the total volumes (missing landforms, limited considered depth, ...) and should therefore be considered as a minimum (but I may have misunderstood the final result). May be this should be made more clear and less affirmative ?

[Response: acknowledged. We revised the discussion on this by further integrating the estimates of small slumps that fell outside the scope of our detection models.](#)

“It could be argued that a large number of small areas undergoing RTS (area < 10,000 m², small elevation changes) not considered in this study could lead to underestimations of carbon release. To estimate this effect we used area-frequency distribution modeling and comparisons with detailed manual RTS inventories to model the total RTS population size and volumetrics of small areas undergoing RTS (Figs. S11a, S12), resulting in an additional 2,244 to 12,514 areas or 18.0 to 63.8 × 10⁶ m³ (see Methods). The inclusion of small RTS increases the total Northern Hemisphere volume estimate from 317.0 × 10⁶ m³ to a range of 335 to 381 × 10⁶ m³, and shows how the predicted impact of undetected smaller RTS on carbon release is small (only 5-17% of the estimated total volume). Furthermore, the impact of missed RTS from the deep learning detection is likely low, around 20%, since the recall is high, 77% to 80% (Table S3). Nevertheless, the carbon release might be underestimated considering that some RTS might be missed due to a less prominent rainbow pattern if it had only a single year of retrogression during the observation period or if it was affected by surrounding noises”.

Sources of error and analysis limitations

This sub-title is a bit peculiar at this place in the manuscript ! The following text is neither a conclusion nor a real uncertainty analysis.

This part should rather be placed right after the results, and before the discussion. It should best integrate the whole text S1, which is important for understanding.

Response: we now integrate and further expand on the whole text S1 in the discussion.

Figures

Most figures are OK and informative.

Figure 4

The figure 4a is small and contains too much information. Remove single dots and leave only the pie diagrams ?

Response: We now remove the pie diagrams.

Figure 5

Figure 5b: is the linear fit really significant ? Visually there is no fit at all ...

Response: We now add gray shading to show the standard deviation of the fit. In our discussion, we acknowledge the weak correlation, e.g, we have: “a positive yet weak correlation was observed...”

Supplementary material

Good and useful for understanding.

The part on “Sources of error ...” should be more explicitly referred to in the text, since it is essential for understanding the limitations of the study.

Response: We now integrate the whole text S1 to the main text (Discussion).

References and reference list

There is insufficient check of citations and reference list:

- Many references cited in the text are missing in the reference list (highlighted red in the edited PDF).

- For several references there is a wrong date, or an ambiguous date (e.g. missing precision for 2022a, 2022b) (year highlighted red in PDF).

[Response: Revised as suggested. See the following point-to-point response \(For detailed comments in the PDF file\).](#)

- Many references in the list are not cited. They have to be removed from the list (highlighted in red in the list).

[Response: Revised as suggested](#)

- Formatting of references in the list: many journal titles are not in italic (highlighted green in the PDF).

[Response: References formatting is revised to be consistent with the journal style.](#)

- Some publications years do not correspond between text and reference list (year highlighted purple in PDF).

- Some references near the end of the list are not in alphabetical order (highlighted with purple line number in PDF).

[Response: Revised as suggested](#)

For detailed comments in the PDF file:

Line 42: ‘Moreover’ -> ‘However’

[Response: Revised as suggested.](#)

Line 71: Smith et al., 2023

[Response: Reference added.](#)

Line 74, 83: IPCC,2023

[Response: Citation corrected.](#)

Line 77: Ward Jones et al., 2022; Kokelj et al., 2023

[Response: Corrected.](#)

Line 79, 90: Turetsky et al., 2020

[Response: Reference added.](#)

Line 88: Sahkil et al., 2021

[Response: Reference removed.](#)

Line 96: Zwieback et al., 2018

[Response: Reference added.](#)

Line 97: Bernhard et al., 2022 a or b?

[Response: Corrected as: Bernhard et al., 2022a](#)

Line 110, 177: Nitze et al., accepted

[Response: Updated as Nitze et al., 2024.](#)

Line 115: Burn and Lewkowicz, 1990

[Response: Reference added.](#)

Line 179: Kjær and Krüger, 2001

[Response: Reference added.](#)

Line 263: Edlund and Garneau, 2000

[Response: Reference added.](#)

Lines 331-332: Turetsky et al., 2020

Response: [Reference added.](#)

Line 343, 355: Shakil et al., (2020)

Response: [Reference added.](#)

Line 395: Dai and Howat, 2018

Response: [Reference is now added.](#)

Line 610, 1384: Welch, 1967

Response: [Reference is now added.](#)

Line 634, 687: Bernhard et al., (2022) “a or b?”

Response: [Corrected as: Bernhard et al., \(2022a\)](#)

Line 644: Segal et al.; (2016) “a, b or c?”

Response: [It is Segal et al., 2016a. Corrected in the new citation.](#)

Line 683: Larsen et al., 2010; Tseng et al., 2013

Response: [Reference is now added.](#)

Line 726: Wessel and Smith, 1991

Response: [Reference is included.](#)

Philippe Schoeneich

Reviewer #1 (Remarks to the Author):

I wish to thank the authors for their careful consideration of my and other reviewers' comments. I see that the manuscript is now improved in its methodological clarity and overall structure of the presentation. As a minor comment, I think that it could be considered if the text on the contribution of small RTS disturbance ($<10,000 \text{ m}^2$) to carbon release in the discussion should be already presented in the results section ("Soil organic carbon release"). Moreover, around lines 418-427 in the discussion there is a lot of repeating information from the Results, section "RTS Activity in Response to Climate Trends". Apart from these notions, I recommend publication.

Response: Thank you for your constructive suggestions.

We now move the text about small RTS to the results section.

As suggested, we delete the repeating details: "... (from 40% to 88% of RTS binned by latitude with $r \geq 0.5$ as latitudes increased from 71°N to 82°N). Precipitation-driven RTS decreased linearly from 60% to 13% with increasing latitude from 60°N to 82°N ."

We keep a brief summary of results as a transition to the discussion of the explanation of the phenomenon.

Reviewer #1 (Remarks on code availability):

I am not sufficiently qualified to assess the code in the provided format and language.

Reviewer #2 (Remarks to the Author):

Revision Review "Volumetric quantifications and dynamics"

The manuscript by Dai et al quantify retrogressive thaw slumps (RTS) in the northern permafrost area in terms of location, time of development and volume. The data is generated by comparing multitemporal DEMs based on ArcticDEM and identifying RTS based on a machine-learning algorithm. The data set not only identified polygons of RTS, but also the volume of the landform. This in turn could be used to estimate possible carbon release due to permafrost thaw and erosion. This combination makes the manuscript and its data base highly valuable for the research community, and this study is as far as we know unique and absolute worth publishing.

We now have received a revised version of the manuscript, and we are pleased to see that many of our suggestions from the initial review has been followed, and that most of our comments are incorporated. The paper appears now much better, with a new title, better division of results and discussion, a improved discussion of landforms, incorporation of a statistical analysis of previous years temperatures and major improvements of many figures. However, some issues remained:

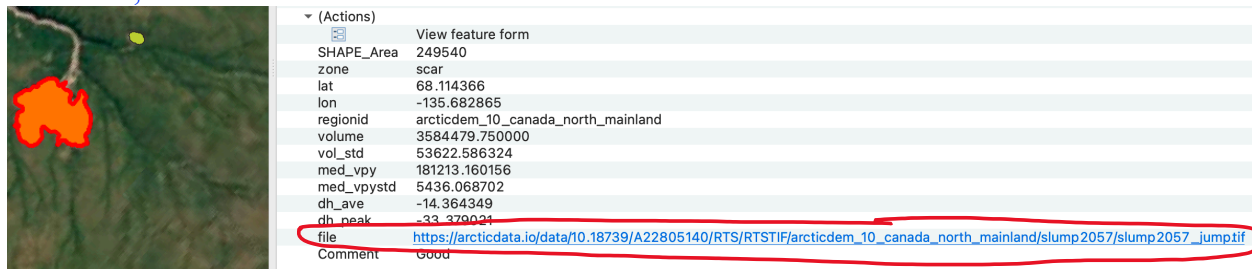
1. Download of the shape files: It is now possible that the individual RTS can be loaded into a GIS, but they are simply numbered, so you have to download each one individually to find the one for a certain region. Maybe the authors could divide it into subfolders with the individual regions. In addition, we couldn't load the shapefiles with the coordinates for all slides. Could possibly be due to the fact that e.g. the PRJ file could not be downloaded (opens only in a

separate table).

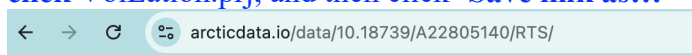
Response:

We now group individual RTS based on regions.

We also add a README.txt file, and add the ftp link for each RTS in ArcticRTS.shp for easy download, as shown in the screenshot below:



To download VolLatlon.prj file from <https://arcticdata.io/data/10.18739/A22805140/RTS/> , **right click** VolLatlon.prj, and then click ‘Save link as...’



Index of /data/10.18739/A22805140/RTS

Name	Last modified	Size	Description
Parent Directory	-	-	-
ArcticRTS.cpg	2025-01-03 21:32	5	
ArcticRTS.dbf	2025-06-04 15:32	797K	
ArcticRTS.prj	2025-06-04 15:31	145	
ArcticRTS.qmd	2025-01-03 21:32	678	
ArcticRTS.shp	2025-01-03 21:32	3.5M	
ArcticRTS.shx	2025-01-03 21:32	22K	
README.txt	2025-06-04 15:32	2.8K	
RTSTIF/	2025-06-04 20:41	-	
VolLatlon.cpg	2025-01-03 21:33	5	
VolLatlon	Open Link in New Tab		
VolLatlon	Open Link in New Window		
VolLatlon	Open Link in Incognito Window		
VolLatlon	Save Link As...		
VolLatlon	Copy Link Address		
VolLatlon	Copy		

2. The discussion is lengthy, and readability would be improved if divided into subsections with own headings, e.g. “Limitations and uncertainties”, “Volumetric yield estimates”, “Relation to climate trends”.

Response: Subsection headings have now been added.

3. Concerning climate relationships, avoid the term “climatic trend” for a 10-year time series. Determining real trends in climate demands at least 30 years, and we are still critical to the significance of figure 5b, even if the figure has improved. It is also normal to give R2-values (not only simple r-values). R2 would give below 25% of explanation, which is not much. It does not mean that climate is not influencing, but you cannot show it statistically significantly on a certain

level in your study.

Response: We change 'climatic trend' to 'climate variability'.

We agree that the significance of Figure 5b is limited, and we now delete it and the related discussions.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: Thank you for your constructive suggestions.

Reviewer #4 (Remarks to the Author):

The main issues raised by my review were correctly addressed in the new manuscript version. This can thus be published.

Response: We appreciate your recommendation.

Reviewer #5 (Remarks to the Author):

My comments focus on machine learning section only. Overall, the methodology presented in the manuscript is rigorous, reliable, and represents a well-justified technical approach.

1. In this manuscript, deep learning primarily serves as a tool to reduce manual workload, utilizing existing adapted methods. The main novelty lies in integrating change detection and classification into a unified framework, with prior knowledge incorporated via three selected feature bands.

Response: Thank you for your positive feedback.

2. For machine learning-based land surface information extraction, training sample distribution is critical. As noted in the manuscript, performance varies across land cover types, indicating limited spatial transferability of the method. Thus, training samples should comprehensively located in the study area.

Response: Thanks for pointing out the importance of training sample distribution. As shown in our Figure S7, our training data distribution indeed covers various regions in the Arctic. We now add your comments to the Methods section: "The performance differences across land cover types indicate limited spatial transferability of our method."

3. The two-stage iterative training aligns with active learning principles, which is meaningful. While model ensemble yields marginal improvement, the method (e.g., voting or other techniques) should be clarified. What is the size of sample blocks?

Response: Thank you for the comment. We now clarify the ensemble method by adding: "The ensemble model is retrieved by applying a logical AND operator to all individual model outputs."

We also add description about the size of sample blocks in the previous section (Step 3a) when we describe data preparation: “The image sample size ranges from approximately 10,000 m² to 500,000 m².”

4. Using 10% of samples for accuracy validation may slightly overestimate performance compared to ground truth, but this does not undermine the conclusions.

Response: Thank you. We will consider exploring different validation sample size percentages in future work.

5. The core identification relies on predefined rules (e.g., elevation change, area thresholds). Machine learning here functions as a labor-saving tool rather than a solution driver, so high methodological novelty is unnecessary.

Response: We agree with your comments.

Reviewer #5 (Remarks on code availability):

The authors provided a README document and a tutorial for reproducibility. The workflow primarily relies on cited third-party tools, with minimal custom-developed components. Consequently, I did not conduct in-depth testing or evaluation of the code.

Response: We appreciate your review of our code and your comments.

Comments to

Pan-Arctic Volumetric Quantifications and Dynamics of Large Permafrost-thaw Driven Mass Wasting Events

By Chunli Dai et al.

General comment

I am not at all a specialist of neither remote sensing nor deep learning methods. Therefore my comments will concentrate on general considerations and formal issues.

The paper offers an overview and pan arctic quantification of volume loss and soil organic carbon loss by retrogressive thaw slumps obtained through deep learning on a comprehensive DEM dataset.

Compared to the already very rich literature on the subject, partly produced by the author team members themselves, the novelty of the contribution can be summarized as follows:

- First evaluation of the phenomenon on a pan-arctic scale (previous works were done on regional levels)
- First quantification of soil loss volumes on a decadal and annual time scales (previous works focussed on areas)
- First estimation of soil organic carbon released by these slumps
- The systematic use of the ArcticDEM data set.

The methodological approach seems to be robust. It uses a high resolution dataset, robust and recognized co-registration and extraction algorithms, several field evidence datasets for calibration and validation. The discussion on uncertainties, error sources and limitations is honestly done (text and mainly supplementary information) and gives a realistic idea of the reliability of the results.

The methodological developments (online methods) shows however that the approach needs a whole set of adjustments and model runs, and still gives approximative results. The results still need a huge work of visual/manual correction, which was done by the team. From this point of view, the paper should possibly be regarded rather as a methodological development paper than for its quantitative results.

Some specific comments.

Title

The title doesn't really reflect the content, and is written in too general terms. I suggested a more precise wording, like:

Pan-arctic quantification of total decadal and annual volume loss and soil organic carbon release by large retrogressive thaw-slumps during the decade 2012-2022.

In fact, the work doesn't quantify the total mass loss, but only the retreat during the considered decade. And it is explicitly restricted to retrogressive thaw slumps, excluding other mass wasting processes. This should be reflected in the title.

Area estimates – lines 161-164

Significance of percentages given in brackets is unclear: ... *in areas with high ground ice content (1.37 ha 1000 km⁻²; Fig. S10; > 20 %)*, ... etc.

20 % of what ? ice-content ? of the data set ? ... ?

Discussion – RTS Activity in response to Climate Trends – lines 252-280 + fig 4 and caption + fig 5 caption + fig S13

Text gives thresholds at 0.5, 0.3, -0.3, -0.5 for both temperature and precipitation. Same in figure 4 caption, and figure 5 caption

But in figure legends of figure 4, figure 5 and figure S13, the thresholds are given at 0.6, 0.2, -0.2, -0.6.

Correct either the figures or the text and caption, according to the thresholds used for the pie charts.

Soil organic carbon release

The discussion on possible much lower SOC release than previously modelled is interesting. However, after reading the methodological part, and regarding the large limitations and uncertainties, it seems that your quantification rather underestimates the total volumes (missing landforms, limited considered depth, ...) and should therefore be considered as a minimum (but I may have misunderstood the final result). May be this should be made more clear and less affirmative ?

Sources of error and analysis limitations

This sub-title is a bit peculiar at this place in the manuscript ! The following text is neither a conclusion nor a real uncertainty analysis.

This part should rather be placed right after the results, and before the discussion. It should best integrate the whole text S1, which is important for understanding.

Figures

Most figures are OK and informative.

Figure 4

The figure 4a is small and contains too much information. Remove single dots and leave only the pie diagrams ?

Figure 5

Figure 5b: is the linear fit really significant ? Visually there is no fit at all ...

Supplementary material

Good and useful for understanding.

The part on “Sources of error ...” should be more explicitly referred to in the text, since it is essential for understanding the limitations of the study.

References and reference list

There is insufficient check of citations and reference list:

- Many references cited in the text are missing in the reference list (highlighted red in the edited PDF).
- For several references there is a wrong date, or an ambiguous date (e.g. missing precision for 2022a, 2022b) (year highlighted red in PDF).
- Many references in the list are not cited. They have to be removed from the list (highlighted in red in the list).
- Formatting of references in the list: many journal titles are not in italic (highlighted green in the PDF).
- Some publications years do not correspond between text and reference list (year highlighted purple in PDF).
- Some references near the end of the list are not in alphabetical order (highlighted with purple line number in PDF).

Reviewer:

Philippe Schoeneich
Université Grenoble Alpes / France