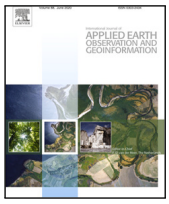
International Journal of Applied Earth Observation and Geoinformation 137 (2025) 104419

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



# Detecting mass wasting of Retrogressive Thaw Slumps in spaceborne elevation models using deep learning

Kathrin Maier [a,*], Philipp Bernhard [b], Sophia Ly [c], Michele Volpi [c], Ingmar Nitze [d], Shiyi Li [a], Irena Hajnsek [a,e]

[a] Department of Environmental Engineering, ETH Zurich, 8093 Zurich, Switzerland
[b] Gamma Remote Sensing, 3073 Guemligen, Switzerland
[c] Swiss Data Science Center, ETH Zurich and EPFL, 8050 Zurich, Switzerland
[d] Permafrost Research Section, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, 14473 Potsdam, Germany
[e] Microwaves and Radar Institute, German Aerospace Centre (DLR) e.V., 82234 Wessling, Germany

## ARTICLE INFO

Dataset link: https://doi.org/10.3929/ethz-b-000718475, https://github.com/kathrinmaier/dem-rts-segmentation

*Keywords:*
Retrogressive Thaw Slumps
Mass wasting
InSAR
Digital elevation model
TanDEM-X
Deep learning

## ABSTRACT

Climate change has led to stronger warming in the Arctic, causing higher ground temperatures and extensive permafrost thaw. Retrogressive Thaw Slumps (RTSs) represent one of the most rapid and considerable geomorphological changes in permafrost regions, occurring when ice-rich permafrost is exposed and thaws. However, large-scale quantification of RTS-related mass wasting in Arctic permafrost landscapes is currently lacking, despite its importance to understand impacts on local environments and the global permafrost carbon cycle. Generating differential digital elevation models (dDEMs) from TanDEM-X single-pass Interferometric SAR (InSAR) observations enables us to quantify volume changes induced by rapid permafrost thaw. To extend this capability across the entire Arctic permafrost region, automation in data processing and RTS detection is essential. This study introduces a method that employs deep learning on InSAR-derived dDEMs to map RTSs and quantify volume changes from RTS activity. We chose eleven study sites with a total area of 71 400 $km^2$ to reflect the diverse character of Arctic environments for model training, testing, and inference. Our trained UNet++ model delivers a scalable solution for mapping RTSs and quantifying mass wasting towards a pan-Arctic scale, achieving segmentation accuracies of 0.58 (Intersection over Union) and classification accuracies of 0.75 (F1) on previously unseen test sites, with volume change estimates from model predictions being within ± 20% of the actual values. We found a total of almost 5000 RTSs active between 2010 and 2021 with volume change rates between 40.75 $m^3yr^{-1}km^2$ for sites in the Siberian to 1164.11 $m^3yr^{-1}km^2$ in the Canadian Arctic.

## 1. Introduction

Climate change has accelerated Arctic warming, causing widespread permafrost thaw (Biskaborn et al., 2019; Olefeldt et al., 2016). Thermokarst development stems from disturbances such as high summer temperatures, heavy precipitation, hydrological changes, tundra fires, and human activities (Grosse et al., 2011). Retrogressive Thaw Slumps (RTSs) are one extreme form of hill slope thermokarst, arising from slope failure following the thawing of ice-rich permafrost (Kokelj et al., 2017). In summer, RTSs extend upslope as massive ice on the headwall thaws, moving the thawed material downslope and creating complex and heterogeneous landforms (Burn and Lewkowicz, 1990). RTSs are prevalent throughout the Arctic, with sizes varying from 0.001 to 1 $km^2$ (Nesterova et al., 2024). Their headwalls can retreat by up to tens of metres annually (Kokelj et al., 2015; Lacelle et al., 2015; Ramage et al., 2018). RTSs often cluster regionally, affecting ecosystems and hydrology by altering ground thermal regimes, affecting sediment and geochemical fluxes, and influencing landscape dynamics and carbon cycles (Turetsky et al., 2020). The rise in ground temperature in recent decades has intensified RTS activities, as shown by the increase in the number and size of RTSs, faster retreat rates, and the increase in mass wasting (Schuur et al., 2015; Lewkowicz and Way, 2019; Ward Jones et al., 2019; Segal et al., 2016; Runge et al., 2022; Bernhard et al., 2022a).

Large-scale RTS mapping and mass wasting quantification are crucial to understanding the impacts of RTSs on ecosystems and the carbon cycle. Advances in satellite remote sensing facilitate large-scale monitoring of RTS dynamics (Lantz and Kokelj, 2008). Previous research has primarily focused on optical and infrared satellite imagery to (semi-)automatically detect RTSs and track their planimetric changes over

* Corresponding author.
*E-mail address:* maierk@ethz.ch (K. Maier).

https://doi.org/10.1016/j.jag.2025.104419
Received 25 November 2024; Received in revised form 29 January 2025; Accepted 10 February 2025
Available online 22 February 2025

time, using indicators such as vegetation loss and soil disturbance in the scar zone (Yang et al., 2023; Nitze et al., 2021; Huang et al., 2022; Xia et al., 2022, 2024; Lin and Knudby, 2023; Runge et al., 2022). RTSs typically exhibit headwalls and scar zones characterised by bare ground, resulting from the removal of previously intact vegetation. However, the complex and heterogeneous nature of RTSs in different Arctic landscapes can make this distinction uncertain, and it may not always directly indicate the actively eroding area of an RTS. Therefore, optical image analysis alone is not sufficient to identify the active erosion area and, more crucially, the associated erosion volume given the three-dimensional nature of RTS processes. In contrast, time-series analysis of Digital Elevation Models (DEMs) reveals topographical changes, allowing a direct estimate of the area affected by RTS activity, as well as quantification of material erosion volumes (Bernhard et al., 2020; Van Der Sluijs et al., 2023; Kokelj et al., 2021). To measure this volume, differential DEMs (dDEMs) can be created by subtracting two DEMs obtained at different observation times. However, DEMs with suitable spatial and temporal resolution that cover the entire Arctic are scarce. One notable example is the ArcticDEM product, which features DEMs generated by optical stereo photogrammetry (Porter et al., 2018). Although ArcticDEM delivers high spatial resolution, its temporal coverage and accuracy vary, posing challenges for RTS analysis (Dai et al., 2024; Huang et al., 2023). Optical data collection is generally challenged by cloud cover and low sunlight in high latitudes, reducing image quality and availability. An alternative method to generate suitable DEMs employs Interferometric Synthetic Aperture Radar (InSAR) from the German TanDEM-X satellite mission. This approach delivers DEMs with a spatial resolution of approximately 10 m and a vertical accuracy of 2–3 m in flat terrain (Bojarski et al., 2021; Krieger et al., 2007). The effectiveness of single-pass InSAR in RTS mapping and volume change quantification has been proven with TanDEM-X DEMs at several Arctic study sites (Bernhard et al., 2020, 2022a). However, to the authors' knowledge, no quantitative study on the amount of mobilised material from RTS activity has been carried out for large spatial scales. This task requires reliable and automated RTS detection methods for dDEMs. Traditional machine learning encounters difficulties (Bernhard et al., 2020), but advances in deep learning have substantially transformed computer vision tasks in remote sensing. Transfer learning utilises pre-trained deep learning models, thus requiring less training data and providing an improved apprehension of spatial context (Zhu et al., 2017). Semantic segmentation with convolutional neural networks effectively detects RTSs in high-resolution optical satellite images (<5 m). The best performing models achieved mean IoU scores from 0.58 with UNet++ (Nitze et al., 2021) to 0.71 with UNet3+ (Yang et al., 2023), and F1 scores from 0.8 with UNet (Witharana et al., 2022) to 0.85 with DeepLabV3+ (Huang et al., 2022) in the Arctic and the Beiluhe region in China. One study used the ArcticDEM for semi-automated RTS detection in the Arctic using an object detection model combined with image segmentation techniques and manual filtering and validation of the final predictions (Huang et al., 2023).

In this study, we used DEMs that were generated from bistatic InSAR observations for Arctic RTS mapping. The objective is to evaluate the potential of automated semantic segmentation with established deep learning models for DEM-based detection of RTS activity and quantification of material erosion volumes that allow for upscaling to large permafrost regions. Specifically, we:

1. manually annotated RTS instances on dDEMs across multiple Arctic study sites and objectively assessed labelling accuracy among different domain experts to establish a comparable baseline;
2. trained and tested several deep learning models on dDEMs for the binary semantic segmentation of RTSs;
3. evaluated the accuracy of active area and volume change rate predictions to ensure high-quality, large-scale monitoring of RTS mass wasting.
4. predicted RTS activity including quantification of volume change for eleven study sites with a total area of 71 400 $km^2$ between 2010 and 2021 based on TanDEM-X derived dDEMs.

## 2. Materials and methods

We combined DEMs and deep learning to map RTS mass wasting activity in Arctic permafrost. First, we outline the selected study sites and the availability of TanDEM-X observations. We then describe the InSAR workflow for generating dDEMs and the elevation error estimation. We address label creation and annotation uncertainties, detail model selection, input data, model training, evaluation metrics, and post-processing. Finally, we introduce the estimation of the active erosion area and volume change from the model predictions as well as the uncertainty quantification. The workflow for this study can be found in Fig. 1.

### 2.1. Study sites and data availability

We selected 11 study sites encompassing a total area of 71 400 $km^2$ spread across the Arctic permafrost region (Fig. 2). The study focuses on Arctic locations with rapid permafrost thaw, representing diverse landscapes, climates, and ground ice content. RTS activity has previously been studied on dDEM data at all selected sites, therefore DEM-based RTS labels partly exist and can be reused for this study (Bernhard et al., 2022a,b). North American sites including parts of the Peel Plateau (Peel) (Kokelj et al., 2015; Segal et al., 2016), the Mackenzie River Delta (Tuktoyaktuk) (Kokelj et al., 2009), Noatak (Nitze et al., 2021), Banks (Lewkowicz and Way, 2019) and Ellesmere Island (Ward Jones et al., 2019) are known to be hotspots for RTS activity with large hillslope or coastal RTSs and high mass wasting activity. Siberian study sites including parts of Gydan, Yamal, and Taymyr (north (N), south-west (SW), south-east (SE)) and Chukotka peninsulas experienced less RTS activity over the past decade, with mainly small and shallow lakeshore RTSs (Nitze et al., 2018; Nesterova et al., 2020). However, during a heatwave in the summer of 2020, many new RTSs emerged in Siberia, especially in N Taymyr (Bernhard et al., 2022b). All sites are part of the continuous permafrost zone (Obu et al., 2019) within the Arctic tundra or the boreal-tundra transition zone with a ground ice content greater than 10% (Brown et al., 2002). Peel, Banks, Ellesmere, Chukotka, and N Taymyr have rugged terrain; the others are relatively flat with abundant lakes. TanDEM-X covers all regions within three observation times. We generated dDEMs for 2010 to 2016 (ascending orbit) and 2010 to 2020/2021 (ascending/descending orbits).

### 2.2. InSAR processing and DEM generation

Using pairs of SAR observations from the TanDEM-X mission, we generated approximately 1400 DEMs that span the years 2010 to 2021 at the selected study sites (Figs. 2, 1a). We orthorectified the SAR observations based on the TanDEM-X 12 m DEM product. Using Gamma Remote Sensing software (Werner et al., 2000), we generated DEMs using standard bistatic InSAR processing (Fritz et al., 2011). Furthermore, we adopted the permafrost-specific processing steps outlined in Bernhard et al. (2020). Each SAR observation results in a DEM, a coherence map, a layover/shadow map, and an incidence angle map.

All DEM products were reprojected to a common coordinate system, WGS 84/NSIDC Sea Ice Polar Stereographic North (EPSG: 3413), and resampled to 10 m spatial resolution (Fig. 1b). To improve data handling and storage, we applied a tiling scheme to all DEM products with 10 × 10 km patches with small spatial overlap to eliminate edge effects. For several DEMs that overlap in a given year, we calculated a weighted average per pixel based on coherence values. We assigned the year corresponding to the start of winter to the averaged DEMs and subtracted the earlier DEM from the later DEM. For all DEM calculations, we performed a coregistration with the Python library

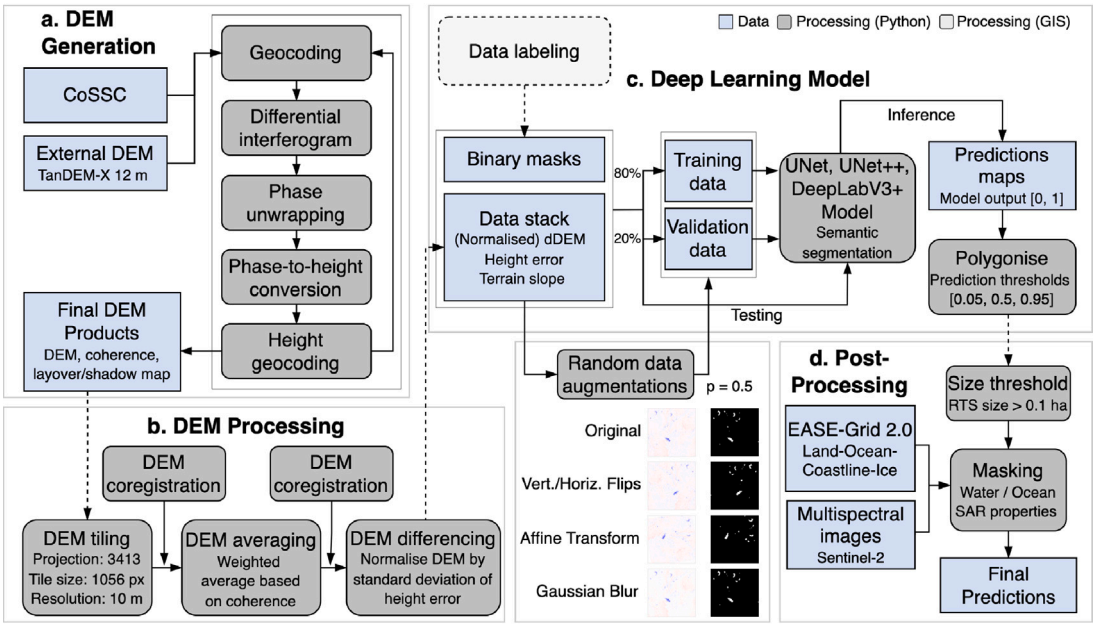



**Fig. 1.** Workflow: a. DEM generation through bistatic InSAR processing. b. DEM processing included resampling, reprojection, tiling, averaging, differencing, and normalisation. c. Manual RTS annotations on normalised dDEMs are used to train deep learning models on a four-channel data stack (normalised dDEM, dDEM, elevation error, and terrain slope). d. Postprocessing includes size thresholding and masking (low SAR quality areas and water bodies). While all parts (a–d) are implemented automatically, no fully automatic pipeline currently integrates them.

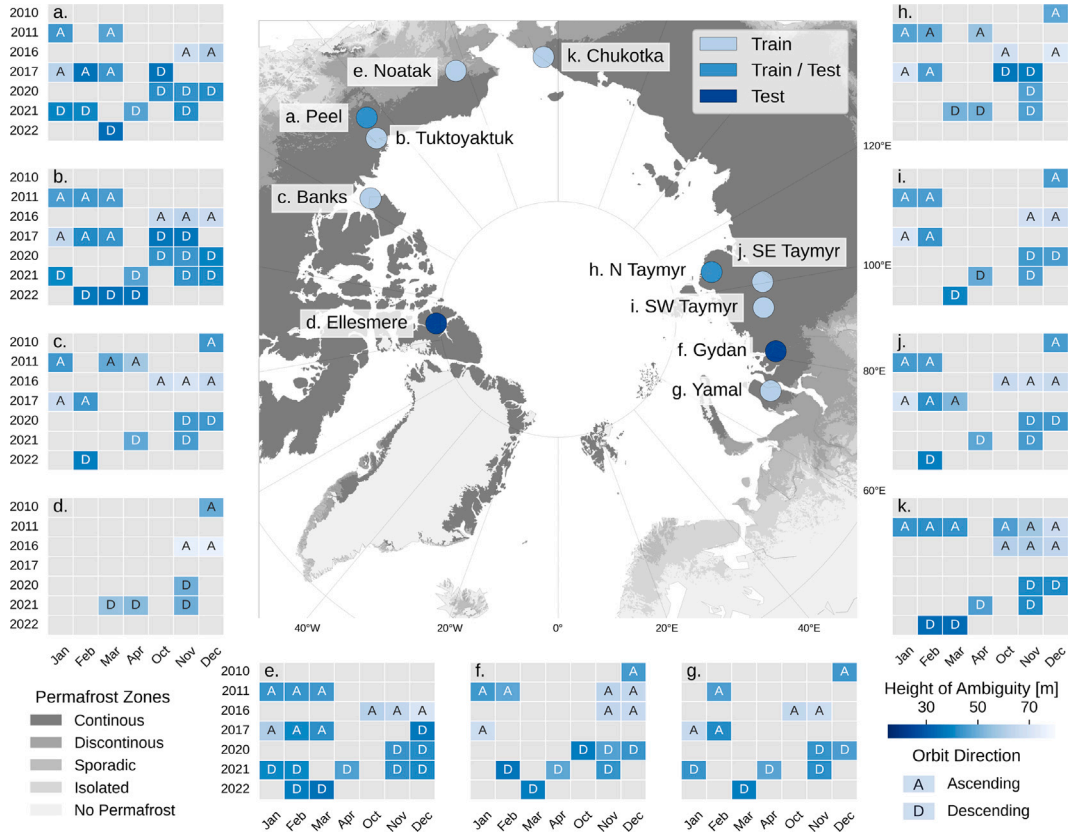



**Fig. 2.** Arctic study sites for RTS detection and quantification of volume change: Seven study sites were solely used for model training and hyperparameter selection (training and validation) and two solely for model testing, on previously unseen data. Two sites (Peel and N Taymyr) were split into geographically separate regions to accommodate both model training and testing. DEMs are produced from TanDEM-X observations from 2010 to 2021 during winter months and with Height of Ambiguity (HoA) between 15 and 80 m. Observations until 2016 were acquired in ascending (A), later observations in descending orbit direction (D).

AROSICS (Scheffler et al., 2017) and removed tilts with the xDEM library (Hugonnet et al., 2021). Even small shifts can greatly affect the quality of the dDEMs. Subpixel shifts occurred in same-orbit DEMs, while different-orbit DEMs required corrections of up to 20 pixels.

### 2.3. Uncertainty estimation of InSAR-derived DEMs

For the uncertainty estimation of InSAR-derived DEMs, we are following a systematic approach that addresses three key aspects: careful selection of observations based on temporal and observation property constraints, masking of unreliable regions affected by radar imaging artefacts, and normalisation of elevation differences with respect to estimated errors.

InSAR processing enables the quantification of the vertical accuracy for each pixel of the generated DEM. The estimated elevation values are subject to both random and systematic errors. To mitigate the influence of systematic errors and accuracy constraints, we only selected TanDEM-X observations that met specific criteria. In general, observations with a large HoA result in DEMs with lower height sensitivity. An upper limit of 80 m has been shown to be reasonable for RTS monitoring (Bernhard et al., 2020; Martone et al., 2012). Additionally, we set a lower limit of 15 m, as small HoA values can hinder phase unwrapping. Most observations used in this study have HoA values between 30 and 70 m (Fig. 2). Wet snow during the melting season can cause noticeable elevation changes and decrease coherence, thus reducing the accuracy of elevation measurements (Nagler and Rott, 2000). During the summer months, the growth of vegetation can cause significant errors due to volume decorrelation (Zwieback et al., 2018). Therefore, we only used observations from outside the snow melt and summer seasons (October to April). Given the low average winter temperatures at all study sites, we anticipate a dry snowpack and a full propagation of radar waves to the ground during the winter season (Leinss and Bernhard, 2021; Millan et al., 2015; Bernhard et al., 2020). Furthermore, we created a SAR quality mask that defines areas affected by layover/shadow, low coherence (<0.3) and high local incidence angles (>1 rad). Elevation estimates in these areas are not reliable and should therefore not be considered for volume change estimation.

Random errors arise primarily from volume decorrelation and low backscatter intensities in the SAR acquisitions. These errors can be estimated by converting the coherence $\gamma$, a measure of interferometric phase quality, into an estimate of the elevation error $\sigma_h$ for each produced DEM

$$\sigma_h = \frac{\sqrt{1-\gamma^2}}{\gamma\sqrt{2L}} \frac{\mathrm{HoA}}{2\pi} \tag{1}$$

where $L$ is the number of looks used ($L = 4$) for the generation of the interferogram to reduce speckle noise and HoA is the Height of Ambiguity, a measure of the height sensitivity (Krieger et al., 2007; Rosen et al., 2000; Rodriguez and Martin, 1992).

TanDEM-X generated DEMs contain inherent noise, and RTS headwall heights can be close to the expected standard error of the elevation change. This can make it difficult to distinguish RTS-induced elevation changes from background noise. We perform a normalisation of the elevation difference ($\mathrm{dDEM_{norm}}$) by the estimated elevation error

$$\mathrm{dDEM_{norm}} = \frac{h_1 - h_2}{\sqrt{(\sigma_{h_1}^{\gamma_1})^2 + (\sigma_{h_2}^{\gamma_2})^2}} \tag{2}$$

where $h_1$ and $h_2$ are the measured elevations of the later and the earlier DEM, respectively. Significant elevation changes are more discernible, while regions with high measurement errors and potentially erroneous elevation changes are suppressed (Bernhard et al., 2020).

### 2.4. DEM-based RTS labelling

Human error and subjectivity typically play a role in the label creation process. Annotations not only convey information on interpretation, but drive the quality of training and generalisation of the segmentation model (Nitze et al., 2024). In addition to satisfying the need for generating enough labels for model training, we tried to quantify the labelling subjectivity through an assessment of independent annotations on the same data from three domain experts. RTS features were manually labelled in normalised dDEMs because significant elevation changes can be better separated from background noise. A set of predefined guidelines helped to standardise the annotation process, including instructions for data visualisation such as common colour scales and data ranges, when to use optical imagery to validate RTS activity, and how to digitise the shapes (Supplement S1). By selecting an area of 100 $\mathrm{km^2}$ in each of the study sites used for model testing, we assessed how well the experts (E1, E2, E3) agree on their annotations. For model training, validation, and testing, the expert E2 digitised a total of 3614 RTSs (training and validation: 1743, testing: 1871). Where existing labels were present (Bernhard et al., 2022a), E2 reviewed and adjusted them as necessary.

### 2.5. Deep learning for DEM-based RTS detection

We assessed deep learning models commonly used for semantic segmentation tasks in remote sensing applications. Here, we present our approach to the preparation of input data, model training and evaluation, as well as the implemented post-processing strategy (Fig. 1c). We generally distinguish between the terms model training, validation, and testing. Training a model involves fitting the model to a training dataset by adjusting its parameters, aiming to minimise the difference between prediction and reference. Validation is part of the training stage and refers to the process of evaluating the model's performance on a separate dataset that was not used during training, allowing, for example, for hyperparameter tuning. In contrast, model testing involves assessing the model's performance on a completely unseen dataset to provide an unbiased evaluation of the model's generalisation ability.

#### 2.5.1. Data

Following the manual creation of RTS annotations, we assigned the study sites to the training phase (which includes validation) or to the testing phase. We used two study sites, Peel and N Taymyr, for both training and testing. Hence, we divided them into spatially distinct sections for each phase. The DEM data were organised into 10×10 km tiles (Fig. 1b). For the sites used in model training, we manually separated the data into training and validation tiles, ensuring that there was no spatial overlap to prevent accuracy overestimation due to data autocorrelation. 80% of the entire dataset used in the training stage was assigned to training and 20% to validation. We converted the RTS polygons into binary segmentation masks and trained deep learning models using four input channels: 1. normalised dDEM, 2. dDEM, 3. elevation error estimate, and 4. terrain slope (Horn, 1981) (Fig. 3). Typically, deep learning models for semantic segmentation are designed for optical images with three input channels. Initially, we ran models with a single input (3×$\mathrm{dDEM_{norm}}$) and various combinations of DEM data with three input channels. However, we found that including all available data as separate channels improved model performance. We used tiles of 512 × 512 pixels and normalised the data to a similar value range. Details about the study sites and reference labels can be found in Supplement S2 and data selection and normalisation parameters in S3.

#### 2.5.2. Model training

We trained three deep learning models and performed hyperparameter optimisation. We limited our selection to established Convolutional Neural Network architectures, namely UNet, UNet++, and DeepLabV3+, which have shown promising results in RTS detection based on optical satellite imagery (Yang et al., 2023; Nitze et al., 2021; Huang et al., 2022; Xia et al., 2022). Our automated pipeline was implemented with PyTorch (Ansel et al., 2024), TorchGeo (Stewart et al., 2022), and PyTorch Segmentation Models (Iakubovskii, 2019) (Fig. 1c). The models were trained on one GPU (NVIDIA RTX A6000) and limited to 200 epochs with a batch size of 8. We chose Dice loss as the loss function and Intersection over Union (IoU) as the validation

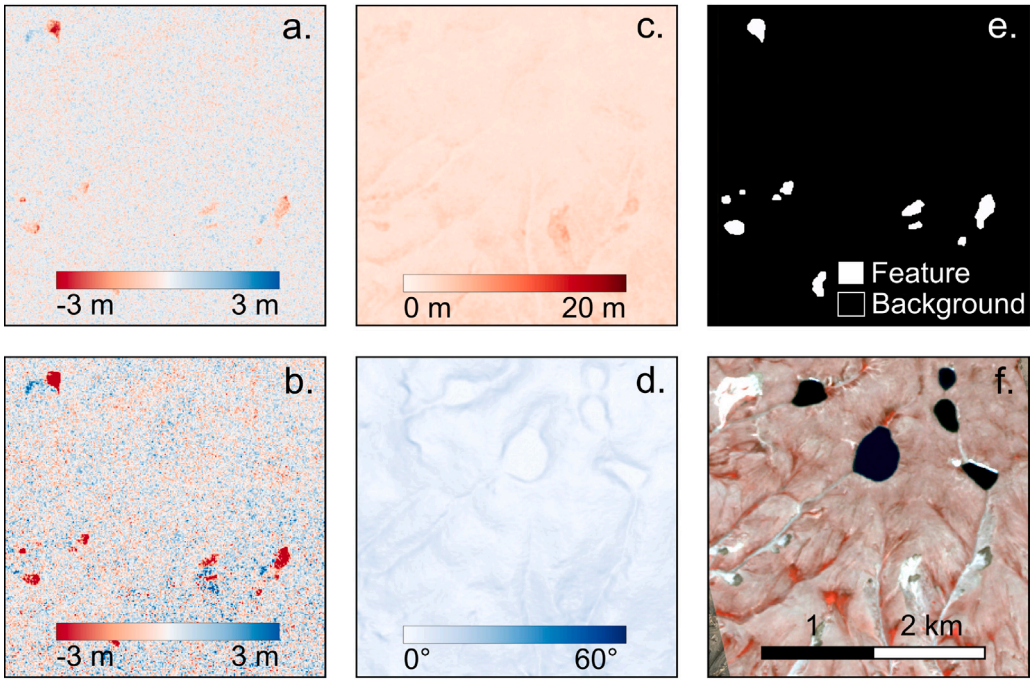



**Fig. 3.** Model input with four channels (a–d): normalised dDEM, dDEM, elevation error estimated from InSAR coherence, and terrain slope. Binary segmentation mask based on manual RTS labels (e) and optical false-colour composite image from Sentinel-2 on 19/07/2017 (f). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

accuracy metric. We optimised the hyperparameters by testing different optimisers, learning rates, and learning rate schedulers (Supplement S4). An early stopping criterion of 50 epochs was applied to avoid overfitting the models. We used the smallest ResNet architecture (ResNet18) as the backbone for all models, as we found no improvement in model performance with larger backbones (ResNet34, ResNet50, ResNet101). Fine-tuning the model with weights from the Imagenet dataset (Deng et al., 2009) led to improvements in accuracy and training time compared to the initialisation of random weights. Freezing the backbone did not improve the validation accuracy in any test, so we kept the backbone trainable.

We trained the segmentation models at nine study sites where seven sites were solely used for the training stage (training and validation), and two sites were split into spatially separate training and testing regions. The models were trained with dDEMs from 2010 to 2016 (TP1) with same-orbit observations (ascending) (Fig. 2). We excluded regions within the study sites without RTS activity to avoid an excessively high proportion of negative samples. Even then, only 0.05% of the total area show RTS activity (Supplement S2). This indicates that the dataset has an extreme class imbalance, which can impair the detection performance.

Data augmentation is a widely employed technique to artificially expand the volume of annotated data during model training and is particularly effective in mitigating data scarcity, a critical challenge in the context of data-intensive deep learning paradigms. For images, it involves transforming, synthesising, or degrading existing data to artificially increase the size of the dataset. Its purpose is to create diverse mini-batch samples that help the model improve generalisation. We chose four commonly used augmentation techniques: vertical and horizontal flips, affine transformation, and Gaussian blur (kernel size: 3 pixels). These operations were applied on-the-fly with a probability of 0.5 that the data is augmented and a probability of 0.5 that each augmentation is applied sequentially (Fig. 1).

### 2.5.3. Model evaluation

We tested the models' transferability to unseen regions and dDEMs from longer time periods at four study sites. Two study sites, Ellesmere and Gydan, were used to assess geographic transferability to regions completely separate from the training data. Ellesmere, in the high Canadian Arctic, has minimal vegetation, coastal erosion, and many large RTSs. Gydan, a Siberian tundra site, has flat terrain, abundant lakes, and fewer small RTSs. We separated parts of two study sites, Peel and N Taymyr, from the training samples to use as independent, yet similar, test datasets. Peel has a complex topography with steep slopes and rivers. N Taymyr was tested with both DEMs from a descending orbit (2017–2020, TP3) to evaluate the influence of the SAR viewing geometry on model performance. We also tested the models over different time periods to assess their ability to generalise and handle differences in RTS size and elevation change magnitudes.

To evaluate both segmentation and classification performance, we chose a set of pixel-level and detection-based metrics. All pixels of the model output (prediction) are compared to the manually generated labels (reference), and hence categorised as either true positives (TP), false positives (FP), true negatives (TN), or false negatives (FN) for the pixel-level assessment. The classification performance is assessed at the RTS feature level. Connected positively classified pixels of the model predictions are considered as a potential RTS instance. If a predicted RTS instance spatially intersects with the reference RTS label, it is counted as TP since even small intersections help to roughly localise RTSs. Due to the high class imbalance, the pixel accuracy that incorporates TN is not an appropriate metric. Instead, we assessed the models' performance using metrics commonly employed for imbalanced segmentation tasks with

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{3}$$

$$\text{Precision } P = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall } R = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F1} = \frac{2 \times P \times R}{P + R}. \tag{6}$$

The detection IoU is calculated by averaging the pixel IoU of all TP detections.

#### 2.5.4. Postprocessing strategy

Rasters conveying probability values from 0.0 to 1.0 constitute the output of a deep learning model, indicating the likelihood that a pixel belongs to an RTS feature. We transformed the probability rasters into binary prediction maps using thresholds of 0.05, 0.5, and 0.95 to investigate the precision of the trained models. As a second step, the binary prediction rasters are vectorised to polygonal features. Typical of imbalanced segmentation, models tend to overpredict the positive class. To address this, we developed a post-processing routine to reduce the number of FP (Fig. 1d). DEMs derived from InSAR generally exhibit larger elevation errors over water bodies, leading to apparent elevation changes in dDEMs (Section 2.3). Rugged and mountainous areas can also degrade the quality of the generated DEMs due to layover and shadow effects and high local incidence angles in SAR imagery dependent on the acquisition geometry. These regions are likely to be incorrectly classified as RTSs because of their comparable elevation change, dimensions, and shapes. We created water body masks using the temporal average of cloud- and snow-free stacks of Sentinel-2 imagery and a Normalised Difference Water Body Index (NDWI) threshold ranging from 0.0 to 0.2, depending on the landscape characteristics. The water body mask is complemented by the EASE-Grid 2.0 Land-Ocean-Coastline-Ice Masks dataset (Brodzik and NSIDC, 2013) for ocean masking. We excluded any predictions that intersect with the water body or the SAR quality mask by more than 50%. Considering the resolution of the DEM and typically reported RTS sizes, e.g. in Nesterova et al. (2024), we implemented a size threshold of $1000\,m^2$, to exclude small polygons that are unlikely to represent correct RTS occurrences (Huang et al., 2021). Furthermore, we excluded polygons with a minimum elevation change below 2 m according to the typical sensitivity to elevation change of DEMs derived from TanDEM-X observations.

### 2.6. Model inference and volume change estimation

We applied the trained UNet++ model to the TP2 dDEMs (2010–2021) at all study sites to detect and delineate active RTSs. To advance beyond quantifying changes in the RTS area, as is possible in optical satellite image analysis, we estimated the volume of material eroded during TP2 based on the derived dDEMs. Elevation changes around zero indicate stable terrain, while negative and positive values reflect mass loss and gain, respectively. We calculate the volume change $\delta v$ based on the average elevation change for the dDEM pixels belonging to a predicted RTS feature

$$\delta v_{RTS} = \frac{\sum_i^{n_{pix}} h_1(i) - h_2(i)}{n_{pix}} \cdot a_{RTS} \tag{7}$$

where $a_{RTS}$ is the area of the predicted RTS feature and $n_{pix}$ is the number of all pixels belonging to the predicted RTS feature. Furthermore, we quantify the uncertainty of the volume change estimate based on the elevation error $\sigma_h$ with

$$\sigma_{v_{RTS}} = \frac{\sqrt{\sum_i^{n_{pix}} \sigma_h(i)^2}}{n_{pix}}. \tag{8}$$

For study sites where reference labels for TP2 are present (Peel, Ellesmere, Gydan), we performed an accuracy assessment of the estimated volume changes from the RTS features predicted by UNet++. Therefore, we summarise the deviations in the area and volume change for the three study sites. To ensure comparability of the results, we calculated area and volume change density rates by normalising the estimates by time period and size of the study site, following the approach of Kokelj et al. (2017). In addition, we conducted a manual visual quality check to exclude obvious false predictions resulting from artificial elevation changes, such as those caused by infrastructure or mining. This process ensured a high-quality dataset suitable for further analysis.

## 3. Results

We generated DEMs from about 1400 TanDEM-X observations between 2010 and 2021 and manually annotated more than 3500 RTS instances at 11 Arctic study sites. We trained three segmentation models to detect RTS-induced elevation changes in dDEMs and assessed both classification and segmentation performance. Based on the best-performing model, we produced a dataset containing RTS polygons and mass-wasting quantities for all study sites between 2010 and 2021.

### 3.1. Baseline accuracies from manual RTS labelling

In remote sensing-based detection tasks, models are typically evaluated on the basis of their performance of replicating human labelling rather than actual feature recognition. Our assessment of the influence of subjectivity on labelling tasks establishes an upper limit of the achievable accuracy of the model. Although all experts followed common annotation guidelines, the analysis shows notable variances in labelling agreement between study sites and dDEM time periods among experts (Fig. 4). The average classification accuracy is 0.63 IoU and 0.76 F1, while segmentation performance is lower with 0.55 IoU and 0.70 F1. The experts showed a high agreement (pixel IoU > 0.7, detection F1 > 0.8) in Peel and Ellesmere (Fig. 4a–c, g–i). These sites feature large RTSs and long time intervals between DEMs, leading to evident elevation changes and therefore distinguishable characteristics. Fig. 4g shows one common problem with manual labelling: E2 and E3 label one RTS feature, while E1 labels two separate instances. RTSs are complex landscape features that can potentially merge over time, resulting in the formation of intricate and dynamic landforms. The resolution of the imagery plays a crucial role, with a lower resolution complicating the delineation of separate entities. In addition, there is a degree of variability in the labelling precision among the experts. The number of vertices used by the different experts indicates the level of detail provided (Fig. 4c, f, i, l). In Gydan and N Taymyr the expert consensus was lower than in Peel and Ellesmere (Fig. 4d–f, j–l). The dataset for N Taymyr contained DEMs from 2017 to 2020, leading to less pronounced elevation changes (Fig. 4j). This resulted in reduced agreement on the number of identified RTSs and segmentation performance. The disagreement was highest in Gydan with a segmentation IoU of only 0.31. The few existing RTSs caused high variability in the experts' annotations. Despite similar training, some experts labelled only areas with significant elevation loss and visible scar zones in optical images, while others also considered smaller and less pronounced features (Fig. 4d, j).

### 3.2. Deep learning model performance

We allowed all models to train for a maximum of 200 epochs with the option to stop training when the validation accuracy did not improve for 50 epochs to prevent overfitting. Under optimal hyperparameter settings, UNet and UNet++ converged in about 150 epochs, achieving a pixel IoU of 0.79 and a dice loss of 0.22. DeepLabV3+ trained more slowly, plateauing at around 150 epochs with a pixel IoU of 0.74 and a loss of 0.25, not fully converging at the end of training (Supplement S5). We tested the models with and without post-processing (Table 1). The best performing model was UNet++ (0.55 pixel IoU, 0.71 detection F1), followed by UNet (0.54, 0.70) and DeepLabV3+ (0.50, 0.70). We observed that UNet and DeepLabV3+ predicted more positive pixels than UNet++, resulting in slightly lower recall values. However, UNet++'s superior F1 score indicates that it avoided more FP despite missing some TP. Post-processing greatly mitigated FP across all models, improving UNet++'s performance by up to 4%.

In the following, we further analyse the overall best-performing model (Table 2). UNet++ achieved the highest performance for TP1 in Ellesmere and for TP3 in N Taymyr, with DEM combinations of the

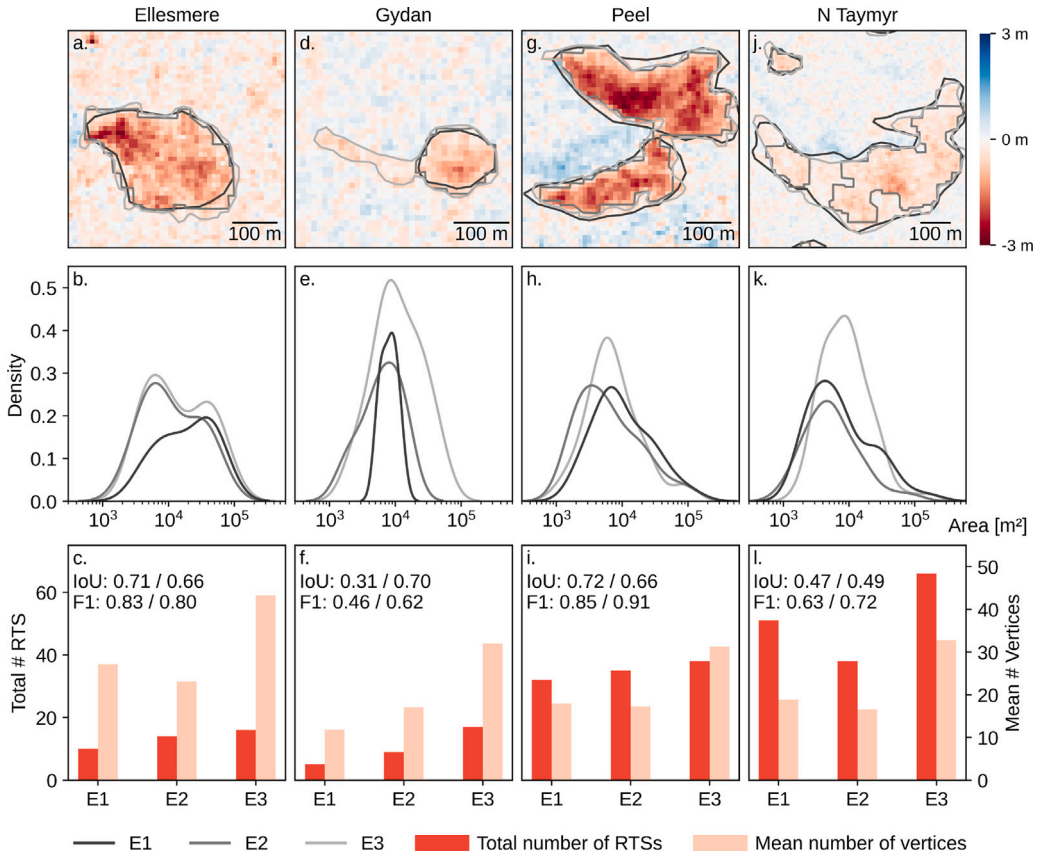



**Fig. 4.** Examples of manual RTS labels on normalised dDEMs, label size distribution, total number of labels, and average number of vertices in Ellesmere (a–c), Gydan (d–f), Peel (g–i), and N Taymyr (j–l). The first value of IoU and F1 (c, f, i, l) represent average segmentation, the second the classification performance between the experts. The experts vary in agreement between the test sites with high agreement in Ellesmere and Peel and low agreement in N Taymyr and Gydan. Patterns in the number of vertices and labels, and the size distribution highlights the impact of different annotation styles.

**Table 1**
Segmentation (pixel) and classification (detection) performance before and after post-processing for UNet, UNet++, and DeepLabV3+. UNet++ performed best, followed by UNet and DeepLabV3+. Post-processing had a positive influence on all metrics and models.

| Model | Post-processing | Pixel | | Detection | |
|---|---|---|---|---|---|
| | | IoU | F1 | IoU | F1 |
| UNet | Before | 0.54 ± 0.19 | 0.68 ± 0.18 | 0.69 ± 0.06 | 0.70 ± 0.14 |
| | After | 0.55 ± 0.18 | 0.69 ± 0.17 | 0.71 ± 0.05 | 0.72 ± 0.12 |
| UNet++ | Before | 0.55 ± 0.16 | 0.70 ± 0.15 | 0.71 ± 0.04 | 0.71 ± 0.13 |
| | After | 0.58 ± 0.16 | 0.72 ± 0.15 | 0.73 ± 0.04 | 0.75 ± 0.12 |
| DeepLab V3+ | Before | 0.50 ± 0.14 | 0.66 ± 0.13 | 0.57 ± 0.09 | 0.70 ± 0.10 |
| | After | 0.52 ± 0.14 | 0.67 ± 0.12 | 0.62 ± 0.06 | 0.74 ± 0.09 |

**Table 2**
Performance statistics of best-performing model (UNet++, post-processed) for all test sites and time ranges.

| Test site | Time period | Pixel | | | | Detection | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | IoU | F1 | P | R | IoU | F1 | P | R |
| Ellesmere | TP1 | 0.75 | 0.86 | 0.90 | 0.82 | 0.80 | 0.83 | 0.92 | 0.76 |
| | TP2 | 0.69 | 0.82 | 0.83 | 0.80 | 0.73 | 0.83 | 0.87 | 0.79 |
| Gydan | TP1 | 0.26 | 0.41 | 0.72 | 0.28 | 0.67 | 0.52 | 0.72 | 0.41 |
| | TP2 | 0.42 | 0.59 | 0.69 | 0.51 | 0.68 | 0.68 | 0.68 | 0.68 |
| Peel | TP1 | 0.60 | 0.75 | 0.71 | 0.80 | 0.73 | 0.75 | 0.73 | 0.76 |
| | TP2 | 0.63 | 0.78 | 0.72 | 0.84 | 0.73 | 0.74 | 0.69 | 0.80 |
| North Taymyr | TP3 | 0.68 | 0.81 | 0.81 | 0.81 | 0.75 | 0.91 | 0.92 | 0.91 |

same orbit. Both sites are characterised by abundant large RTSs. In contrast, Gydan displayed the lowest accuracy, yet maintained high precision, effectively predicting many RTS instances but missing a considerable number. Fig. 5a shows that this shortfall decreased when predicting over longer time intervals. In TP2, the pixel IoU and the detection F1 increased by more than 15%, and the recall by more than 20%. Peel's TP1 data helps to understand the model behaviour in regions similar to training areas, the impact of extended time intervals between DEMs and of complex terrain. The metrics are balanced with slight improvements over an extended time interval between DEMs. Fig. 5b and c visualise the good performance of reducing FP during post-processing with SAR quality and water body masks in Peel and Gydan.

Notably, the probability masks from UNet++ exhibit high precision compared to the reference data, with a strong spatial gradient at the edges. With different thresholds for converting probability maps to binary predictions (Fig. 5d), the variation in the prediction mask was minimal. We therefore selected a threshold of 0.5 for all tests and model inference.

### 3.3. Quantification of RTS mass wasting from model predictions

Based on the trained UNet++ model, we predicted RTS activity at all study sites over the last decade (TP2). Based on the polygonised and post-processed predictions and the dDEMs of TP2, we calculated RTS area and volume change rates. We quantified the error in the volume change estimates that arises from the mismatch between prediction and reference labels (Table 3). The volume change rates deviate less than the area, probably because overestimated pixels near the RTS boundaries contribute only with relatively low elevation changes. The predictions in Ellesmere and Gydan slightly underestimate (−3 and −13%, respectively), while the predictions in Peel overestimate the

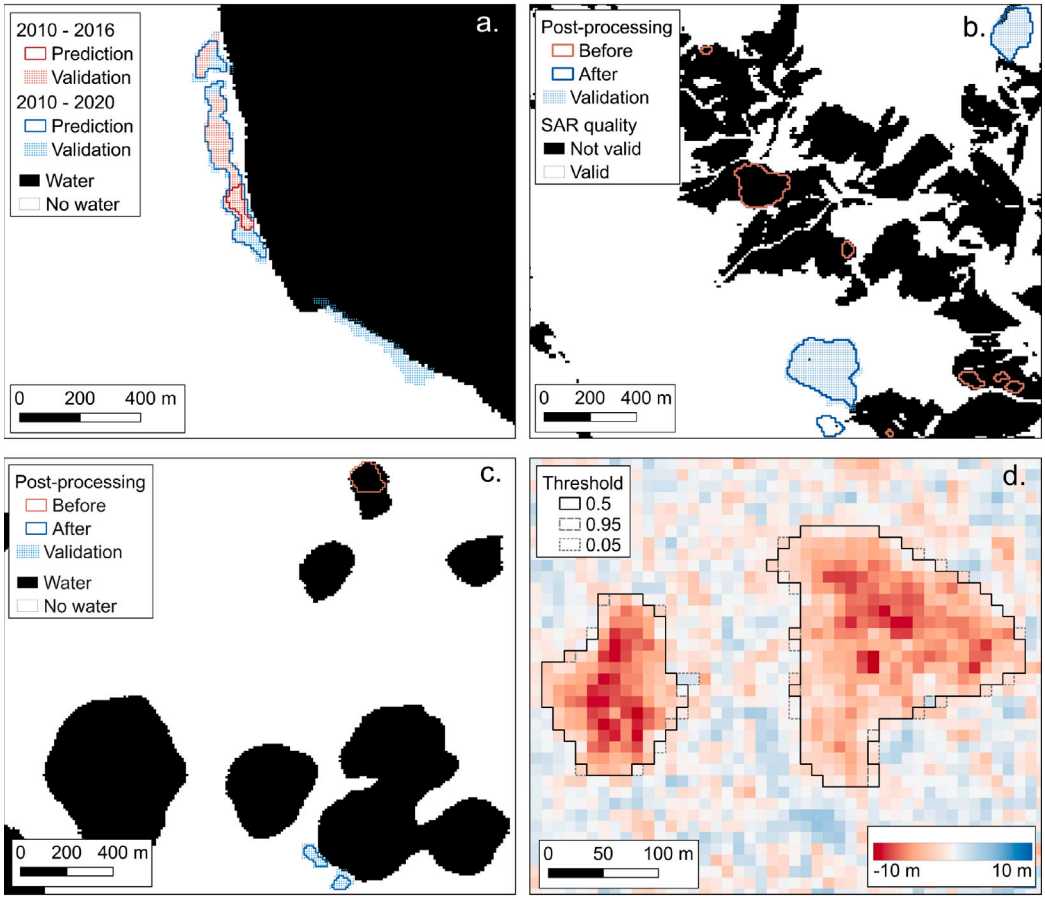



**Fig. 5.** Post-processing, model precision, and limitations of predictions: a. Small and shallow RTSs are difficult to detect (Gydan), b. SAR quality mask reduces FP in areas with high local slope angles (Peel), c. masking water bodies reduces FP detections in water-rich test sites (Gydan), d. high precision in model prediction masks (Ellesmere).

**Table 3**
Total volume change from RTS predictions and reference labels for TP2.

| Test site | Type | Area change rate [$1 \times 10^4$ m$^2$ yr$^{-1}$] | Volume change rate [$1 \times 10^4$ m$^3$ yr$^{-1}$] |
|---|---|---|---|
| Peel | Reference | 46.68 | 300.7 ± 14.9 |
| | Prediction | 57.48 | 343.5 ± 17.6 |
| Ellesmere | Reference | 40.16 | 148.8 ± 13.3 |
| | Prediction | 38.40 | 144.9 ± 12.1 |
| Gydan | Reference | 14.81 | 38.8 ± 5.0 |
| | Prediction | 11.88 | 33.9 ± 4.8 |

volume change (+14%). However, considering the elevation error inherent to the TanDEM-X derived DEMs, these deviations are within the uncertainty bounds of presented method.

Fig. 6 illustrates the RTS mass-wasting characteristics of predicted RTSs for TP2. We manually verified the post-processed predictions and corrected errors from height changes related to human activity such as infrastructure and mining, as well as sediment-rich and small water bodies missing in the water body mask. This manual filtering should not be considered a comprehensive dataset review, but rather a minor quality enhancement for subsequent use, as we removed less than 10% of the total RTS features. N Taymyr in Siberia has the highest number of RTSs, followed by Peel and Banks in the Canadian Arctic (Fig. 6c, d). Despite the high abundance of RTSs, N Taymyr has lower mass-wasting rates compared to Banks and Peel, which show the highest changes in area and volume per unit area (Fig. 6a, b, e, f). In general, Siberia experienced lower RTS activity during the last decade, with smaller and shallower RTSs than North America aligning with the results of previous studies (Bernhard et al., 2022a; Lewkowicz and Way, 2019; Nitze et al., 2018). The area and volume change distributions show no substantial differences between the two regions. Banks, Peel and SW Taymyr exhibit slightly wider ranges, whereas N Taymyr, Chukotka, and Noatak show narrower ranges in the distributions (Fig. 6g, Supplement S6).

## 4. Discussion

We developed a deep learning-based method to accurately map and monitor RTS mass wasting in DEM time-series data over large spatial scales. The best model (UNet++) showed good segmentation (pixel IoU: 0.58 ± 0.16, F1: 0.72 ± 0.15) and classification performance (detection IoU: 0.73 ± 0.04, F1: 0.75 ± 0.12). One weakness of the method is its limited ability to detect small and shallow RTSs, which is probably attributed to the spatial resolution and height sensitivity of TanDEM-X-derived DEMs. However, these features also present challenges for detection models trained with high-resolution optical images with reported accuracies comparable to those of our method (Yang et al., 2023; Nitze et al., 2021). In order to investigate the impact of this limitation, elevation change measurements with higher spatial resolution and vertical accuracy such as DEMs produced by LiDAR or UAV photogrammetry surveys would be needed. However, the same temporal baseline between the DEMs and a reasonable sample size in different geographic settings would be necessary to produce an accurate and robust analysis. Unfortunately, to the authors' knowledge, no study exists that would allow for such a comparison. Generally, we observed that the dDEMs derived from observations taken further apart in time yielded better accuracy and facilitated the detection of RTSs by capturing more significant height changes (Fig. 5 a) that aligned with the findings of Bernhard et al. (2020).

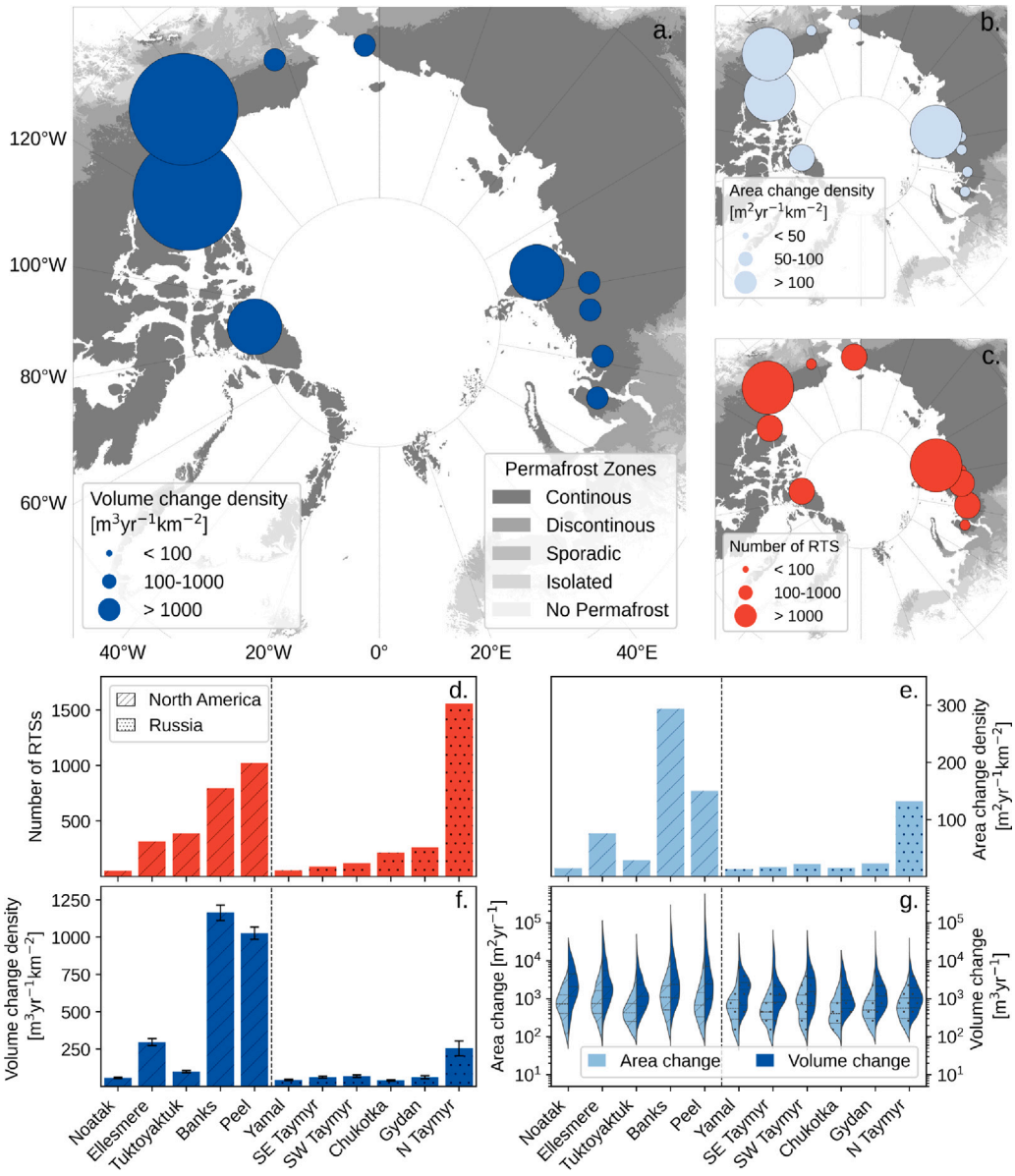



**Fig. 6.** RTS mass wasting during 2010–2021: Number of RTSs (c, d), Area change density (b, e), Volume change density (a, f), distribution of area and volume change (g). Most RTSs were found in N Taymyr, while both area and volume change are highest in Banks and Peel indicating high mass wasting activity from large and deep RTSs.

Due to variations in study designs and data, the comparison of our method with existing RTS detection studies is difficult. Yang et al. (2023) reports a classification IoU of 0.57 with a UNet3+ model trained with high-resolution optical images at distributed Arctic sites and tested at unseen locations in Yamal and Gydan. Their highest IoU of 0.76 occurred when testing on sites analogous to training sites, which is not directly comparable to the conditions of our study. Binary segmentation may excel with simpler data, such as dDEMs, compared to optical data and generalise better. Optical images are strongly influenced by variable Arctic vegetation, while X-Band InSAR dDEMs are mostly unaffected by land cover in low-vegetated tundra. However, SAR observations typically face challenges in complex topography as in Peel. However, the detection performance did not decrease in this study site likely due to effective post-processing (Fig. 5b). Huang et al. (2023) tested DEM-based RTS detection on ArcticDEM data that allows higher spatial resolution but quality issues due to the limitations of stereo optical imagery. The study counted 2494 active RTSs in the entire Arctic between 2008 and 2017, approximately half of those detected in our 71 400 km$^2$ study region. The authors note that they might underestimate RTS occurrence, yet direct comparison is difficult because of differences in observation periods. In particular, Siberia's summer 2020 heat wave triggered the formation of many new RTSs on the Taymyr Peninsula, potentially skewing comparisons (Bernhard et al., 2022b). Through the comparison of expert annotations, we set a better baseline to evaluate model performance. At RTS-rich test sites, the model and experts show good performance (pixel IoU 0.75 vs. 0.72, detection F1 0.80 vs. 0.91). Differences between experts can be attributed to individual human perceptions, interpretations, and variations in the labelling style. Although familiar with DEM-based RTS monitoring and following the same guidelines, experts can interpret landforms differently. This aligns with the findings of Nitze et al. (2024), which revealed even greater differences when comparing the annotations of experts from different permafrost research domains. Both experts and the model face similar challenges with small and sparse RTSs in Gydan, which yield lower accuracies. It is crucial to recognise that human bias in the generation of training data significantly influences the quality of model generalisation and the corresponding predictions. Consequently, the performance of the model is first and foremost constrained by the accuracy and precision of the human annotators, potentially replicating

any biases present in the annotations. Despite this, our findings demonstrate that common deep learning models produce relatively accurate predictions for complex segmentation tasks on remote sensing data, as the achieved accuracies align with the performance levels of expert annotations. However, it is important to note that these predictions contain a certain margin of error, making them more suitable for rough estimates on large scales rather than for precise analysis on local RTS levels.

We attempted to quantify the influence of DEM quality on model performance, although this proved challenging. We selected N Taymyr (TP3) to evaluate the model on descending-descending orbit DEMs, compared to TP1 (ascending-ascending orbit) and TP2 (ascending-descending orbit). However, TP3 spans only three years, compared to six and ten years for TP1 and TP2, respectively. The results suggest that the combination of DEMs from different orbit directions does not significantly impact the performance of the model, although the experts exhibited greater disagreement about the data for TP3. Nevertheless, we cannot isolate the influence of larger time periods from observation properties due to limited time series data. However, during visual inspection of the predictions, we observed that quality issues stemming from satellite acquisitions, such as observations from certain viewing geometries over steep slopes, varying orbit directions of the dDEM pairs, and elevation errors over water bodies, contribute to most FP. Although our developed post-processing strategy mitigates some of these issues (Table 1), smaller rivers or lakes with a high sediment content or artificial elevation changes from infrastructure or mining result in incomplete removal of FP. One potential solution is to perform hard negative mining and explicitly augment the negative training set with examples of these FP. Unlike optical data where geomorphological context and spectral textures allow the model to distinguish human-made from thaw-induced elevation changes, in dDEMs these structures closely resemble RTSs, even to the human eye. Incorporating negative samples that are too similar to actual RTS instances can degrade model performance (Yang et al., 2023), suggesting that these FP may be inherent in the presented method. Nevertheless, we demonstrated that the missing (FN) and overestimated (FP) RTS predictions have a relatively small influence on the total estimated material erosion volume at a regional scale. Our results suggest that the proposed method, with its quantified uncertainties, can provide reasonable large-scale estimates of RTS-induced mass wasting over decadal time frames.

The ability to monitor RTSs with DEMs is hindered by the limited access to comprehensive global high-resolution DEM datasets. TanDEM-X observations offer an opportunity to investigate the mass-wasting activities of RTSs. In optical imagery, RTSs are identified by vegetation disturbances, leaving ambiguity in pinpointing the actively eroding parts of the RTS. Although this approach can observe changes in RTS area, it lacks the ability to quantify volume change, which is crucial to understanding impacts on local hydrology, soil organic carbon mobilisation, and related climate feedback mechanisms. However, in comparison to optical satellite data, which provides at least annual observations, TanDEM-X-generated DEMs do not offer adequate temporal resolution to comprehensively analyse complex and polycyclic RTS dynamics, as generally only two to three DEMs per decade are accessible. Future satellite missions should be designed to enhance RTS monitoring, with the aim of producing high-resolution DEMs for permafrost areas annually (Hajnsek et al., 2022). A complementary approach involves integrating DEM and optical data for comprehensive RTS monitoring: Our automated DEM-based RTS detection method can be used to advance the study of scaling relationships between area and volume, as examined by Van Der Sluijs et al. (2023) and Bernhard et al. (2022a), allowing for more precise temporal resolution to improve the quantification of RTS mass wasting in extensive permafrost regions.

## 5. Conclusions

We evaluated semantic segmentation models based on deep learning, which were trained on dDEMs obtained from TanDEM-X InSAR data. Our method facilitates accurate monitoring of RTS mass wasting, extending beyond the traditional 2D analysis based on optical remote sensing imagery. The best performing model (UNet++) demonstrates robust generalisation across a large dataset that covers key areas of the Arctic permafrost domain, with segmentation (pixel IoU: $0.58 \pm 0.16$, F1: $0.72 \pm 0.15$) and classification performance (detection IoU: $0.73 \pm 0.04$, F1: $0.75 \pm 0.12$) matching or exceeding those of comparable studies. We discussed and compared our results with the accuracy of three experts manually labelling RTSs. The agreement of experts is within the range of the performance of the deep learning model, providing an important context for interpreting the results. Longer temporal spans of dDEMs are advantageous for predicting RTSs. They potentially exhibit larger elevation changes and more extensive active areas, enhancing in turn the detectability of RTSs. By inferring the trained model to the entire study region with an area of $71\,400\,\mathrm{km}^2$, we found RTS volume change rates ranging from 40.75 $\mathrm{m}^3\ \mathrm{yr}^{-1}\ \mathrm{km}^2$ (Chukotka) to 1164.11 $\mathrm{m}^3\ \mathrm{yr}^{-1}\ \mathrm{km}^2$ (Banks) between 2010 and 2021. Our presented method introduces a novel approach for automatically detecting RTS activity using spaceborne elevation data, enabling the empirical quantification of material erosion due to RTS mass wasting across Arctic permafrost environments. All components of the workflow are fully automated; however, integrating these components into a cohesive pipeline will be a focus of future work. To scale this approach to the entire Arctic region, adaptations to the pipeline and robust handling of very large datasets will be essential. The generated dataset, with the computed RTS area and volume change rates, presents a unique opportunity to investigate the drivers and impacts of RTS mass wasting in diverse permafrost environments. The derived data can facilitate further exploration of the interconnections with climate, topography, and carbon cycle feedback, advancing our understanding of rapid permafrost thaw processes.

## CRediT authorship contribution statement

**Kathrin Maier:** Writing – original draft, Visualization, Validation, Software, Methodology, Data curation, Conceptualization. **Philipp Bernhard:** Writing – review & editing, Methodology, Conceptualization. **Sophia Ly:** Writing – review & editing, Investigation, Data curation. **Michele Volpi:** Writing – review & editing, Methodology. **Ingmar Nitze:** Writing – review & editing. **Shiyi Li:** Writing – review & editing. **Irena Hajnsek:** Writing – review & editing, Supervision, Resources, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We would like to thank Gianna Giovanoli for her additional help in the RTS annotations process, as well as the German Aerospace Centre for the provision of the data taken from the TanDEM-X mission and the European Space Agency for the Sentinel-2 data provision. IN was funded by German Federal Ministry for Economic Affairs and Climate Action (ML4Earth50EE2201C), European Space Agency (CCI+Permafrost), National Science Foundation (awards #1927872 and #2052107) and google.org Impact Challenge (Permafrost Discovery Gateway).

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jag.2025.104419.

## Data availability

The dataset containing the spatial extents of the study sites, manual annotations of RTS on differential DEMs, and the model predictions is available at https://doi.org/10.3929/ethz-b-000718475. The code and example data for model training and testing and the RTS mass-wasting calculation can be found at https://github.com/kathrinmaier/dem-rts-segmentation. TanDEM-X CoSSC data can be requested from the German Aerospace Centre.

## References


Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C.K., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M.Y., Suk, H., Zhang, S., Suo, M., Tillet, P., Zhao, X., Wang, E., Zhou, K., Zou, R., Wang, X., Mathews, A., Wen, W., Chanan, G., Wu, P., Chintala, S., 2024. PyTorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In: Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2. pp. 929–947. http://dx.doi.org/10.1145/3620665.3640366.

Bernhard, P., Zwieback, S., Bergner, N., Hajnsek, I., 2022a. Assessing volumetric change distributions and scaling relations of retrogressive thaw slumps across the Arctic. Cryosphere 16 (1), 1–15. http://dx.doi.org/10.5194/tc-16-1-2022.

Bernhard, P., Zwieback, S., Hajnsek, I., 2022b. Accelerated mobilization of organic carbon from retrogressive thaw slumps on the Northern Taymyr Peninsula. Cryosphere 16, 2819–2835. http://dx.doi.org/10.5194/tc-2022-36.

Bernhard, P., Zwieback, S., Leinss, S., Hajnsek, I., 2020. Mapping retrogressive thaw slumps using single-pass TanDEM-X observations. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 13, 3263–3280. http://dx.doi.org/10.1109/JSTARS.2020.3000648.

Biskaborn, B.K., Smith, S.L., Noetzli, J., Matthes, H., Vieira, G., Streletskiy, D.A., Schoeneich, P., Romanovsky, V.E., Lewkowicz, A.G., Abramov, A., Allard, M., Boike, J., Cable, W.L., Christiansen, H.H., Delaloye, R., Diekmann, B., Drozdov, D., Etzelmüller, B., Grosse, G., Guglielmin, M., Ingeman-Nielsen, T., Isaksen, K., Ishikawa, M., Johansson, M., Johannsson, H., Joo, A., Kaverin, D., Kholodov, A., Konstantinov, P., Kröger, T., Lambiel, C., Lanckman, J.-P., Luo, D., Malkova, G., Meiklejohn, I., Moskalenko, N., Oliva, M., Phillips, M., Ramos, M., Sannel, A.B.K., Sergeev, D., Seybold, C., Skryabin, P., Vasiliev, A., Wu, Q., Yoshikawa, K., Zheleznyak, M., Lantuit, H., 2019. Permafrost is warming at a global scale. Nat. Commun. 10 (1), 264. http://dx.doi.org/10.1038/s41467-018-08240-4.

Bojarski, A., Bachmann, M., Boer, J., Kraus, T., Wecklich, C., Steinbrecher, U., Tous-Ramon, N., Schmidt, K., Klenk, P., Grigorov, C., Schwerdt, M., Zink, M., 2021. TanDEM-X long-term system performance after 10 years of operation. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 14, 2522–2534. http://dx.doi.org/10.1109/JSTARS.2021.3055546.

Brodzik, M., NSIDC, 2013. EASE-grid 2.0 land-ocean-coastline-ice masks derived from Boston University global MODIS/Terra land cover data. http://dx.doi.org/10.5067/VY2JQZL9J8AQ, [dataset].

Brown, J., Ferrians, O., Heginbottom, J., Melnikov, E., 2002. Circum-arctic map of permafrost and ground-ice conditions, version 2. http://dx.doi.org/10.7265/SKBG-KF16, [dataset].

Burn, C., Lewkowicz, A., 1990. Canadian landform examples - 17 retrogressive thaw slumps. Can. Geogr. 34 (3), 273–276. http://dx.doi.org/10.1111/j.1541-0064.1990.tb01092.x.

Dai, C., Howat, I.M., Van Der Sluijs, J., Liljedahl, A.K., Higman, B., Freymueller, J.T., Ward Jones, M.K., Kokelj, S.V., Boike, J., Walker, B., Marsh, P., 2024. Applications of ArcticDEM for measuring volcanic dynamics, landslides, retrogressive thaw slumps, snowdrifts, and vegetation heights. Sci. Remote. Sens. 9, 100130. http://dx.doi.org/10.1016/j.srs.2024.100130.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, Li Fei-Fei, 2009. ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. http://dx.doi.org/10.1109/CVPR.2009.5206848.

Fritz, T., Rossi, C., Yague-Martinez, N., Rodriguez-Gonzalez, F., Lachaise, M., Breit, H., 2011. Interferometric processing of TanDEM-X data. In: IEEE International Geoscience and Remote Sensing Symposium. pp. 2428–2431. http://dx.doi.org/10.1109/IGARSS.2011.6049701.

Grosse, G., Harden, J., Turetsky, M., McGuire, A.D., Camill, P., Tarnocai, C., Frolking, S., Schuur, E.A.G., Jorgenson, T., Marchenko, S., Romanovsky, V., Wickland, K.P., French, N., Waldrop, M., Bourgeau-Chavez, L., Striegl, R.G., 2011. Vulnerability of high-latitude soil organic carbon in North America to disturbance. J. Geophys. Res. 116, http://dx.doi.org/10.1029/2010JG001507.

Hajnsek, I., Aðalgeirsdóttir, G., Bartsch, A., Cassola, M., Fischer, G., Jesswein, K., Grosse, G., Haas, C., Huber, S., Kääb, A., Junsu, K., Krieger, G., Mössinger, A., Montpetit, B., Münzenmayer, R., Otto, T., Papathanassiou, K., Almeida, F., Rott, H., Zonno, M., 2022. The ka-band interferometric radar mission proposal for cold environments - SKADI. In: Proc. Adv. RF Sensors Remote Sensing Instrum.. ARSI.

Horn, B., 1981. Hill shading and the reflectance map. Proc. IEEE 69 (1), 14–47. http://dx.doi.org/10.1109/PROC.1981.11918.

Huang, L., Lantz, T.C., Fraser, R.H., Tiampo, K.F., Willis, M.J., Schaefer, K., 2022. Accuracy, efficiency, and transferability of a deep learning model for mapping retrogressive thaw slumps across the Canadian Arctic. Remote. Sens. 14 (12), 2747. http://dx.doi.org/10.3390/rs14122747.

Huang, L., Liu, L., Luo, J., Lin, Z., Niu, F., 2021. Automatically quantifying evolution of retrogressive thaw slumps in Beiluhe (Tibetan Plateau) from multi-temporal CubeSat images. Int. J. Appl. Earth Obs. Geoinf. 102, 102399. http://dx.doi.org/10.1016/j.jag.2021.102399.

Huang, L., Willis, M.J., Li, G., Lantz, T.C., Schaefer, K., Wig, E., Cao, G., Tiampo, K.F., 2023. Identifying active retrogressive thaw slumps from ArcticDEM. ISPRS J. Photogramm. Remote Sens. 205, 301–316. http://dx.doi.org/10.1016/j.isprsjprs.2023.10.008.

Hugonnet, R., Mannerfelt, E., Dehecq, A., Knuth, F., Tedstone, A., 2021. xDEM. http://dx.doi.org/10.5281/ZENODO.4809698.

Iakubovskii, P., 2019. Segmentation models pytorch. URL: https://github.com/qubvel/segmentation_models.pytorch.

Kokelj, S.V., Kokoszka, J., van der Sluijs, J., Rudy, A.C.A., Tunnicliffe, J., Shakil, S., Tank, S.E., Zolkos, S., 2021. Thaw-driven mass wasting couples slopes with downstream systems, and effects propagate through Arctic drainage networks. Cryosphere 15 (7), 3059–3081. http://dx.doi.org/10.5194/tc-15-3059-2021.

Kokelj, S.V., Lantz, T.C., Kanigan, J., Smith, S.L., Coutts, R., 2009. Origin and polycyclic behaviour of tundra thaw slumps, Mackenzie Delta region, Northwest Territories, Canada. Permafr. Periglac. Process. 20 (2), 173–184. http://dx.doi.org/10.1002/ppp.642.

Kokelj, S.V., Lantz, T.C., Tunnicliffe, J., Segal, R., Lacelle, D., 2017. Climate-driven thaw of permafrost preserved glacial landscapes, northwestern Canada. Geology 45 (4), 371–374. http://dx.doi.org/10.1130/G38626.1.

Kokelj, S., Tunnicliffe, J., Lacelle, D., Lantz, T., Chin, K., Fraser, R., 2015. Increased precipitation drives mega slump development and destabilization of ice-rich permafrost terrain, northwestern Canada. Glob. Planet. Change 129, 56–68. http://dx.doi.org/10.1016/j.gloplacha.2015.02.008.

Krieger, G., Moreira, A., Fiedler, H., Hajnsek, I., Werner, M., Younis, M., Zink, M., 2007. TanDEM-X: A satellite formation for high-resolution SAR interferometry. IEEE Trans. Geosci. Remote Sens. 45 (11), 3317–3341. http://dx.doi.org/10.1109/TGRS.2007.900693.

Lacelle, D., Brooker, A., Fraser, R.H., Kokelj, S.V., 2015. Distribution and growth of thaw slumps in the Richardson Mountains–Peel Plateau region, northwestern Canada. Geomorphology 235, 40–51. http://dx.doi.org/10.1016/j.geomorph.2015.01.024.

Lantz, T.C., Kokelj, S.V., 2008. Increasing rates of retrogressive thaw slump activity in the Mackenzie Delta region, N.W.T., Canada. Geophys. Res. Lett. 35 (6), L06502. http://dx.doi.org/10.1029/2007GL032433.

Leinss, S., Bernhard, P., 2021. TanDEM-X: Deriving InSAR height changes and velocity dynamics of great aletsch glacier. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 14, 4798–4815. http://dx.doi.org/10.1109/JSTARS.2021.3078084.

Lewkowicz, A.G., Way, R.G., 2019. Extremes of summer climate trigger thousands of thermokarst landslides in a high arctic environment. Nat. Commun. 10 (1), 1329. http://dx.doi.org/10.1038/s41467-019-09314-7.

Lin, Y., Knudby, A.J., 2023. A transfer learning approach for automatic mapping of retrogressive thaw slumps (RTSs) in the western Canadian Arctic. Int. J. Remote Sens. 44 (6), 2039–2063. http://dx.doi.org/10.1080/01431161.2023.2195571.

Martone, M., Bräutigam, B., Rizzoli, P., Gonzalez, C., Bachmann, M., Krieger, G., 2012. Coherence evaluation of TanDEM-X interferometric data. ISPRS J. Photogramm. Remote Sens. 73, 21–29. http://dx.doi.org/10.1016/j.isprsjprs.2012.06.006.

Millan, R., Dehecq, A., Trouve, E., Gourmelen, N., Berthier, E., 2015. Elevation changes and X-band ice and snow penetration inferred from TanDEM-X data of the Mont-Blanc area. In: 8th International Workshop on the Analysis of Multitemporal Remote Sensing Images (Multi-Temp). pp. 1–4. http://dx.doi.org/10.1109/Multi-Temp.2015.7245753.

Nagler, T., Rott, H., 2000. Retrieval of wet snow by means of multitemporal SAR data. IEEE Trans. Geosci. Remote Sens. 38 (2), 754–765. http://dx.doi.org/10.1109/36.842004.

Nesterova, N., Khomutov, A., Kalyukina, A., Leibman, M., 2020. The specificity of thermal denudation feature distribution on Yamal and Gydan Peninsulas, Russia. http://dx.doi.org/10.5194/egusphere-egu2020-746.

Nesterova, N., Leibman, M., Kizyakov, A., Lantuit, H., Tarasevich, I., Nitze, I., Veremeeva, A., Grosse, G., 2024. Review article: Retrogressive thaw slump characteristics and terminology. Cryosphere 18 (10), 4787–4810. http://dx.doi.org/10.5194/tc-18-4787-2024.

Nitze, I., Grosse, G., Jones, B.M., Romanovsky, V.E., Boike, J., 2018. Remote sensing quantifies widespread abundance of permafrost region disturbances across the Arctic and Subarctic. Nat. Commun. 9 (1), 5423. http://dx.doi.org/10.1038/s41467-018-07663-3.

Nitze, I., Heidler, K., Barth, S., Grosse, G., 2021. Developing and testing a deep learning approach for mapping retrogressive thaw slumps. Remote. Sens. 13 (21), 4294. http://dx.doi.org/10.3390/rs13214294.

Nitze, I., Van der Sluijs, J., Barth, S., Bernhard, P., Huang, L., Kizyakov, A., Lara, M.J., Nesterova, N., Runge, A., Veremeeva, A., Ward Jones, M., Witharana, C., Xia, Z., Liljedahl, A.K., 2024. A labeling intercomparison of retrogressive thaw slumps by a diverse group of domain experts. Permafr. Periglac. Process. 2249. http://dx.doi.org/10.1002/ppp.2249.

Obu, J., Westermann, S., Bartsch, A., Berdnikov, N., Christiansen, H.H., Dashtseren, A., Delaloye, R., Elberling, B., Etzelmüller, B., Kholodov, A., Khomutov, A., Kääb, A., Leibman, M.O., Lewkowicz, A.G., Panda, S.K., Romanovsky, V., Way, R.G., Westergaard-Nielsen, A., Wu, T., Yamkhin, J., Zou, D., 2019. Northern Hemisphere permafrost map based on TTOP modelling for 2000–2016 at 1 km$^2$ scale. Earth-Sci. Rev. 193, 299–316. http://dx.doi.org/10.1016/j.earscirev.2019.04.023.

Olefeldt, D., Goswami, S., Grosse, G., Hayes, D., Hugelius, G., Kuhry, P., McGuire, A.D., Romanovsky, V.E., Sannel, A., Schuur, E., Turetsky, M.R., 2016. Circumpolar distribution and carbon storage of thermokarst landscapes. Nat. Commun. 7 (1), 13043. http://dx.doi.org/10.1038/ncomms13043.

Porter, C., Morin, P., Howat, I., Noh, M.-J., Bates, B., Peterman, K., Keesey, S., Schlenk, M., Gardiner, J., Tomko, K., Willis, M., Kelleher, C., Cloutier, M., Husby, E., Foga, S., Nakamura, H., Platson, M., Wethington, M., Williamson, C., Bauer, G., Enos, J., Arnold, G., Kramer, W., Becker, P., Doshi, A., D'Souza, C., Cummens, P., Laurier, F., Bojesen, M., 2018. ArcticDEM, Version 3. http://dx.doi.org/10.7910/DVN/OHHUKH.

Ramage, J.L., Irrgang, A.M., Morgenstern, A., Lantuit, H., 2018. Increasing coastal slump activity impacts the release of sediment and organic carbon into the Arctic Ocean. Biogeosciences 15 (5), 1483–1495. http://dx.doi.org/10.5194/bg-15-1483-2018.

Rodriguez, E., Martin, J., 1992. Theory and design of interferometric synthetic aperture radars. IEEE Proc. F Radar Signal Process. 139 (2), 147. http://dx.doi.org/10.1049/ip-f-2.1992.0018.

Rosen, P., Hensley, S., Joughin, I., Li, F., Madsen, S., Rodriguez, E., Goldstein, R., 2000. Synthetic aperture radar interferometry. Proc. IEEE 88 (3), 333–382. http://dx.doi.org/10.1109/5.838084.

Runge, A., Nitze, I., Grosse, G., 2022. Remote sensing annual dynamics of rapid permafrost thaw disturbances with LandTrendr. Remote Sens. Environ. 268, 112752. http://dx.doi.org/10.1016/j.rse.2021.112752.

Scheffler, D., Hollstein, A., Diedrich, H., Segl, K., Hostert, P., 2017. AROSICS: An automated and robust open-source image co-registration software for multi-sensor satellite data. Remote. Sens. 9 (7), 676. http://dx.doi.org/10.3390/rs9070676.

Schuur, E.a.G., McGuire, A.D., Schädel, C., Grosse, G., Harden, J.W., Hayes, D.J., Hugelius, G., Koven, C.D., Kuhry, P., Lawrence, D.M., Natali, S.M., Olefeldt, D., Romanovsky, V.E., Schaefer, K., Turetsky, M.R., Treat, C.C., Vonk, J.E., 2015. Climate change and the permafrost carbon feedback. Nature 520 (7546), 171–179. http://dx.doi.org/10.1038/nature14338.

Segal, R.A., Lantz, T.C., Kokelj, S.V., 2016. Acceleration of thaw slump activity in glaciated landscapes of the Western Canadian Arctic. Env. Res. Lett. 11 (3), http://dx.doi.org/10.1088/1748-9326/11/3/034025.

Stewart, A.J., Robinson, C., Corley, I.A., Ortiz, A., Ferres, J.M.L., Banerjee, A., 2022. TorchGeo: deep learning with geospatial data. In: Proceedings of the 30th International conference on Advances in Geographic Information Systems. pp. 1–12. http://dx.doi.org/10.1145/3557915.3560953.

Turetsky, M.R., Abbott, B.W., Jones, M.C., Anthony, K.W., Olefeldt, D., Schuur, E.A.G., Grosse, G., Kuhry, P., Hugelius, G., Koven, C., Lawrence, D.M., Gibson, C., Sannel, A.B.K., McGuire, A.D., 2020. Carbon release through abrupt permafrost thaw. Nat. Geosci. 13 (2), 138–143. http://dx.doi.org/10.1038/s41561-019-0526-0.

Van Der Sluijs, J., Kokelj, S.V., Tunnicliffe, J.F., 2023. Allometric scaling of retrogressive thaw slumps. Cryosphere 17 (11), 4511–4533. http://dx.doi.org/10.5194/tc-17-4511-2023.

Ward Jones, M.K., Pollard, W.H., Jones, B.M., 2019. Rapid initialization of retrogressive thaw slumps in the Canadian high Arctic and their response to climate and terrain factors. Env. Res. Lett. 14 (5), 055006. http://dx.doi.org/10.1088/1748-9326/ab12fd.

Werner, C., Wegmüller, U., Strozzi, T., Wiesmann, A., 2000. Gamma SAR and interferometric processing software. In: European Space Agency (Special Publication) ESA SP. Vol. 461, pp. 211–219.

Witharana, C., Udawalpola, M.R., Liljedahl, A.K., Jones, M.K.W., Jones, B.M., Hasan, A., Joshi, D., Manos, E., 2022. Automated detection of retrogressive thaw slumps in the high arctic using high-resolution satellite imagery. Remote. Sens. 14 (17), 4132. http://dx.doi.org/10.3390/rs14174132.

Xia, Z., Huang, L., Fan, C., Jia, S., Lin, Z., Liu, L., Luo, J., Niu, F., Zhang, T., 2022. Retrogressive thaw slumps along the Qinghai–Tibet Engineering Corridor: a comprehensive inventory and their distribution characteristics. Earth Syst. Sci. Data 14 (9), 3875–3887. http://dx.doi.org/10.5194/essd-14-3875-2022.

Xia, Z., Liu, L., Mu, C., Peng, X., Zhao, Z., Huang, L., Luo, J., Fan, C., 2024. Widespread and rapid activities of retrogressive thaw slumps on the Qinghai-Tibet plateau from 2016 to 2022. Geophys. Res. Lett. 51 (17), e2024GL109616. http://dx.doi.org/10.1029/2024GL109616.

Yang, Y., Rogers, B.M., Fiske, G., Watts, J., Potter, S., Windholz, T., Mullen, A., Nitze, I., Natali, S.M., 2023. Mapping retrogressive thaw slumps using deep neural networks. Remote Sens. Environ. 288, 113495. http://dx.doi.org/10.1016/j.rse.2023.113495.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci. Remote. Sens. Mag. 5 (4), 8–36. http://dx.doi.org/10.1109/MGRS.2017.2762307.

Zwieback, S., Kokelj, S.V., Günther, F., Boike, J., Grosse, G., Hajnsek, I., 2018. Sub-seasonal thaw slump mass wasting is not consistently energy limited at the landscape scale. Cryosphere 12 (2), 549–564. http://dx.doi.org/10.5194/tc-12-549-2018.