Article

A Comparison of Satellite Imagery Sources for Automated Detection of Retrogressive Thaw Slumps

Heidi Rodenhizer *, Yili Yang, Greg Fiske, Stefano Potter, Tiffany Windholz, Andrew Mullen, Jennifer D. Watts and Brendan M. Rogers

Woodwell Climate Research Center, Falmouth, MA 02540, USA; yyang@woodwellclimate.org (Y.Y.); gfiske@woodwellclimate.org (G.F.); spotter@woodwellclimate.org (S.P.); twindholz@woodwellclimate.org (T.W.); amullen@woodwellclimate.org (A.M.); jwatts@woodwellclimate.org (J.D.W.); brogers@woodwellclimate.org (B.M.R.)
* Correspondence: hrodenhizer@woodwellclimate.org

Abstract: Retrogressive thaw slumps (RTS) are a form of abrupt permafrost thaw that can rapidly mobilize ancient frozen soil carbon, magnifying the permafrost carbon feedback. However, the magnitude of this effect is uncertain, largely due to limited information about the distribution and extent of RTS across the circumpolar region. Although deep learning methods such as Convolutional Neural Networks (CNN) have shown the ability to map RTS from high-resolution satellite imagery (≤ 10 m), challenges remain in deploying these models across large areas. Imagery selection and procurement remain one of the largest challenges to upscaling RTS mapping projects, as the user must balance cost with resolution and sensor quality. In this study, we compared the performance of three satellite imagery sources that differed in terms of sensor quality and cost in predicting RTS using a Unet3+ CNN model and identified RTS characteristics that impact detectability. Maxar WorldView imagery was the most expensive option, with a ground sample distance of 1.85 m in the multispectral bands (downloaded at 4 m resolution). Planet Labs PlanetScope imagery was a less expensive option with a ground sample distance of approximately 3.0–4.2 m (downloaded at 3 m resolution). Although PlanetScope imagery was downloaded at a higher resolution than WorldView, the radiometric footprint is around 10–12 m, resulting in less crisp imagery. Finally, Sentinel-2 imagery is freely available and has a 10 m resolution. We used 756 RTS polygons from seven sites across Arctic Canada and Siberia in model training and 63 RTS polygons in model testing. The mean IoU of the validation and testing data sets were 0.69 and 0.75 for the WorldView model, 0.70 and 0.71 for the PlanetScope model, and 0.66 and 0.68 for the Sentinel-2 model, respectively. The IoU of the RTS class was nonlinearly related to the RTS Area, showing a strong positive correlation that attenuated as the RTS Area increased. The models were better able to predict RTS that appeared bright on a dark background and were less able to predict RTS that had higher plant cover, indicating that bare ground was a primary way the models detected RTS. Additionally, the models performed less well in wet areas or areas with patchy ground cover. These results indicate that all imagery sources tested here were able to predict larger RTS, but higher-quality imagery allows more accurate detection of smaller RTS.

Keywords: Arctic; permafrost; retrogressive thaw slump; deep learning; remote sensing

1. Introduction

As the Arctic warms roughly four times faster than the global average [1], permafrost is thawing, causing these ecosystems to release CO_2 and CH_4 into the atmosphere and accelerate climate change [2,3]. While gradual increases in active layer thickness, the seasonally thawed layer at the top of the soil profile, have been well documented across many sites in the Arctic [4–15], melting of excess ice contained within permafrost introduces the potential for ground subsidence and thermokarst, the formation of depressions caused

by differential rates of subsidence across the landscape [16–21]. Although field-based studies on the impact of thermokarst processes on carbon cycling are too spatially and temporally limited to provide circumpolar estimates [17,22–27], initial modeling efforts suggest that thermokarst processes could double the warming impact of gradual permafrost thaw [19]. Additionally, abrupt thaw accelerates lateral export of particulate and dissolved organic carbon in water [28,29] and negatively impacts the travel and subsistence activities of Arctic residents [30]. The combined global and local repercussions of this permafrost carbon feedback make it important to understand where and how quickly permafrost is thawing. Despite being widespread, thermokarst formation is challenging to detect due to its sporadic nature in both time and space [31–37], meaning that circumpolar maps of thermokarst disturbances do not exist. In part due to this, thermokarst is not represented in Earth system models [38].

Although some forms of thermokarst, such as retrogressive thaw slumps (RTS), can be visible in satellite imagery, and machine learning models have improved automated detection of these features, mapping these features across the circumpolar region still poses considerable challenges [39]. RTS are formed when thermoerosion exposes ice-rich permafrost on a slope, forming a steep headwall above a highly disturbed slump floor where bare ground is exposed [40,41]. Once exposed, the headwall continues to thaw and erode, growing retrogressively up the slope at a rate of up to tens of meters per year [42–46]. RTS can be active for decades before they stabilize and vegetation can re-establish [47]. A number of remote sensing data types and techniques, such as LiDAR, InSAR, and multispectral imagery, have been used to identify RTS [48–52]. Typically, these data sources have been combined with manual image interpretation by researchers for RTS detection in bounded study areas [33,42,45,50,52–59]. LiDAR and InSAR methods rely on the elevational profile or elevational changes of RTS, while multispectral imagery relies on the distinctive appearance of RTS due to steep headwalls that cast dark shadows and exposed ground in the slump floor. Multispectral satellite imagery is particularly suited to RTS mapping due to the greater availability of imagery with circumpolar coverage at sufficiently high temporal and spatial resolutions to detect these small (typically <10 ha) disturbances. For these reasons, methods for automating the detection of RTS from multispectral imagery have been developed in recent years [39,60–65].

Various multispectral satellite imagery products have been used for mapping RTS with trade-offs between sensor quality, spatial resolution, temporal resolution, and cost. Two proprietary satellite imagery products that have been commonly used in recent efforts to map RTS features are WorldView (produced by Maxar) and PlanetScope (produced by Planet Labs) [39,61–63]. WorldView imagery is available at up to 1.8 m spatial resolution for the multispectral sensor and 0.46 m for the panchromatic sensor and has a revisit time of up to 1.1 days. While WorldView imagery is very high quality, the cost can be prohibitive at continental or global scales, although choosing a lower resolution for download reduces the cost. The PlanetScope constellation of approximately 130 CubeSats provides daily global coverage with a ground sample distance of 3.0–4.2 m (reprocessed to 3 m resolution in the orthorectified products). These satellites provide less crisp images than larger satellites like WorldView, but this is balanced by faster revisit times and lower costs. Open-access imagery has lagged behind proprietary imagery in spatial and temporal resolution, but Sentinel-2 imagery may be sufficient for mapping RTS features at large scales. With a 10 m resolution in the visible and near-infrared bands and a revisit time of 5 days, the spatial resolution of Sentinel-2 multispectral imagery may be sufficient to map many RTS features, although it would likely miss many smaller features. Because this imagery is freely available and easy to access in Google Earth Engine (GEE) [66], the barriers to use are minimal.

In this study, we compared the performance of WorldView, PlanetScope, and Sentinel-2 imagery in mapping RTS features using a convolutional neural network (CNN) in order to help inform the selection of imagery for future automated RTS mapping and identified characteristics of RTS which impact their detectability. We expected that WorldView

imagery would provide the best RTS classification, followed by PlanetScope, and then Sentinel-2, as WorldView imagery appears significantly more crisp than the other imagery sources. As performance is not the only consideration, however, we also aimed to determine if the performance of either PlanetScope or Sentinel-2 imagery was sufficient to justify the lower (or nonexistent) cost of these imagery sources. To this end, we compared the performance of each image type across RTS feature sizes, determined the RTS feature size at which reliable results can be obtained from each, and identified characteristics of RTS that can improve or hinder model predictions.

2. Materials and Methods

2.1. Study Regions

We selected seven sites across Arctic Canada and Russia where RTS have previously been identified (Figure 1; see Nitze et al. 2021; Yang et al. 2023 for further details) [39,63]. These sites span a broad range of environmental conditions from taiga to Arctic desert. Additionally, a subset of RTS features had already been manually delineated at these sites for model training and validation. In Arctic Canada, the study regions included Banks Island, Herschel Island, Horton Delta, and Tuktoyaktuk Peninsula. In Russia, the sites included Kolguev Island, a small section of the Lena River, and the Yamal and Gydan Peninsulas. Banks Island, Horton Delta, the Yamal and Gydan Peninsulas, and the Lena River site tend to have RTS formation along lakes or rivers, while Herschel Island, Tuktoyaktuk Peninsula, and Kolguev Island tend to have RTS formation along the coast. Herschel Island and the Yamal and Gydan Peninsulas have the smallest RTS features (typically <1 ha), while Banks Island has the largest features (exceeding 10 ha in some cases). All regions were used in model training, and all sites except the Lena River and Kolguev Island were used during model testing.

Figure 1. Map of the locations of RTS features used in model testing. The Arctic Circle is shown as a dashed line. Regions used only in model training are shown in gray, while all other regions, which were used in model training and testing, are coded by color.

2.2. Data

2.2.1. WorldView

We purchased 10,551 individual WorldView images taken during the summer between 2003 and 2020 (75% later than 2015) at 4 m resolution (Imagery ©2017–2021 Maxar, Westminster, CO, USA; [63]) from imagery with a ground sample distance of 1.85 m. No pre-processing of the images was deemed necessary. A single composite image in the visible bands (RGB) was created for each region in Google Earth Engine.

2.2.2. PlanetScope

Composite images were created from PlanetScope PSScene surface reflectance imagery (ortho_analytic_4b_sr and ortho_analytic_8b_sr bundles) for each study region [67]. This imagery has a ground sample distance of approximately 3.0–4.2 m, depending on the altitude and satellite version, and is available for download at 3 m resolution. However, the full width at half maximum (FWHM) of PlanetScope ranges from 3.59 to 3.70 pixels in the visible bands [68], indicating that the light included in the reflectance values of each pixel is actually sourced from an approximately circular area with a diameter of at least 10 m. We selected images for download programmatically and applied additional pre-processing to improve geometric and radiometric calibration. Pre-processing was completed in Python, as the geometric and radiometric calibration algorithms we used have been implemented in Python, and compositing was completed in GEE. All code for PlanetScope download and pre-processing is available on GitHub (https://github.com/whrc/planet_processing).

Images were selected based on year, season, and metadata provided by Planet Labs and supplemented with metadata calculated from daily moderate resolution imaging spectroradiometer (MODIS) imagery. For each composite image, we downloaded enough individual images to cover the entire region ten times. All images were taken during the period between spring snowmelt (the snow-free date was determined from MODIS) and the end of September. We started by downloading images from July 31st and sequentially added images from earlier and later in the growing season as needed to achieve our desired image coverage. Years were selected to correspond to the dates of imagery used in manual RTS delineation: 2018 for the Yamal/Gydan region and 2018–2019 for the remaining regions. Cloud cover metadata provided by Planet Labs was used to further filter imagery for download (max 40% cloud cover, but typically much less) and was supplemented by cloud cover estimates from concurrent MODIS imagery, as there were a number of Planet images with 100% cloud cover that were marked as 0% cloud cover in the Planet metadata, and MODIS data could often catch these falsely marked images. While this method had the potential to falsely mark clear PlanetScope images as cloudy based on the MODIS metadata, we found that far more cloudy images were removed in this step than clear ones, thereby avoiding the download of large quantities of cloudy imagery.

Following download, PlanetScope images were processed to improve geometric and radiometric calibration and combined into composite images (Figure S1). Within each region, each image was geometrically aligned to a common Sentinel-2 composite image from the same time period using the Automated and Robust Open-Source Image Co-Registration (AROSICS) algorithm [69]. Second, variation in the magnitude of surface reflectance values between images was reduced by radiometrically calibrating to the Sentinel-2 composite using the Iteratively Reweighted Multivariate Alteration Detection (IRMAD) algorithm (Figure S2). This algorithm identifies invariant pixels that can be used to develop a linear regression between the two images [70,71]. The calibrated images were finally combined into composite RGB+Near-Infrared (NIR) images (bands 1,2, 3, and 4 for the four-band product, and bands 2, 4, 6, and 8 for the eight-band product) for each region by taking the median surface reflectance value of each band on a cell-by-cell basis.

2.2.3. Sentinel-2

A single composite image was made from Sentinel-2 (European Space Agency) data in GEE (https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2

_SR_HARMONIZED#description, accessed on 15 June 2024). Surface reflectance images between days of year 180 (end of June) and 273 (end of September) from 2017–2021 were selected, as there were insufficient cloud-free pixels in the 2018–2019 Sentinel-2 data to create a composite image which corresponded to the RTS delineation dates. A cloud filter was applied to the images using a threshold of 15% from the Sentinel-2: Cloud Probability layer in GEE, and a composite RGB+NIR image was created by taking the median value on a cell-by-cell basis across all remaining images.

2.2.4. Additional Data Sources

In addition to the red, green, and blue (RGB) bands from one of the three imagery sources, we also included five other bands in the imagery used for model training: Near Infrared (NIR), Normalized Difference Vegetation Index (NDVI; $(NIR - R)/(NIR + R)$), Normalized Difference Water Index (NDWI; $(G - NIR)/(G + NIR)$), and elevation data (Figure S3). NIR was derived from the 10 m Sentinel-2 band 8 (WorldView and Sentinel-2 models) or the 3 m PlanetScope (PlanetScope model) band 4 (four-band PSB and PS2.SD sensors) or 8 (eight-band PSB.SD sensors). NDVI and NDWI were derived from Sentinel-2 imagery as additional layers to represent vegetation growth and open water. Two additional layers related to elevation were derived from the ArcticDEM [72]. First, we calculated relative elevation from the ArcticDEM by subtracting a mean filter (30 m kernel) from the raw elevation values. This layer was intended to enhance the signal of the characteristic elevation difference of RTS features, which is often smaller than larger-scale topographic features such as ridges and valleys. Second, we created an enhanced shaded relief layer to highlight the slope and headwall of RTS features. The ArcticDEM is available at 2 m resolution; therefore, all other image bands were upsampled to 2 m resolution prior to model training in order to match the resolution of the ArcticDEM.

2.3. RTS Digitization

RTS feature outlines were manually digitized in ArcGIS Pro using the Esri base imagery for use in model training, validation, and testing, as previously described in Yang et al. [63]. Manually searching for and digitizing RTS features from satellite imagery is a highly time-consuming task as these features are small, spatially and temporally sporadic, and highly changeable. Therefore, we were limited to 888 RTS polygons that were available for use in model training, validation, and testing, even after pooling observations with another lab [39]. Although some of these RTS polygons were previously delineated from individual PlanetScope images by Nitze et al. [39], all polygons were re-delineated from the Esri base imagery for this study to ensure consistency in methods between study regions and to allow the use of higher-resolution imagery in determining "ground truth" than could be used in the rest of the study (Esri base imagery cannot be downloaded for research purposes).

2.4. Deep Learning Model

We trained convolutional neural network models on three different imagery sources to perform semantic segmentation of RTS features. Each model used the Unet3+ architecture [61] with the EfficientnetB7 backbone [73], as this was the best-performing model in Yang et al. [63]. Each model used the eight bands of input data described in Section 2.2, with the only difference between the three models being the source of the RGB base map: WorldView, PlanetScope, or Sentinel-2 (Figure S3). Satellite images and additional data sources were cropped into 256 × 256 pixels using eight-band tiles containing RTS features. All image bands were normalized to the range of 0 to 1. We tested more than 50 combinations of different hyperparameters, loss functions, optimizers, batch sizes, and learning rates manually and adopted the best set for use in model training. We used Tensorflow [74] on Google Colaboratory (A100 GPU on Colab Pro+) to train our model. A total of 756 RTS scenes (85%) were used in training combined with image augmentation techniques (scaling, flipping, affine transformation, elastic transformation, degradation, and dropout) to alleviate overfitting. We used a geographically separated set of RTS polygons for valida-

tion during the model training process (69 features; 8%). The trained models were then used to predict RTS on a geographically separated set of 63 testing image tiles (7%) that the model had not seen before. We manually selected the training, validation, and testing data sets to ensure geographically separated samples because it otherwise would have been possible for RTS that appeared in multiple tiles to be included in more than one of the training, validation, or testing groups. Model segmentation results (probability of RTS) of the testing tiles were reclassified using a threshold of 0.5 to produce maps containing RTS and background classes. For more details on the deep learning methods, please see Yang et al. [63].

2.5. Testing/Analyses

All analyses were completed in R [75], primarily using the terra [76], sf [77,78], and tidyverse packages [79]. All code used in analyzing the model testing results is publicly available on GitHub (https://github.com/whrc/rts_data_comparison).

Intersection over union (IoU) was used to evaluate model performance. IoU is calculated by taking the area of the intersection of a single predicted class and the ground truth for that class divided by the area of the union of the same predicted class and the ground truth. The mean IoU was calculated by taking the mean of the IoU for the RTS class and the IoU of the background class for each model. Although each RTS feature had its own data tile, there were cases where multiple RTS features appeared within a single tile. To ensure that only one RTS feature was being analyzed within each tile, we masked tiles with multiple RTS features such that only pixels that were closer to the RTS feature of interest (the RTS feature for which that tile was created) than to any other RTS feature were included in the calculation. The mean IoU is used for comparison with other models and to allow for test images without RTS to be scored. However, the mean IoU is less useful for evaluating individual RTS because the extent of background pixels is typically much larger than the extent of RTS pixels (Figure S4). Often, this caused mean IoU to be much higher than the RTS IoU, particularly in tiles with small, undetected, or poorly-predicted RTS. Therefore, we used the RTS IoU in the following analyses to ensure that the results reflected the ability of the CNN models to identify RTS specifically.

We determined how RTS feature area impacts the RTS IoU scores by fitting asymptotic curves using the following equation:

$$\text{RTSIoU} = \frac{\text{A} * \text{RTS Area}}{\text{B} + \text{RTS Area}} \tag{1}$$

for each imagery type, where A is a horizontal asymptote representing the maximum (predicted) RTS IoU and B is a vertical asymptote representing the minimum RTS Area included in the nonlinear model.

Various thresholds were calculated to understand how large an RTS feature must be for reliable detection across imagery types (Figure S5). First, we defined the detection threshold as the RTS Area at which the predicted IoU exceeded 0.5, as this reflects the threshold at which 50% of the union of the prediction and ground-truth overlap and is a concrete threshold that researchers can base their imagery decisions on. Although this is a somewhat arbitrary threshold value, we chose to use a value of 0.5 because it is halfway between 0 (the model missed the feature) and 1 (perfect agreement between prediction and ground truth), and the exact value of the threshold does not impact the order of the imagery rankings. Second, we defined the model convergence threshold as the RTS Area at which the slope of the predicted IoU approached zero (slope = 1×10^{-5}; less than 0.1 increase in the IoU per 1 ha), indicating that an increase in the RTS Area beyond this threshold did not improve model predictions. Finally, the maximum RTS IoU was defined as the value of parameter A (horizontal asymptote) from the fit asymptotic curves.

To investigate the characteristics of the imagery that allowed certain features to be predicted better than others, we compared the input data based on the quality of model predictions. We first normalized for the relationship between RTS Area and the IoU based

on the expectation that small RTS features would be too small to be easily detectable in satellite imagery. Individual RTS features for which the RTS IoU fell above the 50% confidence interval (CI) of the nonlinear model were considered to have a "high" prediction quality (i.e., better prediction than expected for the size of the RTS feature), features which fell below the 50% CI were considered to have a "low" prediction quality (i.e., worse prediction than expected for the size of the RTS feature), and all others were considered "expected". For each of these classes, we compared the values of the individual input data layers (e.g., relative elevation) within the RTS outline and in background pixels in order to identify environmental conditions in the imagery that contribute to model performance.

In addition to the RTS Area, RTS shape may impact both the visibility of a feature within different imagery products and the ability of the models to detect those features. For example, long, narrow features might remain invisible in imagery despite covering a large area, while smaller, round features could be visible. Therefore, we analyzed the additional effect that RTS shape has on the relationship between the RTS Area and the RTS IoU using linear models with interaction terms. To quantify shape in a single value, we calculated the Polsby–Popper score [80], which measures the compactness of a shape. Scores approaching 0 have a small area and long perimeter, while a score of 1 indicates a perfect circle. The RTS IoU was modeled using a maximum model that included log-transformed RTS Area, RTS shape, and the interaction between the two. Akaike Information Criterion (AIC) and the coefficient of determination (r^2) were calculated to compare the models.

3. Results

3.1. General Metrics of Model Performance

All three CNN models performed well, with mean IoU scores ≥ 0.68 in testing. Nevertheless, higher-quality images resulted in better model predictions. The mean validation and testing IoU scores across all features were 0.69 and 0.75 for the WorldView model, 0.70 and 0.71 for the PlanetScope model, and 0.66 and 0.68 for the Sentinel-2 model, respectively (Table 1). Testing scores were higher than validation scores due most likely to the small validation and testing data set sizes and high variability in detection difficulty between RTS. In general, the predictions reflected the shape of the ground-truth polygons quite well (Figure 2), although the RTS Area tended to be slightly overestimated, and linear models of ground-truth vs. predicted Polsby–Popper shape showed a statistically insignificant relationship (Figure S6). IoU scores for the RTS class averaged 0.37 for WorldView, 0.30 for PlanetScope, and 0.28 for Sentinel-2, highlighting the difference in performance between imagery types more strongly than the mean IoU (Table 1; Figure 3). The poorer RTS IoU scores indicated that high mean IoU scores could be achieved due to the (usually) larger extent of background pixels within testing tiles, even where the models failed to predict RTS features well. Additionally, the low RTS IoU scores seemed to be caused in large part by small features (by area) that went undetected rather than poor delineation of features that were detected, as RTS IoU scores rose to 0.48, 0.47, and 0.38 for WorldView, PlanetScope, and Sentinel-2, respectively, when undetected features were omitted. Out of 63 testing features, the WorldView model missed 15 (24%), the Sentinel-2 model missed 17 (27%), and the PlanetScope model missed 23 (37%; Table 1). Across all models, all of the undetected features were 0.46 ha or smaller, and >80% of undetected features were 0.15 ha or smaller, compared to a mean RTS Area of 0.93 ha and a median RTS Area of 0.27 ha, indicating that image resolution was insufficient to detect the smallest features (Figure S7). Additionally, we saw evidence of false positives, when non-RTS Areas were sometimes predicted to be RTS by the models (e.g., the WorldView snow example in Figure 2).

Table 1. Metrics of model performance. "Validation" refers to tiles used in the validation steps of model training, and "testing" refers to testing tiles on which predictions were made.

	Validation			Testing				
Imagery	Mean IoU	Mean IoU	RTS IoU	Median RTS IoU	Undetected RTS (Count (%))	Undetected RTS Max Area (ha)	Detection Threshold (ha)	Model Convergence Threshold (ha)
WorldView	0.75	0.76	0.37	0.36	15 (24%)	0.46	0.7	4.69
PlanetScope	0.71	0.75	0.30	0.25	23 (37%)	0.38	1.01	5.09
Sentinel-2	0.68	0.70	0.28	0.23	17 (27%)	0.29	1.51	4.62
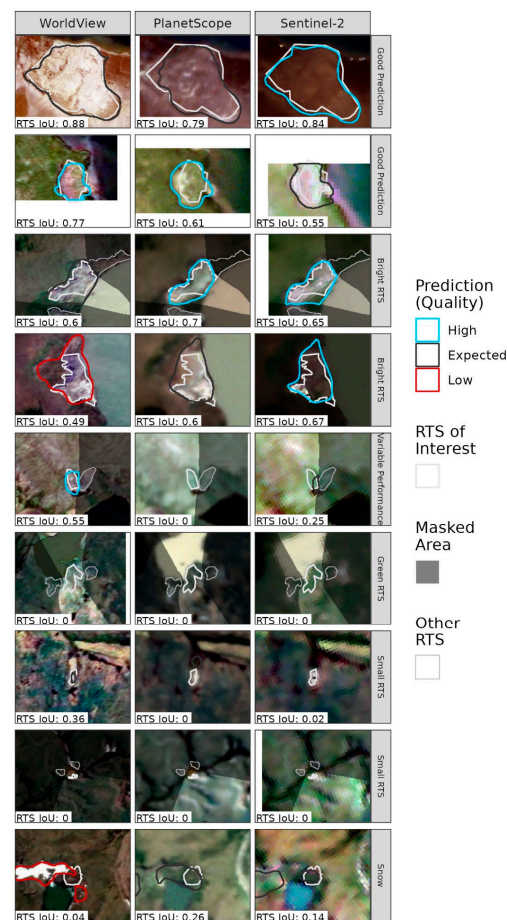
Figure 2. RGB imagery and prediction outlines for a subset of the 63 RTS testing features. The quality of the prediction relative to feature size is indicated by the color of the prediction outline. The RTS feature of interest is shown in light gray. In cases where there are multiple RTS features within a tile, the other RTS features are shown in a thinner light gray line, and the mask area is shown in a dashed light gray line. Columns show the different imagery sources, and rows show different RTS features, which were selected to display differences in the predictions and imagery. Rows labeled "Good Prediction" show predictions that had some of the highest IoU scores. Rows labeled "Bright RTS" show examples of how bright RTS on a dark background were predicted well in the PlanetScope imagery. The row labeled "Variable Performance" shows predictions that varied significantly among imagery types. The row labeled "Green RTS" shows an RTS with a high plant cover that was undetected in all models. The rows labeled "Small RTS" show some of the smaller RTS features, which were often undetected. The row labeled "Snow" shows one example of how snow in the WorldView image was inaccurately labeled as an RTS feature.
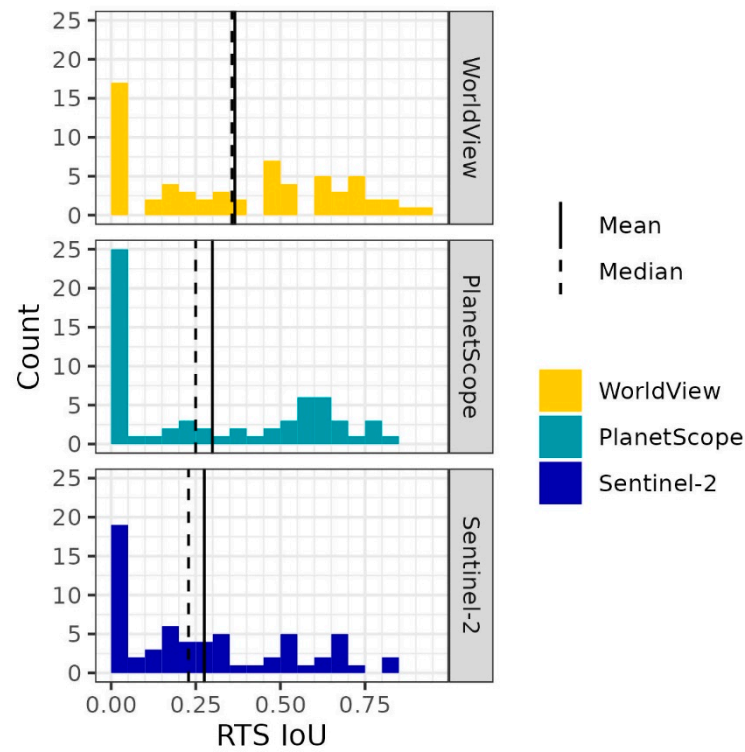
Figure 3. Histograms of the IoU scores of the RTS class across the testing dataset. Mean and median IoU scores are shown as solid and dashed lines, respectively.

3.2. RTS Area and Model Performance

There was a strong, nonlinear relationship between the RTS Area and the RTS IoU. At small sizes, RTS IoU scores increased rapidly as the RTS Area increased, and at larger sizes, the slope of the relationship decreased and eventually plateaued (Figure 4; Table 2). The maximum modeled RTS IoU was 0.83 for the WorldView model, 0.78 for the PlanetScope model, and 0.69 for the Sentinel-2 model (Table 2). The detection threshold (predicted IoU > 0.5) was 0.70 ha for WorldView, 1.01 ha for PlanetScope, and 1.51 ha for Sentinel-2. The model convergence threshold was 4.69 ha for the WorldView model, 5.09 ha for the PlanetScope model, and 4.62 ha for the Sentinel-2 model (Table 1).
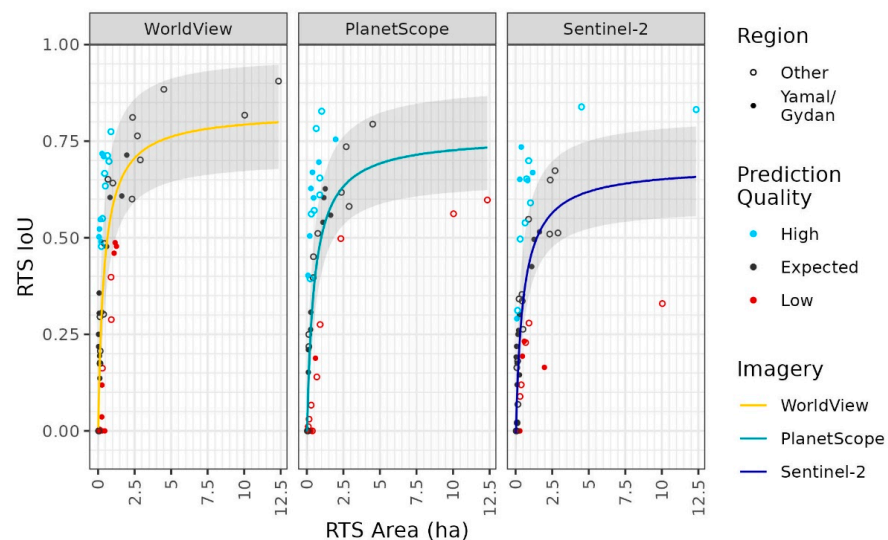
Figure 4. The relationship between RTS Area and RTS IoU. Each point represents the prediction for a single RTS feature. The nonlinear relationship is shown as a solid line, and 50% confidence intervals

are shown in light gray. RTS feature predictions that fell outside of the 50% confidence interval were considered of higher or lower prediction quality than expected, and this is indicated by point color.

Table 2. Parameters of the nonlinear models of RTS IoU by RTS Area for each imagery type.

Imagery	Term	Estimate	SE	*t*-Statistic	*p*-Value
WorldView	K_m	3883.292	1022.569	3.798	<0.001
	V_{max}	0.833	0.077	10.795	<0.001
PlanetScope	K_m	4671.998	1299.752	3.595	0.001
	V_{max}	0.775	0.079	9.799	<0.001
Sentinel-2	K_m	4689.771	1377.305	3.405	0.001
	V_{max}	0.694	0.075	9.275	<0.001

Formula: RTS IoU ~ $V_{max} \times$ RTS Area$(m^2)/(K_m +$ RTS Area$(m^2))$.

3.3. Environmental Drivers of Model Performance

A few RTS features had exceptionally high or low RTS IoU scores for their size, as determined by the 50% CI of the asymptotic models (Figure 4). Specific RTS features tended to have the same or similar prediction quality across imagery types, indicating that there were some characteristics of each feature that made it more or less detectable regardless of imagery type (Figure 5). For example, in the PlanetScope model, RTS features with a high prediction quality had significantly higher surface reflectance values in the visible bands (hereafter "luminance") within the RTS feature than in the background pixels (i.e., bright RTS features on a dark background were detected more easily; Figure 6). This trend was also present in the other two models, although the differences were not statistically significant. Plant cover within RTS also impacted the ability of the model to detect RTS. In both the PlanetScope and Sentinel-2 models, RTS were predicted poorly in cases where plant cover (measured by NDVI and NIR values) within RTS was similar to plant cover in background pixels. Across all imagery types, the models were less able to detect RTS that occurred in areas with variable, or patchy, ground cover (measured by the standard deviation of surface reflectance in the visible bands), indicating that the models struggled to identify the boundary between tundra and RTS pixels when patchy ground cover concealed this transition. Additionally, the PlanetScope model performed poorly in areas with higher NDWI (Figure S8), indicating that wet soils or the presence of lakes could pose a challenge to RTS detection.
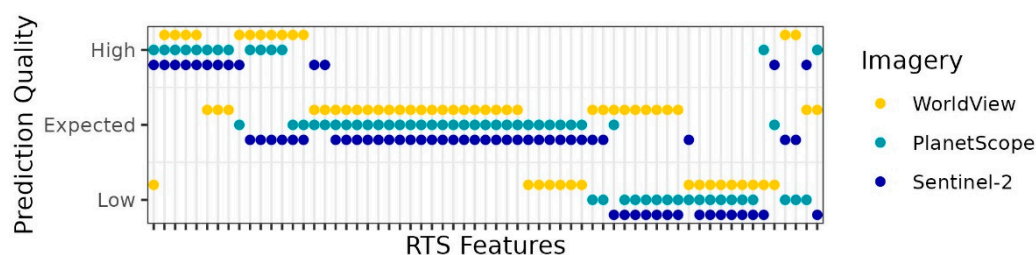

Figure 5. Prediction quality across all imagery types for each RTS feature in the testing dataset. RTS features tended to have the same or similar prediction qualities across imagery types, indicating that there were characteristics of RTS features that made them more or less detectable across imagery types.
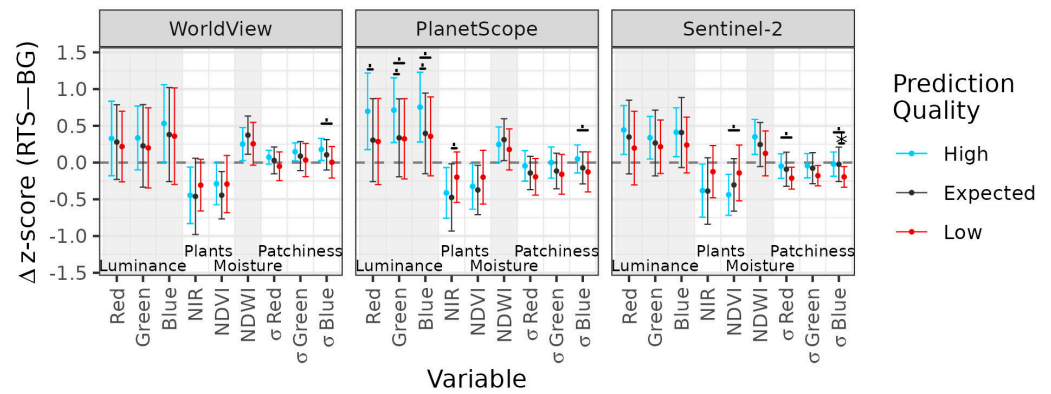
Figure 6. The difference in input data values between RTS and non-water background (BG) pixels (RTS—BG) across prediction quality classes. The points were calculated by first taking the difference in mean pixel values (z-score) between RTS and non-water background (BG) pixels on a tile-by-tile basis and then averaging this value across all 63 testing tiles. The error bars show the standard deviation across tiles. Z-scores were calculated using all pixel values, including water pixels. Relative elevation and shaded relief are not included, as there were no discernable patterns across classes. Statistically different groups are indicated with lines between the two classes and a label for the level of significance ($p < 0.1$: '.', $p < 0.05$: '*').

3.4. Regional Patterns of Model Performance

Prediction quality tended to be lower in the Yamal/Gydan and Banks Island regions than in the other regions within the study (Figure 7). In the Yamal/Gydan region, this seemed to be due primarily to the small size of RTS features (Figure 8). Additionally, RTS features in the Yamal/Gydan region comprised the majority of the undetected features; 73%, 78%, and 82% of the undetected features were located within the Yamal/Gydan region in the WorldView, PlanetScope, and Sentinel-2 models, respectively. The overall poor prediction quality in the Banks Island region did not seem to be related to RTS Area, as RTS features in that region included the largest within the study. However, Banks Island was the most northerly site in this study, and the limited vegetation apparent in the images of this polar desert region may have decreased the difference in appearance between RTS and background tundra pixels. Additionally, the RTS in this region tended to have a less compact morphology than other regions, which may have counteracted the larger size of the RTS somewhat (Figure 8). The high prediction quality of RTS within the Herschel Island region was likely a coincidence due to the small sample size.
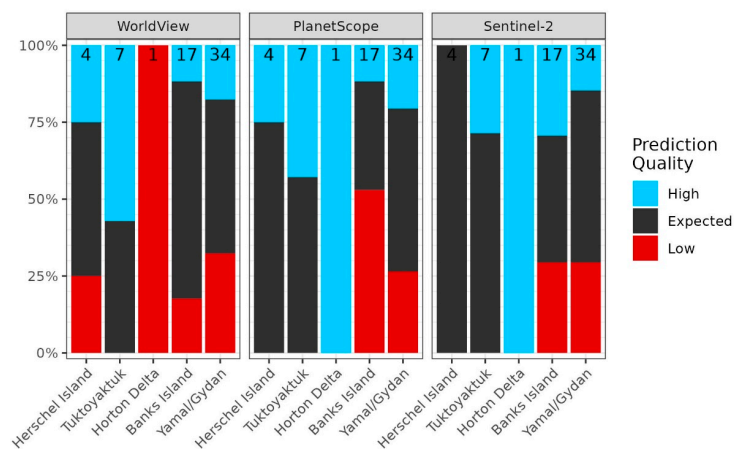

Figure 7. Prediction quality across geographic regions. The percentage of predictions that were high, expected, and low is shown on the Y-axis. The total count of RTS features within each region is indicated at the top of the bars. Banks Island and the Yamal/Gydan region had the highest percentage of low-quality predictions.
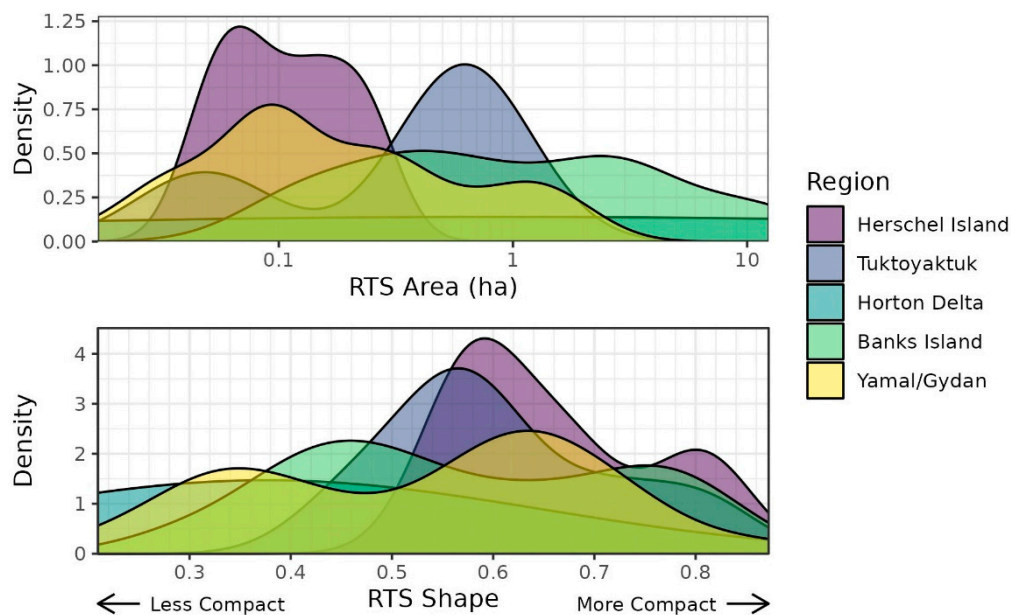
Figure 8. Frequency distributions of RTS Area and RTS shape across regions. Region is indicated with the fill color. RTS Area is shown on a log scale. Smaller RTS shape values indicate less compact shapes, and larger values indicate more compact or circular shapes.

3.5. RTS Shape and Model Performance

In addition to the RTS Area, the RTS shape also influenced the RTS IoU, although the influence of shape was much smaller (Figure S8). Based on AIC, we would have slightly preferred the linear model of the RTS IoU that did not include RTS shape or interactions for each imagery type (Table 3). However, including the RTS shape and the interaction between area and shape showed slightly increased r^2 values for each imagery type. This indicates that the RTS shape and the interaction between area and shape do influence the ability of the three deep learning models to detect RTS, but that with our small sample size, the influence is small enough to be ignored.

Table 3. AIC scores and r^2 values for the linear models of RTS IoU by RTS Area and RTS shape. For each imagery type, all possible models between the intercept-only model and the model with both explanatory variables and their interaction were tested. Using AIC, the model with only RTS Area would be selected for each imagery type (bold). However, the r^2 values indicate that including RTS shape and the interaction between area and shape improves model fit slightly (about 1%) for each imagery type (r^2 bold).

Imagery	Model	AIC	r^2
WorldView	IoU ~ 1	26.52	-
	IoU ~ Area	**−32.752**	**0.622**
	IoU ~ Area + Shape	−30.754	0.622
	IoU ~ Area × Shape	−31.256	**0.637**
PlanetScope	IoU ~ 1	25.699	-
	IoU ~ Area	**−34.658**	**0.628**
	IoU ~ Area + Shape	−34.192	0.637
	IoU ~ Area × Shape	−32.349	**0.638**
Sentinel-2	IoU ~ 1	10.222	-
	IoU ~ Area	**−43.671**	**0.588**
	IoU ~ Area + Shape	−42.324	0.592
	IoU ~ Area × Shape	−41.063	**0.597**

4. Discussion

4.1. Trade-Offs between Imagery Sources

Overall, all three models had mean IoU scores within the range of those reported in other studies that used similar deep learning methods for RTS detection [39,63], although there were considerable differences across the three imagery types. As we hypothesized, WorldView imagery performed the best of the three imagery types across all metrics. The WorldView model had a smaller detection threshold than the other two models and also had a higher maximum RTS IoU, indicating that the WorldView model was the best at predicting RTS across all RTS Areas. At small RTS Areas, this can be explained purely by the sensor specifications; the higher-quality of WorldView imagery makes smaller objects visible and larger objects appear crisper, which allows for improved classification [81]. The fact that this improved performance even for larger RTS that were readily visible in both PlanetScope and Sentinel-2 may suggest that the models relied on smaller characteristics of RTS features that do not scale linearly with RTS Area and are more accurately represented by higher-resolution imagery. For example, the size of the RTS headwall is unlikely to change as rapidly as the area of the entire feature: the headwall length and depth could change somewhat, and sun direction and angle will affect its visibility across images, but the headwall will always remain a narrow strip in satellite imagery. Therefore, all else being equal, WorldView imagery should allow better detection of smaller characteristics of RTS regardless of the total area of the feature.

While all of the models were unable to detect some RTS features, none of the undetected features were larger than 0.46 ha, and the maximum size of undetected features did not correspond to image resolution, although this seems largely related to the small sample size (Figure S6). Interestingly, the PlanetScope model had more undetected RTS features than the Sentinel-2 model despite its higher resolution, higher maximum RTS IoU, and smaller detection threshold. This could be related to the large FWHM of the PlanetScope sensors [68], which results in each pixel reflecting an area on the ground similar to a Sentinel-2 pixel, or it could be related to the magnitude of geometric offsets between each of the imagery sources and the ArcticDEM. It could also be that the Sentinel-2 model actually performed comparably well or better than the PlanetScope model but that the combination of low sample sizes (particularly of large RTS) and one low prediction quality outlier around 10 ha in the Sentinel-2 model was responsible for reducing the maximum predicted IoU relative to the PlanetScope model (Figure 4). Without a larger testing dataset, particularly at large RTS Areas, however, this potential explanation is inconclusive.

Based on our findings, we recommend using the highest resolution imagery possible, taking into account the FWHM or other metrics of line spread of each imagery source. WorldView imagery, which had the smallest radiometric footprint of the options tested here, despite being downloaded at a lower resolution than PlanetScope imagery, was able to detect smaller RTS features and more accurately delineate RTS features of all sizes than the models trained and predicted on the other imagery options. This is due to the large FWHM of the PlanetScope sensors, which results in image clarity that is more similar to Sentinel-2 than WorldView. Even with the 4 m WorldView imagery, though, model performance suffered as the RTS Area decreased. While improvements in the size of the training set or model framework could alleviate this issue somewhat, imagery resolution seems to be the primary limiting factor, and access to higher-resolution imagery could improve model performance considerably.

Recognizing that the cost of high-resolution imagery can be prohibitive, our results can help inform when and where to make compromises. First, we recommend that researchers consider the typical size of RTS features in a given region and compare that to the resolution/quality of possible imagery choices. For example, in an area like the Yamal/Gydan region in this study, where the median RTS Area was 0.12 ha, the cost of higher-quality imagery seems necessary, as even WorldView only had a predicted RTS IoU of 0.18 at 0.12 ha (PlanetScope: 0.13, Sentinel-2: 0.12). On Banks Island, however, where the median RTS Area was 0.91 ha, Sentinel-2 imagery may be sufficient, with a predicted RTS

IoU of 0.43, although higher-quality imagery could still improve performance. Second, if sufficient ground-truth data exist, it would be helpful to consider the total extent of RTS features that are too small to reliably detect. For example, if RTS features smaller than the detection threshold for a specific imagery type compose 50% of the total extent of RTS features within a region, then higher-resolution imagery would be strongly recommended. If, however, the extent of small RTS features is only 5% of the total extent, higher-resolution data may not deliver sufficient improvement in results to be worth the cost. While these recommendations cannot remove all subjectivity from this process, they should decrease the guesswork and help guide researchers in imagery selection.

4.2. RTS Area and Shape

Around 60% of the variability in the RTS IoU was explained by the RTS Area for each of the imagery types, and adding the RTS shape increased this by about 1%, even though AIC scores indicated that the RTS shape should not be included in the linear regressions. We concluded that the RTS shape likely does impact the visibility of RTS features, but that the small sample size in the linear regressions (63 points) meant that the error was too large to conclusively recommend the inclusion of this variable in the regression. Theoretically, the mechanism by which the RTS shape would determine the visibility of RTS, particularly small RTS, is clear: a very elongated shape will have a small width relative to a compact (round) shape, and the width of a shape should determine whether it appears as a discrete feature in the imagery of a certain resolution. Therefore, we expect that there should be an interaction between RTS Area and RTS shape in determining how well RTS features can be predicted from different imagery types, and that this relationship would be more apparent with a larger testing dataset. Researchers could use this as an additional piece of information when deciding on imagery requirements for mapping tasks, particularly in cases where an analysis of the size distribution of RTS does not provide a clear answer.

4.3. Characteristics Affecting RTS Detection

We were not able to directly inspect the patterns that were important to the CNN models because there are few methods for the interpretation of AI models, and we were unable to find a suitable method for our needs [82]. Instead, we employed a relatively simplistic statistical test of individual input data layers to determine important factors for RTS detection. These statistical tests cannot fully explain the complex spatial patterns or relationships between different layers of the input data that are represented within convolutional layers of the CNN model. However, these statistical tests still elucidated important factors in the ability of the CNN models to detect RTS and highlighted environmental conditions that pose challenges to mapping RTS, which future research can address.

Beyond RTS Area and shape, we found that RTS luminance, RTS plant cover, and the nature of the landscape in which RTS occurred were important in determining how well RTS features could be predicted. The importance of luminance in RTS detection was very prominent in a handful of the PlanetScope tiles where RTS features appeared only as a blurry clump of bright pixels (Figure 2). The high luminance of these pixels was likely due to the relatively high reflectance of bare soil, which is a prominent characteristic of RTS [83]. Additionally, higher plant cover within RTS made it harder for the models to detect RTS, consistent with our conclusion that the models relied on bare-ground pixels to identify RTS. For both luminance and variables related to plant cover, it seems plausible that the lack of statistical significance in the WorldView model could be because the model could rely on more complex and detailed spatial patterns that we were not able to directly test. Finally, RTS detection in regions that were particularly wet or had patchy ground cover was more challenging to the PlanetScope model, indicating that areas that look more disturbed make it harder for the model to detect the differences between RTS and background pixels. Therefore, additional training data are needed in challenging areas to improve model performance.

Across all three models, relative elevation and shaded relief derived from the Arctic-DEM did not have a discernible impact on model performance, despite the clear elevation changes associated with RTS. This could be due to insufficient vertical resolution of the ArcticDEM (absolute vertical precision of the ArcticDEM has not been verified), image quality issues such as noise and artifacts in the ArcticDEM [51], and issues with geometric alignment and temporal offsets between the various imagery sources and the ArcticDEM. It is possible that mismatches between input data layers reduced the statistical strength of our tests, that the quality of the elevation data was insufficient, or that these variables were actually less important in the CNN models. With our current methods and data sources, we were unable to definitively disentangle these possibilities, although we suspect that the quality and vertical resolution of the ArcticDEM is the primary reason that elevation did not emerge as an important variable. Luckily, recent improvements in methodology, quantity of input data, and quality control to the ArcticDEM in version 4.1 should alleviate data quality issues in future RTS mapping work [84]. Additionally, it may be possible that different methods for calculating relative elevation or shaded relief maps could improve the usefulness of these layers in predicting RTS. For example, testing different distances over which to calculate the relative elevation, or including multiple layers that each use a different distance, could allow larger or smaller RTS or components of RTS (e.g., headwall, floor) to be better targeted [17].

4.4. Regional Challenges to RTS Detection

Even by using WorldView imagery, the highest quality imagery tested here, 32% of the RTS features in the Yamal/Gydan region were undetected. This number rose to 41% in the Sentinel-2 imagery and to 53% in the PlanetScope imagery. Additionally, there were a higher number of RTS features with low prediction quality and a lower number of RTS features with high prediction quality in this region, indicating that the unusually small size of features in this region was not the only factor causing poor RTS prediction. We originally hypothesized that this might be due to different RTS morphology in the region; if the RTS features in this region tended to be less compact and, therefore, had a narrower width, it would follow that they would be less visible in the imagery. However, this hypothesis proved to be incorrect, as the RTS features in the Yamal/Gydan region were slightly more compact than features in other regions and should, therefore, be slightly more visible on average than RTS of the same size in other regions. Additionally, RTS features on Banks Island tended to have a low prediction quality relative to other features of similar size. This could be explained by the relatively shallow headwall heights of features in this region [33], as this would reduce the visibility of the headwall and its shadow in satellite images or by the sparse Arctic desert vegetation reducing the contrast between RTS and background pixels.

4.5. Challenges Associated with RTS Delineation

In a few cases, we noticed that the RTS predictions seemed more aligned with the imagery they were predicted on than to the ground-truth polygons, indicating that reported IoU scores may have been reduced by inaccurate ground-truth polygons. There are many possible mechanisms through which this can occur, and this will be an ongoing challenge to the automated detection of RTS. One major reason for this is that the ground-truth RTS polygons were delineated on different imagery than was used for predictions (high-resolution ESRI basemap imagery used in delineations cannot be exported and used for other tasks), and sometimes multiple imagery sources were used in this process. This allowed us to use high-quality imagery for delineation and confirm through multiple images that disturbances noted on the ground were actually RTS. However, this also means that there could be differences between images used for delineation and prediction due to misalignment and temporal offsets. Despite our best efforts to address these issues, we found a few examples in which the RTS feature had evidently expanded between the date of the imagery used for delineation and the date of the imagery used for prediction

(see the second row in Figure 2). Additionally, manual delineation of RTS features is an inherently subjective task, and features delineated by different people or from different imagery can differ significantly [85]. Inconsistent delineation could have an adverse effect on model performance.

4.6. Remaining Challenges and Future Improvements

This study was motivated by the reality of limitations to accessing and purchasing high-resolution imagery, which remains a major barrier to conducting large-scale geospatial science for many applications, especially in the rapidly warming permafrost region. Although high-resolution images of the planet are being produced at an unprecedented rate, the time and cost required to access this imagery can still be considerable. Even when funding can be procured for high-resolution imagery, the time required to reach licensing agreements with proprietary imagery providers, download imagery, and provide for data storage slows the pace of research. Although some proprietary imagery can be made available to researchers funded by government grants through agreements between the granting institution and the imagery provider, our experience indicates that this access is typically limited and insufficient for global-scale scientific research. Open access to high-resolution satellite imagery for research purposes (e.g., requiring government-facilitated data buys) would, therefore, rapidly accelerate scientific progress in geospatial fields, with implications for human well-being in a changing world.

There are a number of possible directions for future research to improve model performance and scale-up RTS predictions. First, additional training, validation, and testing polygons are needed to improve model performance [61,63] and increase statistical power in analyses of model predictions. A current effort to pool currently available RTS polygon data created by different lab groups and allow ongoing incorporation of new data as they are created will soon allow researchers access to a much larger dataset for these purposes [86]. As this data set grows, it will be important to ensure that new contributions are targeted to represent the wide variety of environments in which RTS occur. Second, current RTS models have a high false positive rate [39,63,87]. The addition of "negative" training data (i.e., imagery tiles without RTS features) is one step that could limit false positives in future wall-to-wall predictions, but other methods, such as incorporating data layers that represent landscape change, such as LandTrendr [64,88], or using model ensembles to reduce non-RTS noise in model outputs could also be necessary. Third, adherence to standardized methods for RTS delineation is required to decrease differences in RTS delineation between people and lab groups. Guidelines are currently being developed by the International Permafrost Association RTSInTrain Action group [85]. Fourth, the incorporation of the newest ArcticDEM into RTS modeling efforts will reduce errors associated with poor elevation data. Fifth, further image pre-processing methods could be considered to improve image quality and resolution (e.g., point spread deconvolution or creation of super-resolution images) [81,89,90]. Finally, there are still computational challenges associated with deploying these models in wall-to-wall predictions across the Arctic, and research is needed to establish a streamlined workflow for this process. Addressing these challenges will ultimately lead to large-scale mapping of RTS and other thermokarst features, which is desperately needed to monitor the rapidly warming permafrost zone and inform models of permafrost carbon–climate feedbacks.

5. Conclusions

Thermokarst processes are known to exert a large impact on Arctic carbon cycling but have yet to be incorporated into Earth system models, in part due to the lack of geospatial products available for these processes. RTS are one form of thermokarst for which automated mapping seems feasible, but challenges, such as imagery acquisition, remain for developing a circumpolar map. We compared three CNN models trained on different imagery sources for detecting RTS in order to inform choices related to imagery procurement. For these small, sporadic features, we found that it is important to procure the highest

resolution data possible and that considering metrics of radiometric footprint/line spread is imperative in this process. Additionally, we analyzed how the spectral characteristics of RTS features determine variability in model performance. RTS features that appeared bright were easier for the models to detect, while increased plant cover within RTS and wet or patchy ground cover surrounding RTS hindered model performance.

References

1. Rantanen, M.; Karpechko, A.Y.; Lipponen, A.; Nordling, K.; Hyvärinen, O.; Ruosteenoja, K.; Vihma, T.; Laaksonen, A. The Arctic Has Warmed Nearly Four Times Faster than the Globe since 1979. *Commun. Earth Environ.* **2022**, *3*, 168. [CrossRef]
2. Schuur, E.A.G.; Abbott, B.W.; Commane, R.; Ernakovich, J.; Euskirchen, E.; Hugelius, G.; Grosse, G.; Jones, M.; Koven, C.; Leshyk, V.; et al. Permafrost and Climate Change: Carbon Cycle Feedbacks From the Warming Arctic. *Annu. Rev. Environ. Resour.* **2022**, *47*, 343–371. [CrossRef]
3. Miner, K.R.; Turetsky, M.R.; Malina, E.; Bartsch, A.; Tamminen, J.; McGuire, A.D.; Fix, A.; Sweeney, C.; Elder, C.D.; Miller, C.E. Permafrost Carbon Emissions in a Changing Arctic. *Nat. Rev. Earth Environ.* **2022**, *3*, 55–67. [CrossRef]
4. Brown, J.; Hinkel, K.M.; Nelson, F.E. The Circumpolar Active Layer Monitoring (Calm) Program: Research Designs and Initial Results. *Polar Geogr.* **2000**, *24*, 166–258. [CrossRef]
5. Hinkel, K.M.; Nelson, F.E. Spatial and Temporal Patterns of Active Layer Thickness at Circumpolar Active Layer Monitoring (CALM) Sites in Northern Alaska, 1995–2000. *J. Geophys. Res.* **2003**, *108*, 8168. [CrossRef]

6. Streletskiy, D.A.; Shiklomanov, N.I.; Nelson, F.E.; Klene, A.E. Thirteen Years of Observations at Alaskan CALM Sites: Long-Term Active Layer and Ground Surface Temperature Trends. In Proceedings of the Ninth International Conference on Permafrost, Fairbanks, AK, USA, 29 June–3 July 2008; Volume 2, pp. 1727–1732.

7. Nyland, K.E.; Shiklomanov, N.I.; Streletskiy, D.A.; Nelson, F.E.; Klene, A.E.; Kholodov, A.L. Long-Term Circumpolar Active Layer Monitoring (CALM) Program Observations in Northern Alaskan Tundra. *Polar Geogr.* **2021**, *43*, 167–185. [CrossRef]

8. Strand, S.M.; Christiansen, H.H.; Johansson, M.; Åkerman, J.; Humlum, O. Active Layer Thickening and Controls on Interannual Variability in the Nordic Arctic Compared to the circum-Arctic. *Permafr. Periglac. Process.* **2021**, *32*, 47–58. [CrossRef]

9. Duchesne, C.; Smith, S.L.; Ednie, M.; Bonnaventure, P.P. Active Layer Variability and Change in the Mackenzie Valley, Northwest Territories. In Proceedings of the 68th Canadian Geotechnical Conference and Seventh Canadian Conference on Permafrost (GEOQuébec 2015), Quebec, QC, Canada, 20–23 September 2015.

10. Burn, C.R.; Lewkowicz, A.G.; Wilson, M.A. Long-term Field Measurements of Climate-induced Thaw Subsidence above Ice Wedges on Hillslopes, Western Arctic Canada. *Permafr. Periglac. Process.* **2021**, *32*, 261–276. [CrossRef]

11. Nixon, F.M.; Taylor, A.E. Regional Active Layer Monitoring across the Sporadic, Discontinuous and Continuous Permafrost Zones, Mackenzie Valley, Northwestern Canada. In Proceedings of the Seventh International Conference on Permafrost, Yellowknife, NT, Canada, 23–27 June 1998; Volume 55, pp. 815–820.

12. Shiklomanov, N.I.; Streletskiy, D.A.; Nelson, F.E. Northern Hemisphere Component of the Global Circumpolar Active Layer Monitoring (CALM) Program. In Proceedings of the Tenth International Conference on Permafrost, Salekhard, Russia, 25–29 June 2012; Volume 1, pp. 377–382.

13. Smith, S.L.; Wolfe, S.A.; Riseborough, D.W.; Nixon, F.M. Active-Layer Characteristics and Summer Climatic Indices, Mackenzie Valley, Northwest Territories, Canada. *Permafr. Periglac. Process.* **2009**, *20*, 201–220. [CrossRef]

14. O'Neill, H.B.; Smith, S.L.; Duchesne, C. Long-Term Permafrost Degradation and Thermokarst Subsidence in the Mackenzie Delta Area Indicated by Thaw Tube Measurements. In *Cold Regions Engineering 2019*; American Society of Civil Engineers: Quebec City, QC, Canada, 2019; pp. 643–651.

15. Schuur, E.A.G.; Bracho, R.; Celis, G.; Belshe, E.F.; Ebert, C.; Ledman, J.; Mauritz, M.; Pegoraro, E.F.; Plaza, C.; Rodenhizer, H.; et al. Tundra Underlain by Thawing Permafrost Persistently Emits Carbon to the Atmosphere over 15 Years of Measurements. *J. Geophys. Res. Biogeosci.* **2021**, *126*, e2020JG006044. [CrossRef]

16. Olefeldt, D.; Goswami, S.; Grosse, G.; Hayes, D.; Hugelius, G.; Kuhry, P.; McGuire, A.D.; Romanovsky, V.E.; Sannel, A.B.K.; Schuur, E.A.G.; et al. Circumpolar Distribution and Carbon Storage of Thermokarst Landscapes. *Nat. Commun.* **2016**, *7*, 13043. [CrossRef] [PubMed]

17. Rodenhizer, H.; Belshe, F.; Celis, G.; Ledman, J.; Mauritz, M.; Goetz, S.; Sankey, T.; Schuur, E.A.G. Abrupt Permafrost Thaw Accelerates Carbon Dioxide and Methane Release at a Tussock Tundra Site. *Arct. Antarct. Alp. Res.* **2022**, *54*, 443–464. [CrossRef]

18. Philipp, M.; Dietz, A.; Buchelt, S.; Kuenzer, C. Trends in Satellite Earth Observation for Permafrost Related Analyses—A Review. *Remote Sens.* **2021**, *13*, 1217. [CrossRef]

19. 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.; et al. Carbon Release through Abrupt Permafrost Thaw. *Nat. Geosci.* **2020**, *13*, 138–143. [CrossRef]

20. Nelson, F.E.; Anisimov, O.A.; Shiklomanov, N.I. Subsidence Risk from Thawing Permafrost. *Nature* **2001**, *410*, 889–890. [CrossRef] [PubMed]

21. Jorgenson, M.T.; Shur, Y.L.; Pullman, E.R. Abrupt Increase in Permafrost Degradation in Arctic Alaska. *Geophys. Res. Lett.* **2006**, *33*, L02503. [CrossRef]

22. Abbott, B.W.; Jones, J.B. Permafrost Collapse Alters Soil Carbon Stocks, Respiration, CH_4, and N_2O in Upland Tundra. *Glob. Change Biol.* **2015**, *21*, 4570–4587. [CrossRef]

23. Cassidy, A.E.; Christen, A.; Henry, G.H.R. Impacts of Active Retrogressive Thaw Slumps on Vegetation, Soil, and Net Ecosystem Exchange of Carbon Dioxide in the Canadian High Arctic. *Arct. Sci.* **2017**, *3*, 179–202. [CrossRef]

24. Cassidy, A.E.; Christen, A.; Henry, G.H.R. The Effect of a Permafrost Disturbance on Growing-Season Carbon-Dioxide Fluxes in a High Arctic Tundra Ecosystem. *Biogeosciences* **2016**, *13*, 2291–2303. [CrossRef]

25. Hughes-Allen, L.; Bouchard, F.; Laurion, I.; Séjourné, A.; Marlin, C.; Hatté, C.; Costard, F.; Fedorov, A.; Desyatkin, A. Seasonal Patterns in Greenhouse Gas Emissions from Thermokarst Lakes in Central Yakutia (Eastern Siberia). *Limnol. Oceanogr.* **2021**, *66*, S98–S116. [CrossRef]

26. Jensen, A.E.; Lohse, K.A.; Crosby, B.T.; Mora, C.I. Variations in Soil Carbon Dioxide Efflux across a Thaw Slump Chronosequence in Northwestern Alaska. *Environ. Res. Lett.* **2014**, *9*, 025001. [CrossRef]

27. Rodenhizer, H.; Natali, S.M.; Mauritz, M.; Taylor, M.A.; Celis, G.; Kadej, S.; Kelley, A.K.; Lathrop, E.R.; Ledman, J.; Pegoraro, E.F.; et al. Abrupt Permafrost Thaw Drives Spatially Heterogeneous Soil Moisture and Carbon Dioxide Fluxes in Upland Tundra. *Glob. Change Biol.* **2023**, *29*, 6286–6302. [CrossRef] [PubMed]

28. Abbott, B.W.; Jones, J.B.; Godsey, S.E.; Larouche, J.R.; Bowden, W.B. Patterns and Persistence of Hydrologic Carbon and Nutrient Export from Collapsing Upland Permafrost. *Biogeosciences* **2015**, *12*, 3725–3740. [CrossRef]

29. Kokelj, S.V.; Kokoszka, J.; van der Sluijs, J.; Rudy, A.C.A.; Tunnicliffe, J.; Shakil, S.; Tank, S.E.; Zolkos, S. Thaw-Driven Mass Wasting Couples Slopes with Downstream Systems, and Effects Propagate through Arctic Drainage Networks. *Cryosphere* **2021**, *15*, 3059–3081. [CrossRef]

30. Schuur, E.A.G.; Mack, M.C. Ecological Response to Permafrost Thaw and Consequences for Local and Global Ecosystem Services. *Annu. Rev. Ecol. Evol. Syst.* **2018**, *49*, 279–301. [CrossRef]

31. Farquharson, L.M.; Romanovsky, V.E.; Cable, W.L.; Walker, D.A.; Kokelj, S.V.; Nicolsky, D. Climate Change Drives Widespread and Rapid Thermokarst Development in Very Cold Permafrost in the Canadian High Arctic. *Geophys. Res. Lett.* **2019**, *46*, 6681–6689. [CrossRef]

32. Farquharson, L.M.; Mann, D.H.; Grosse, G.; Jones, B.M.; Romanovsky, V.E. Spatial Distribution of Thermokarst Terrain in Arctic Alaska. *Geomorphology* **2016**, *273*, 116–133. [CrossRef]

33. Bernhard, P.; Zwieback, S.; Bergner, N.; Hajnsek, I. Assessing Volumetric Change Distributions and Scaling Relations of Retrogressive Thaw Slumps across the Arctic. *Cryosphere* **2022**, *16*, 1–15. [CrossRef]

34. Jones, B.M.; Grosse, G.; Arp, C.D.; Miller, E.; Liu, L.; Hayes, D.J.; Larsen, C.F. Recent Arctic Tundra Fire Initiates Widespread Thermokarst Development. *Sci. Rep.* **2015**, *5*, 15865. [CrossRef]

35. Jorgenson, M.T.; Douglas, T.A.; Liljedahl, A.K.; Roth, J.E.; Cater, T.C.; Davis, W.A.; Frost, G.V.; Miller, P.F.; Racine, C.H. The Roles of Climate Extremes, Ecological Succession, and Hydrology in Repeated Permafrost Aggradation and Degradation in Fens on the Tanana Flats, Alaska. *J. Geophys. Res. Biogeosci.* **2020**, *125*, e2020JG005824. [CrossRef]

36. Nitze, I.; Cooley, S.W.; Duguay, C.R.; Jones, B.M.; Grosse, G. The Catastrophic Thermokarst Lake Drainage Events of 2018 in Northwestern Alaska: Fast-Forward into the Future. *Cryosphere* **2020**, *14*, 4279–4297. [CrossRef]

37. Rowland, J.C.; Coon, E.T. From Documentation to Prediction: Raising the Bar for Thermokarst Research. *Hydrogeol. J.* **2016**, *24*, 645–648. [CrossRef]

38. Treharne, R.; Rogers, B.M.; Gasser, T.; MacDonald, E.; Natali, S. Identifying Barriers to Estimating Carbon Release From Interacting Feedbacks in a Warming Arctic. *Front. Clim.* **2022**, *3*, 716464. [CrossRef]

39. Nitze, I.; Heidler, K.; Barth, S.; Grosse, G. Developing and Testing a Deep Learning Approach for Mapping Retrogressive Thaw Slumps. *Remote Sens.* **2021**, *13*, 4294. [CrossRef]

40. Kokelj, S.V.; Jorgenson, M.T. Advances in Thermokarst Research: Recent Advances in Research Investigating Thermokarst Processes. *Permafr. Periglac. Process.* **2013**, *24*, 108–119. [CrossRef]

41. Lantuit, H.; Pollard, W.H. Fifty Years of Coastal Erosion and Retrogressive Thaw Slump Activity on Herschel Island, Southern Beaufort Sea, Yukon Territory, Canada. *Geomorphology* **2008**, *95*, 84–102. [CrossRef]

42. Lantuit, H.; Pollard, W.H. Temporal Stereophotogrammetric Analysis of Retrogressive Thaw Slumps on Herschel Island, Yukon Territory. *Nat. Hazards Earth Syst. Sci.* **2005**, *5*, 413–423. [CrossRef]

43. Swanson, D.; Nolan, M. Growth of Retrogressive Thaw Slumps in the Noatak Valley, Alaska, 2010–2016, Measured by Airborne Photogrammetry. *Remote Sens.* **2018**, *10*, 983. [CrossRef]

44. French, H.M. Active Thermokarst Processes, Eastern Banks Island, Western Canadian Arctic. *Can. J. Earth Sci.* **1974**, *11*, 785–794. [CrossRef]

45. Lantz, T.C.; Kokelj, S.V. Increasing Rates of Retrogressive Thaw Slump Activity in the Mackenzie Delta Region, N.W.T., Canada. *Geophys. Res. Lett.* **2008**, *35*, L06502. [CrossRef]

46. Lewkowicz, A.G. Headwall Retreat of Ground-Ice Slumps, Banks Island, Northwest Territories. *Can. J. Earth Sci.* **1987**, *24*, 1077–1085. [CrossRef]

47. Burn, C.R. The Thermal Regime of a Retrogressive Thaw Slump near Mayo, Yukon Territory. *Can. J. Earth Sci.* **2000**, *37*, 967–981. [CrossRef]

48. Liu, Y.; Qiu, H.; Kamp, U.; Wang, N.; Wang, J.; Huang, C.; Tang, B. Higher Temperature Sensitivity of Retrogressive Thaw Slump Activity in the Arctic Compared to the Third Pole. *Sci. Total Environ.* **2024**, *914*, 170007. [CrossRef] [PubMed]

49. Yang, D.; Qiu, H.; Ye, B.; Liu, Y.; Zhang, J.; Zhu, Y. Distribution and Recurrence of Warming-Induced Retrogressive Thaw Slumps on the Central Qinghai-Tibet Plateau. *J. Geophys. Res. Earth Surf.* **2023**, *128*, e2022JF007047. [CrossRef]

50. Lewkowicz, A.G.; Way, R.G. Extremes of Summer Climate Trigger Thousands of Thermokarst Landslides in a High Arctic Environment. *Nat. Commun.* **2019**, *10*, 1329. [CrossRef] [PubMed]

51. 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.; et al. Applications of ArcticDEM for Measuring Volcanic Dynamics, Landslides, Retrogressive Thaw Slumps, Snowdrifts, and Vegetation Heights. *Sci. Remote Sens.* **2024**, *9*, 100130. [CrossRef]

52. Niu, F.; Luo, J.; Lin, Z.; Fang, J.; Liu, M. Thaw-Induced Slope Failures and Stability Analyses in Permafrost Regions of the Qinghai-Tibet Plateau, China. *Landslides* **2016**, *13*, 55–65. [CrossRef]

53. Balser, A.W.; Jones, J.B.; Gens, R. Timing of Retrogressive Thaw Slump Initiation in the Noatak Basin, Northwest Alaska, USA. *J. Geophys. Res. Earth Surf.* **2014**, *119*, 1106–1120. [CrossRef]

54. Kokelj, S.V.; Tunnicliffe, J.; Lacelle, D.; Lantz, T.C.; Chin, K.S.; Fraser, R. Increased Precipitation Drives Mega Slump Development and Destabilization of Ice-Rich Permafrost Terrain, Northwestern Canada. *Glob. Planet. Change* **2015**, *129*, 56–68. [CrossRef]

55. Luo, J.; Niu, F.; Lin, Z.; Liu, M.; Yin, G.; Gao, Z. Inventory and Frequency of Retrogressive Thaw Slumps in Permafrost Region of the Qinghai–Tibet Plateau. *Geophys. Res. Lett.* **2022**, *49*, e2022GL099829. [CrossRef]

56. Segal, R.A.; Lantz, T.C.; Kokelj, S.V. Acceleration of Thaw Slump Activity in Glaciated Landscapes of the Western Canadian Arctic. *Environ. Res. Lett.* **2016**, *11*, 034025. [CrossRef]

57. Swanson, D.K. Permafrost Thaw-related Slope Failures in Alaska's Arctic National Parks, *c*. 1980–2019. *Permafr. Periglac. Process.* **2021**, *32*, 392–406. [CrossRef]

58. Ward Jones, M.K.; Pollard, W.H.; Jones, B.M. Rapid Initialization of Retrogressive Thaw Slumps in the Canadian High Arctic and Their Response to Climate and Terrain Factors. *Environ. Res. Lett.* **2019**, *14*, 055006. [CrossRef]

59. Yin, G.; Luo, J.; Niu, F.; Liu, M.; Gao, Z.; Dong, T.; Ni, W. High-Resolution Assessment of Retrogressive Thaw Slump Susceptibility in the Qinghai-Tibet Engineering Corridor. *Res. Cold Arid Reg.* **2023**, *15*, 288–294. [CrossRef]

60. Huang, L.; Lantz, T.C.; Fraser, R.H.; Tiampo, K.F.; Willis, M.J.; Schaefer, K. Accuracy, Efficiency, and Transferability of a Deep Learning Model for Mapping Retrogressive Thaw Slumps across the Canadian Arctic. *Remote Sens.* **2022**, *14*, 2747. [CrossRef]

61. Huang, L.; Luo, J.; Lin, Z.; Niu, F.; Liu, L. Using Deep Learning to Map Retrogressive Thaw Slumps in the Beiluhe Region (Tibetan Plateau) from CubeSat Images. *Remote Sens. Environ.* **2020**, *237*, 111534. [CrossRef]

62. Witharana, C.; Udawalpola, M.R.; Liljedahl, A.K.; Jones, M.K.W.; Jones, B.M.; Hasan, A.; Joshi, D.; Manos, E. Automated Detection of Retrogressive Thaw Slumps in the High Arctic Using High-Resolution Satellite Imagery. *Remote Sens.* **2022**, *14*, 4132. [CrossRef]

63. Yang, Y.; Rogers, B.M.; Fiske, G.; Watts, J.; Potter, S.; Windholz, T.; Mullen, A.; Nitze, I.; Natali, S.M. Mapping Retrogressive Thaw Slumps Using Deep Neural Networks. *Remote Sens. Environ.* **2023**, *288*, 113495. [CrossRef]

64. Runge, A.; Nitze, I.; Grosse, G. Remote Sensing Annual Dynamics of Rapid Permafrost Thaw Disturbances with LandTrendr. *Remote Sens. Environ.* **2022**, *268*, 112752. [CrossRef]

65. Lin, Y.; Knudby, A.J. A Transfer Learning Approach for Automatic Mapping of Retrogressive Thaw Slumps (RTSs) in the Western Canadian Arctic. *Int. J. Remote Sens.* **2023**, *44*, 2039–2063. [CrossRef]

66. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]

67. Planet Team. Planet Application Program Interface. In *Space for Life on Earth*; Planet Team: San Francisco, CA, USA, 2017.

68. Kim, M.; Park, S.; Anderson, C.; Stensaas, G.L. *System Characterization Report on Planet's SuperDove*; System Characterization of Earth Observation Sensors; U.S. Department of the Interior and U.S. Geological Survey: Renton, VA, USA, 2022.

69. Scheffler, D.; Hollstein, A.; Diedrich, H.; Segl, K.; Hostert, P. AROSICS: An Automated and Robust Open-Source Image Co-Registration Software for Multi-Sensor Satellite Data. *Remote Sens.* **2017**, *9*, 676. [CrossRef]

70. Nielsen, A.A. The Regularized Iteratively Reweighted MAD Method for Change Detection in Multi- and Hyperspectral Data. *IEEE Trans. Image Process.* **2007**, *16*, 463–478. [CrossRef] [PubMed]

71. Nielsen, A.A.; Conradsen, K.; Simpson, J.J. Multivariate Alteration Detection (MAD) and MAF Postprocessing in Multispectral, Bitemporal Image Data: New Approaches to Change Detection Studies. *Remote Sens. Environ.* **1998**, *64*, 1–19. [CrossRef]

72. Porter, C.; Morin, P.; Howat, I.; Noh, M.-J.; Bates, B.; Peterman, K.; Keesey, S.; Schlenk, M.; Gardiner, J.; Tomko, K.; et al. ArcticDEM 2018, Version 3; Harvard Dataverse, V1. Available online: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/OHHUKH (accessed on 15 June 2024).

73. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019.

74. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* **2016**, arXiv:1603.04467.

75. R Core Team R: A Language and Environment for Statistical Computing 2023. R Foundation for Statistical Computing, Vienna, Austria. Available online: https://www.R-project.org/ (accessed on 15 June 2024).

76. Hijmans, R.J. Terra: Spatial Data Analysis; Version: 1.7-29. 2023; Available online: https://CRAN.R-project.org/package=terra (accessed on 15 June 2024).

77. Pebesma, E.; Bivand, R. *Spatial Data Science: With Applications in R*, 1st ed.; Chapman and Hall: London, UK; CRC: Boca Raton, FL, USA, 2023.

78. Pebesma, E. Simple Features for R: Standardized Support for Spatial Vector Data. *R J.* **2018**, *10*, 439. [CrossRef]

79. Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; et al. Welcome to the Tidyverse. *J. Open Source Softw.* **2019**, *4*, 1686. [CrossRef]

80. Polsby, D.D.; Popper, R.D. The Third Criterion: Compactness as a Procedural Safeguard against Partisan Gerrymandering. *Yale Law Policy Rev.* **1991**, *9*, 301. [CrossRef]

81. Huang, C.; Townshend, J.R.G.; Liang, S.; Kalluri, S.N.V.; DeFries, R.S. Impact of Sensor's Point Spread Function on Land Cover Characterization: Assessment and Deconvolution. *Remote Sens. Environ.* **2002**, *80*, 203–212. [CrossRef]

82. Zhang, L.; Zhang, L. Artificial Intelligence for Remote Sensing Data Analysis: A Review of Challenges and Opportunities. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 270–294. [CrossRef]

83. Burn, C.R.; Lewkowicz, A.G. Retrogressive Thaw Slumps. *Can. Geogr. Géographies Can.* **1990**, *34*, 273–276. [CrossRef]

84. Porter, C.; Howat, I.; Noh, M.-J.; Husby, E.; Khuvis, S.; Danish, E.; Tomko, K.; Gardiner, J.; Negrete, A.; Yadav, B.; et al. ArcticDEM, Version 4.1; Harvard Dataverse, V1. 2023. Available online: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/3VDC4W (accessed on 15 June 2024).

85. Nitze, I.; Nesterova, N.; Runge, A.; Veremeeva, A.; Jones, M.W.; Witharana, C.; Xia, Z.; Liljedahl, A.K. A Labeling Intercomparison Experiment of Retrogressive Thaw Slumps by a Diverse Group of Domain Experts. *J-STARS*, 2024; *submitted*.

86. Yang, Y.; Rodenhizer, H.; Rogers, B.M.; Dean, J.; Singh, R.; Windholz, T.; Poston, A.; Zolkos, S.; Fiske, G.; Mullen, A.; et al. ARTS: A Scalable Data Set for Arctic Retrogressive Thaw Slumps. In Proceedings of the EGU General Assembly 2024, Vienna, Austria, 14–19 April 2024.

87.   Huang, L.; Willis, M.J.; Li, G.; Lantz, T.C.; Schaefer, K.; Wig, E.; Cao, G.; Tiampo, K.F. Identifying Active Retrogressive Thaw Slumps from ArcticDEM. *ISPRS J. Photogramm. Remote Sens.* **2023**, *205*, 301–316. [CrossRef]

88.   Kennedy, R.E.; Yang, Z.; Cohen, W.B. Detecting Trends in Forest Disturbance and Recovery Using Yearly Landsat Time Series: 1. LandTrendr—Temporal Segmentation Algorithms. *Remote Sens. Environ.* **2010**, *114*, 2897–2910. [CrossRef]

89.   Benecki, P.; Kawulok, M.; Kostrzewa, D.; Skonieczny, L. Evaluating Super-Resolution Reconstruction of Satellite Images. *Acta Astronaut.* **2018**, *153*, 15–25. [CrossRef]

90.   Lv, Z.; Jia, Y.; Zhang, Q. Joint Image Registration and Point Spread Function Estimation for the Super-Resolution of Satellite Images. *Signal Process. Image Commun.* **2017**, *58*, 199–211. [CrossRef]