

Supplemental Methods and Results for:

Permafrost thaw follows spatially organized, non-linear trajectories in ice-wedge terrain

1. Data Input Tests:

We examined how the post-processing of the satellite imagery impacted U-Net classifier accuracy by running training tests. We trained the U-Net model using the same set of annotated data but with different images processed using different pan-sharpening algorithms. For each type of processed imagery, we trained the U-Net model five times to account for stochasticity associated with random selection of training and validation tiles. Each trained U-Net model was trained with 40 tiles and validated with 10 tiles that had been randomly withheld from model training. We compared the accuracy of the resulting trained U-Net by examining two accuracy metrics, the pixel accuracy value and the mean Intersection-over-Union, using an ANOVA test.

Pixel accuracy is the proportion of pixels in the tile that were accurately identified. Intersection-over-union (IoU) is the proportion of overlap between the predicted pixels and the reference pixels. When the feature of interest (i.e., ice-wedge troughs) are relatively small compared to the input image, pixel accuracy is often high when the model performs well at identifying the non-feature of interest regardless of how well the model performs at identifying the feature of interest; because IoU focuses specifically on the feature of interest, it better represents how well the model performs at identifying that class.

We compared single-band panchromatic imagery with four-band, pan-sharpened multispectral imagery (bands: R, G, B, NIR) produced using five different pan-sharpening algorithms: Brovey, ESRI, Gram-Schmidt, IHS, and Mean. The pan-sharpening process was conducted in ArcGIS Pro using the Pan-Sharpener raster function.

These data input tests were conducted using binary training data (trough/no-trough) to focus specifically on the impact of differing imagery on trough detection ability. Accuracy metrics associated with all the data input tests are reported in Table S1 and visualized in Figure S1. While there was no significant difference between accuracy metrics as determined by an ANOVA test, the IHS and Mean pan-sharpening method produced the highest values for both pixel accuracy and IoU. The IHS method was chosen because the pan-sharpened images it produced contained more visible differentiation in color compared to the Mean method.

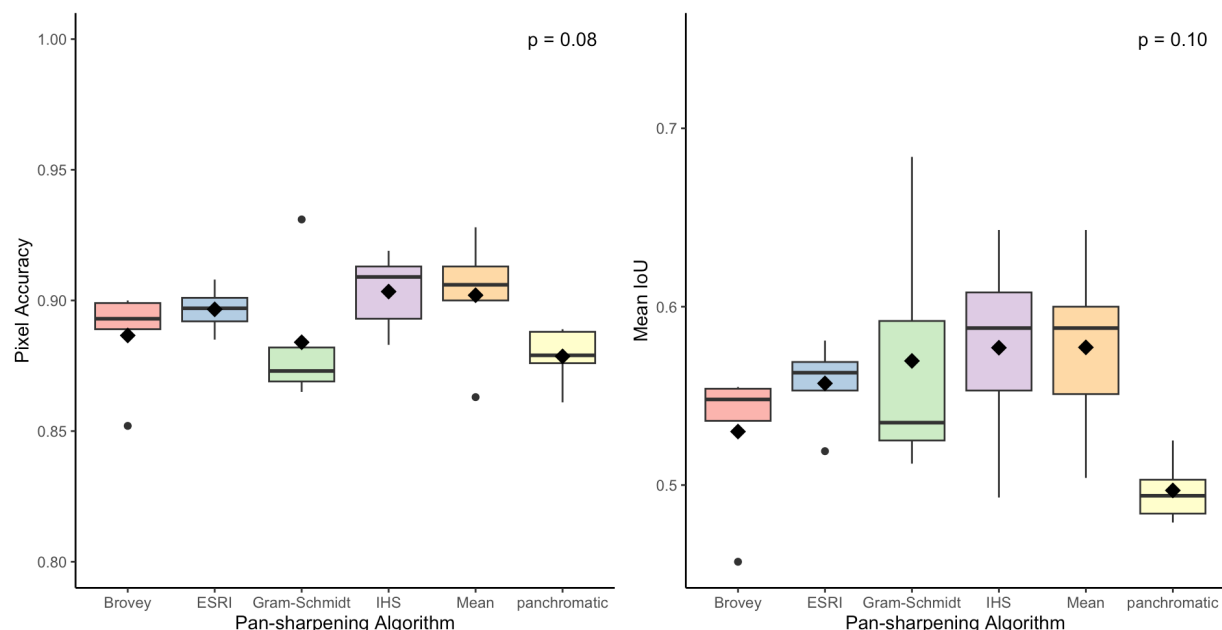


Figure S1: Comparison of accuracy metrics (left: pixel accuracy; right: mean IoU) for four-band, multispectral imagery created using different pan-sharpening methods and the original single-band panchromatic image. P-value calculated with the ANOVA test reported in the top-right corner of each box-plot. n=5 trained models per algorithm.

Table S1: Accuracy values for two metrics, pixel accuracy and mean Intersection-over-Union, all computed with n=5 trained models to account for stochasticity in model training.

Algorithm	Pixel Accuracy			Mean Intersection-over-Union		
	Mean	Median	St. Dev.	Mean	Median	St. Dev.
<i>Brovey</i>	0.887	0.893	0.020	0.530	0.548	0.042
<i>ESRI</i>	0.897	0.897	0.009	0.557	0.563	0.024
<i>Gram-Schmidt</i>	0.884	0.873	0.027	0.570	0.535	0.071
<i>IHS</i>	0.903	0.909	0.015	0.577	0.588	0.057
<i>mean</i>	0.902	0.906	0.024	0.577	0.588	0.052
<i>Panchromatic</i>	0.879	0.879	0.011	0.497	0.494	0.018

Table S2: Satellite images and trough detection accuracy metrics from the final trained U-Net used to produce trough maps. Px Acc: Pixel Accuracy, mIoU: mean Intersection-over-Union.

Date	Px Acc. (%)	mIoU: dry trough	mIoU: flooded trough	Dice: dry trough	Dice: flooded trough
08-15-2006	93.6	0.347	0.467	0.515	0.636
07-02-2011	90.3	0.352	0.699	0.520	0.823
07-07-2014	91.6	0.530	0.618	0.693	0.764
08-05-2023	91.7	0.420	0.602	0.591	0.751

2. Degradation Stage Simulations

We tested seven trough width thresholds and seven trough flooding percentage thresholds to determine what threshold value to use for defining hydrogeomorphic degradation stages. Tested width thresholds were 1m, 1.25m, 1.5m, 1.75m, 2m, 2.25m, and 2.5m (Figure S2). Tested trough flooding percentage thresholds were 2.5%, 5%, 7.5%, 10%, 12.5%, 15%, and 17.5% (Figure S3, S4).

The selection of the final threshold values used to classify degradation stages was informed by simulation results as well as field observations. The 1.5 m trough width threshold was selected to distinguish the undegraded stage, as smaller thresholds failed to capture areas that visibly appeared undegraded in the imagery (Figure S2), while field observation indicated that troughs greater than 1.5 m in width were rarely undegraded. This width threshold is close to previously published remote sensing studies of ice-wedge degradation (Braun and Andresen, 2024; Wickland et al., 2020).

The 7.5% threshold was used to distinguish the initial and advanced stages of degradation. This selected threshold visually appeared appropriate for dividing the initial and advanced stages of degradation in comparison to field observations, and the simulation results indicated that changing the threshold value to either 5% or 10% produced very little change in the overall proportion of the site that was classified as initial or advanced degradation (Figure S7, S8).

Overall, the similarity in spatial patterns and overall trends in the amount of the site classified as each stage through time increased our confidence that small shifts in threshold values would not produce significant impacts on results (Figure S5-S8).

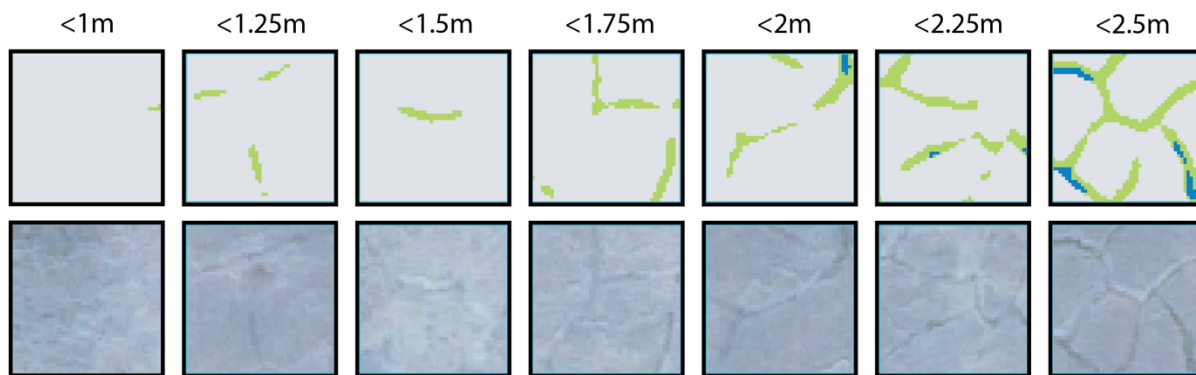


Figure S2: Examples of mapped ice-wedge troughs (top) and satellite images (bottom) for each tested width threshold (columns labeled by upper limit of width threshold; cells with an average trough width less than the threshold are assigned the “Undegraded” stage). Note that the trough width thresholds of 2m and greater include cells that contain water, a clear indicator of degradation that should not be present in the undegraded cells.

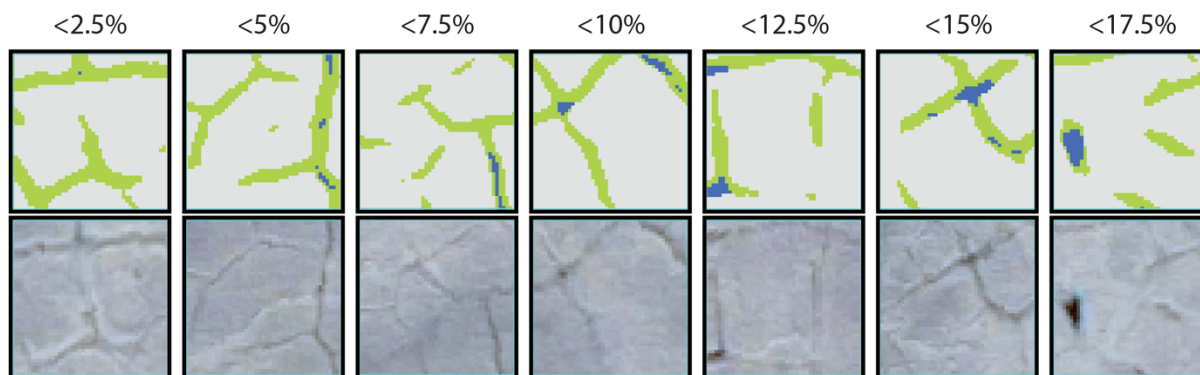


Figure S3: Examples of mapped ice-wedge troughs (top) and satellite images (bottom) for each tested flooding percentage threshold (columns labeled by upper limit of flooding percentage threshold; cells with a flooding percentage less than the threshold are assigned the “Initial Degradation” stage). Note that cells with average widths <1.5 m were excluded.

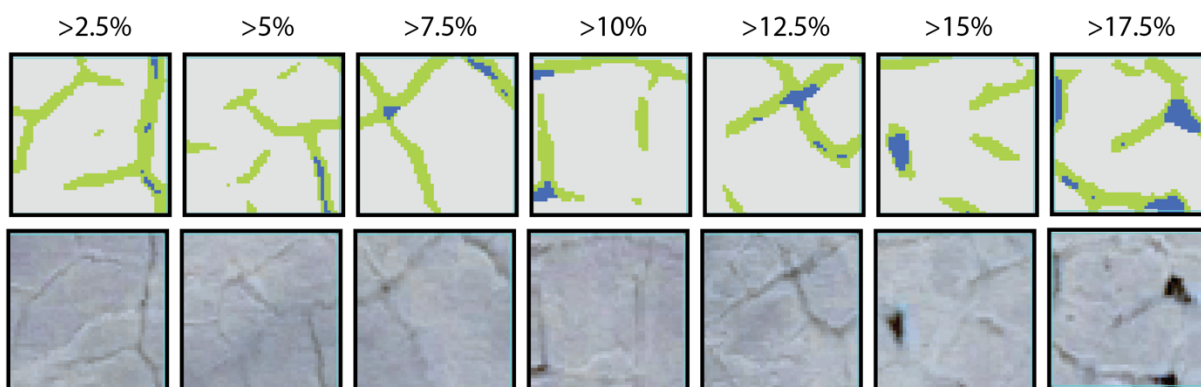


Figure S4: Examples of mapped ice-wedge troughs (top) and satellite images (bottom) for each tested flooding percentage threshold (columns labeled by upper limit of flooding percentage threshold; cells with a flooding percentage greater than the threshold are assigned the “Advanced Degradation” stage). Note that cells with average widths <1.5 m were excluded.

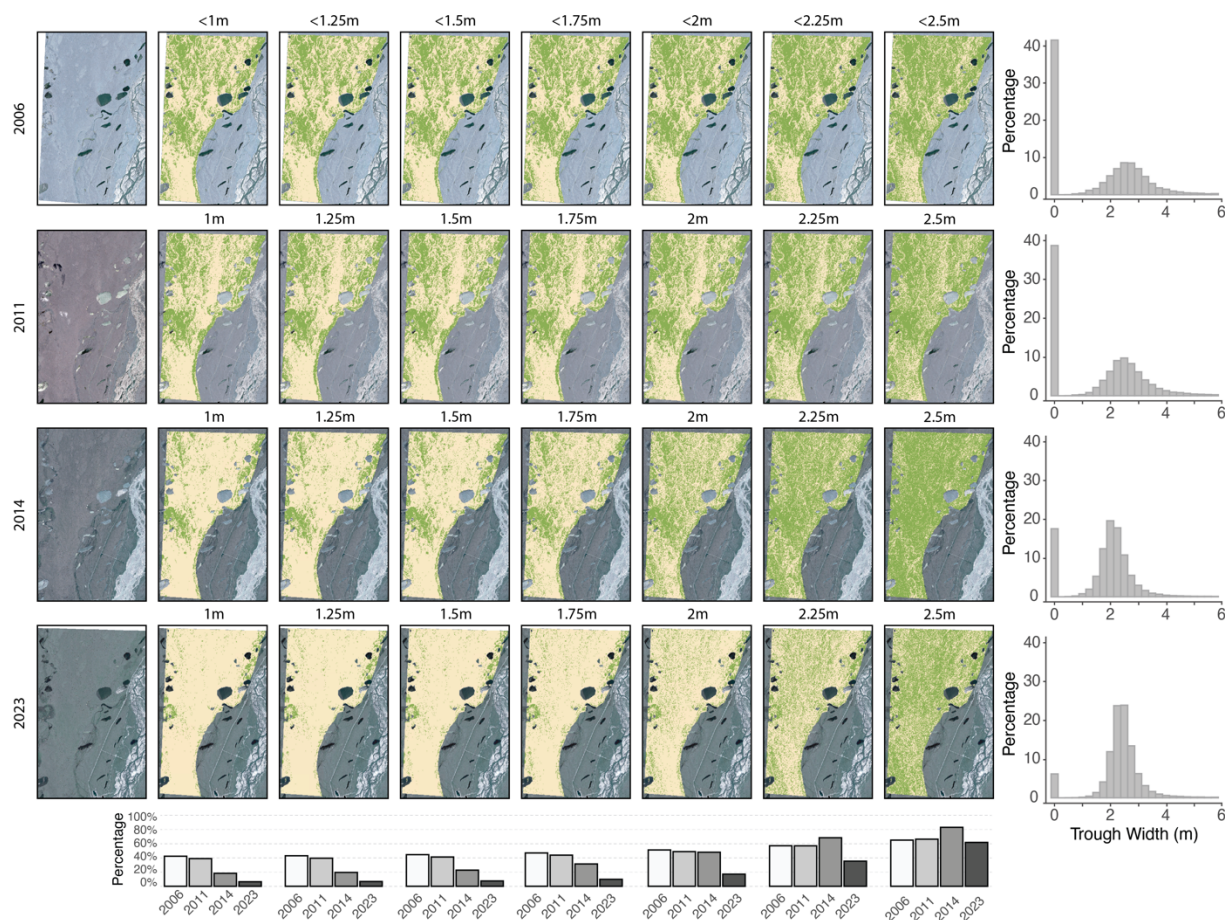


Figure S5: Examining the impact of the width threshold on mapping undegraded ice-wedge stages. Each row of the figure corresponds to one of the four years of analysis, while the columns correspond to the different trough width thresholds used to distinguish the undegraded stages. Portions of the site classified as "Undegraded" are mapped in green. The right-most column depicts the histogram of trough widths for each year. The bottom-most row indicates what portion of the site was classified as "Undegraded" per year using that column's width threshold. Cells that had a trough width of zero were classified as undegraded.

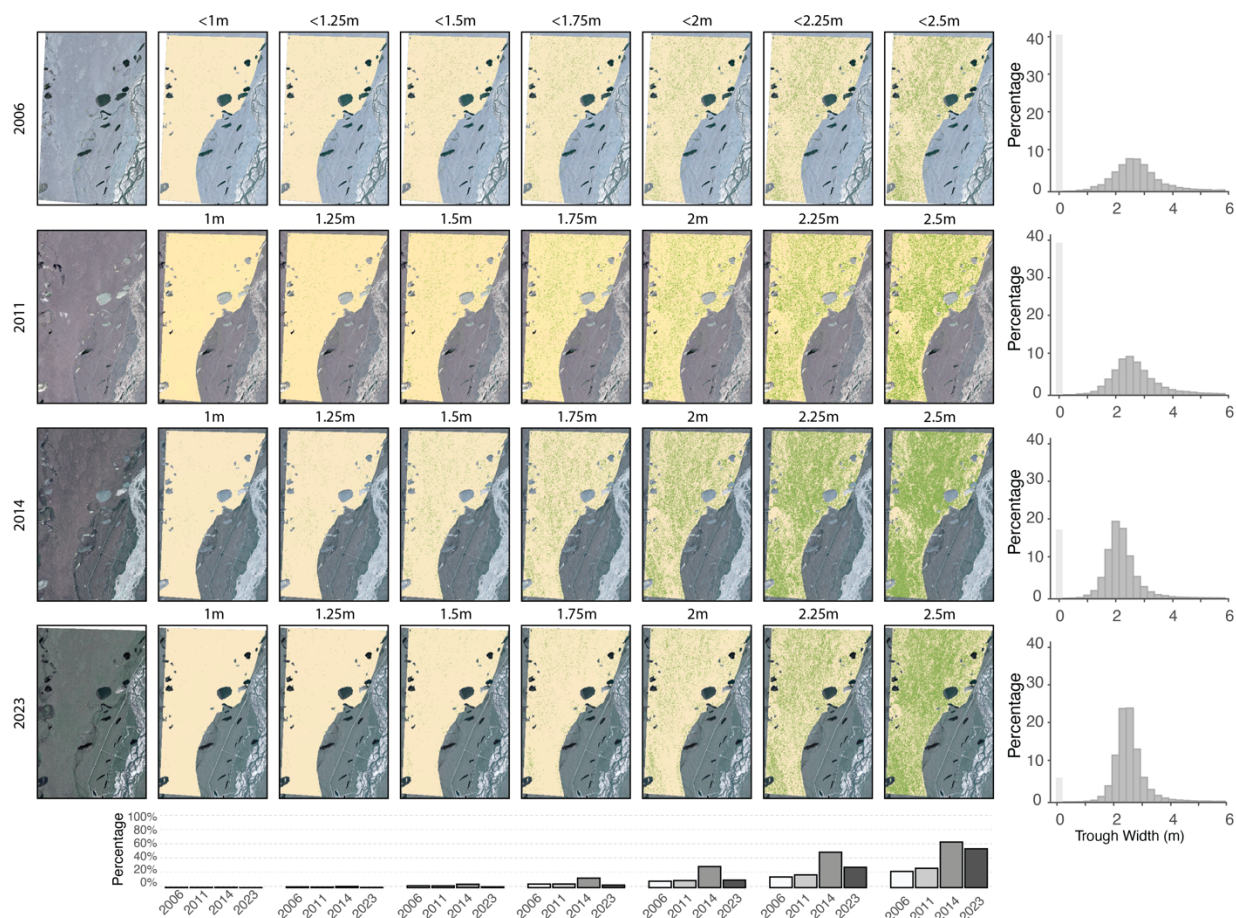


Figure S6: Examining the impact of the width threshold on mapping undegraded ice-wedge stages. Each row of the figure corresponds to one of the four years of analysis, while the columns correspond to the different trough width thresholds used to distinguish the undegraded stages. Portions of the site classified as “Undegraded” are mapped in green. The right-most column depicts the histogram of trough widths for each year. The bottom-most row indicates what portion of the site was classified as “Undegraded” per year using that column’s width threshold. Cells with a trough width of zero were not classified as undegraded.

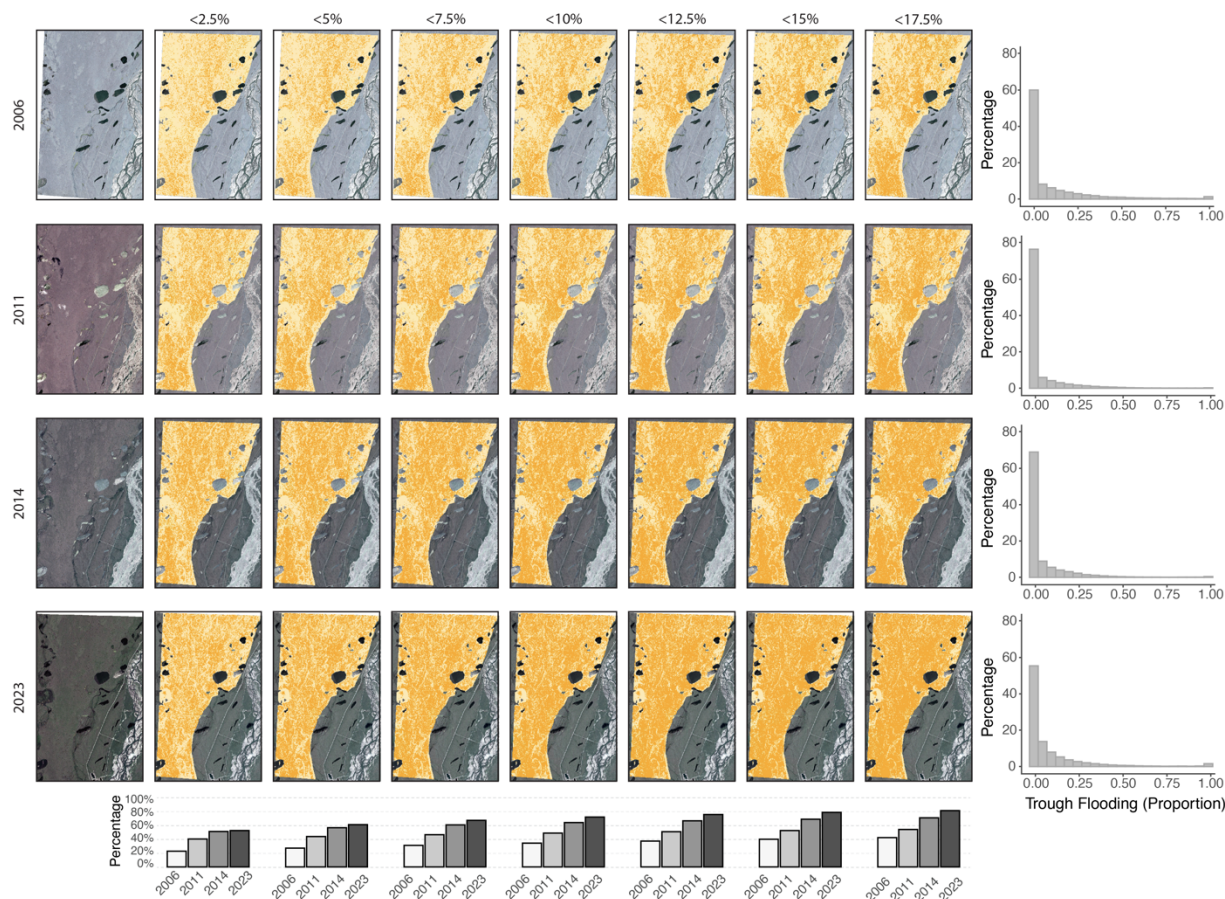


Figure S7: Examining the impact of the flooding percentage threshold on mapping initial degradation ice-wedge stages. Each row of the figure corresponds to one of the four years of analysis, while the columns correspond to the different flooding percentage thresholds used to distinguish the initial degradation stage. Portions of the site classified as "Initial Degradation" are mapped in orange. The right-most column depicts the histogram of flooding percentage for each year. The bottom-most row indicates what portion of the site was classified as "Initial Degradation" per year using that column's width threshold.

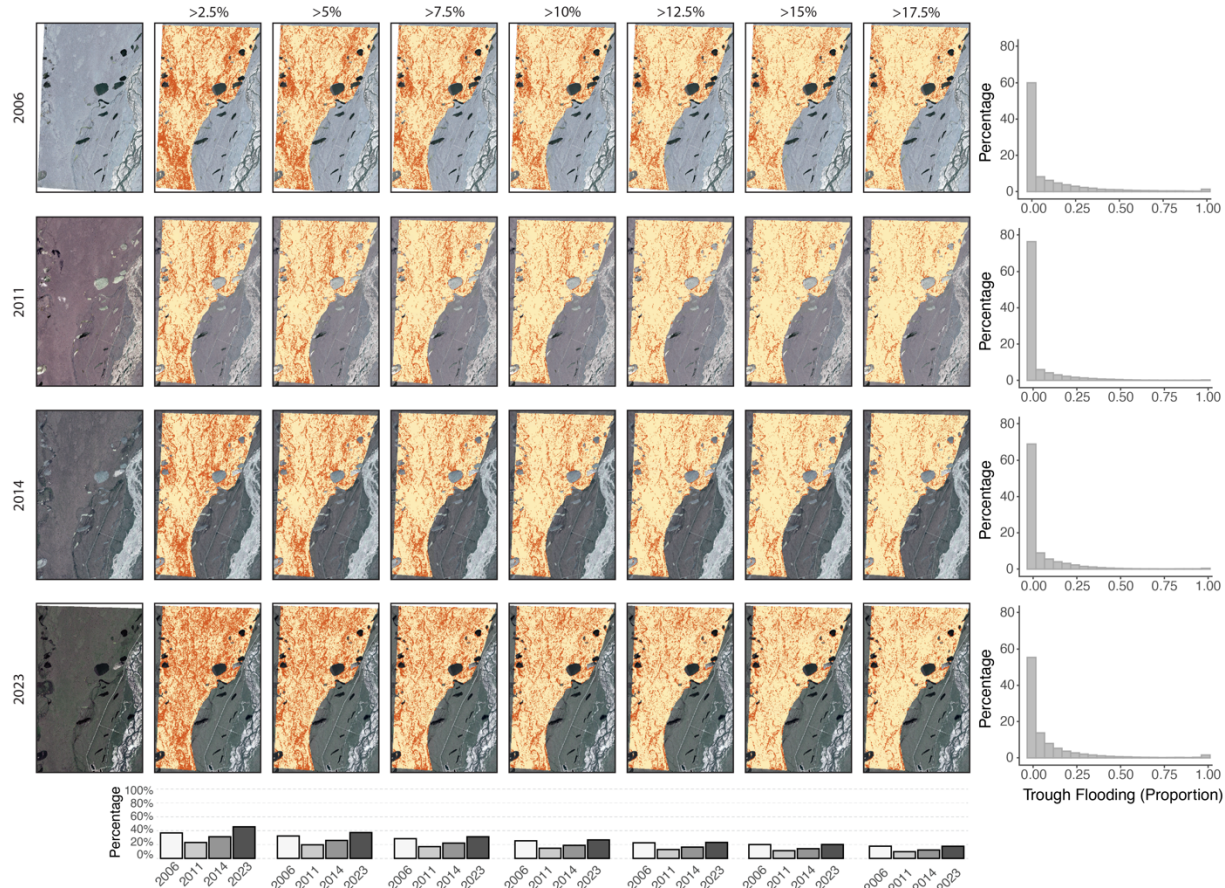


Figure S8: Examining the impact of the flooding percentage threshold on mapping advanced degradation ice-wedge stages. Each row of the figure corresponds to one of the four years of analysis, while the columns correspond to the different flooding percentage thresholds used to distinguish the advanced degradation stage. Portions of the site classified as “Advanced Degradation” are mapped in red. The right-most column depicts the histogram of flooding percentage for each year. The bottom-most row indicates what portion of the site was classified as “Advanced Degradation” per year using that column’s width threshold.

3. Stage Sequence Sensitivity Tests

A small magnitude change in the thresholding metric can trigger a drastic change in hydrogeomorphic degradation stage classification. Such small magnitude changes might deserve to trigger a classification change, but it is also possible that such changes are the product of mapping inaccuracy or co-registration errors between images. We conducted stage sequence sensitivity tests to examine whether such subtle changes are influencing the spatiotemporal patterns we found in ice-wedge degradation stage.

We examined four sensitivities (Table S2). Two sensitivities were related to changes across the width and flooding percentage threshold values. The third and fourth sensitivities examined changes in the presence and absence of trough pixels.

The threshold sensitivities (1 and 2) only accepted a change in degradation stage if the associated shift in threshold metric was greater than 10% of the threshold value. The results of our sensitivity tests (Table S3, S4) show that enacting this sensitivity influenced ~0.1% of our site; this indicates that when a cell is transitioning from one stage to another, that transition is occurring because a substantial (>10% of the threshold value) change in trough metric is detected.

The trough presence sensitivity (3) only accepted a change in the presence or absence of troughs if the associated change in total trough pixels was greater than 1% of the area of the cell. This sensitivity impacted more of the site than the threshold sensitivities; ~5% of the site in 2011 and 2014, and 1.9% in 2023. This sensitivity test is intended to diminish the amount of classification changes triggered by noise in the maps.

The undegraded sensitivity (4) reassigns cells from a “no trough” classification to an “undegraded” classification. A trough detection at a different date indicates that a trough is present at that location and therefore that cell should still be categorized as “undegraded” even if the trough is not visible at that point in time. This sensitivity impacted the largest portion of the site, though the percentage impacted decreased over time from 35.9% in 2006 to 1.6% in 2023. This temporal trend aligns with our observation that the site was highly degrading through time; undegraded cells were transitioning to higher stages of degradation.

Table S3: Descriptions of the four sensitivity tests

Sensitivity	Description
<i>Sensitivity 1</i>	Only accepted changes in stage from <i>undegraded</i> $\diamond$ <i>initial degradation</i> or <i>advanced degradation</i> if the magnitude of change in trough width was greater than 0.15 m (i.e., 10% of the 1.5 m threshold)
<i>Sensitivity 2</i>	Only accepted changes in stage from <i>initial degradation</i> $\diamond$ <i>advanced degradation</i> if the magnitude of change in flooding percentage was greater than 0.75% (i.e., 10% of the 7.5% threshold)
<i>Sensitivity 3</i>	Only accepted changes in stage from <i>no trough</i> $\diamond$ <i>undegraded</i> , <i>initial degradation</i> , or <i>advanced degradation</i> if the magnitude of change in total trough area was greater than 1% of the cell size (9 m ²)
<i>Sensitivity 4</i>	Replaced <i>no trough</i> with <i>undegraded</i> if that cell had a stage assigned for any other year

Table S4: Sensitivity tests resulted in changed stages in the following number of cells:

Year	2006	2011	2014	2023
<i>Sensitivity 1</i>	n/a	25	62	27
<i>Sensitivity 2</i>	n/a	10	12	14
<i>Sensitivity 3</i>	n/a	2814	2992	1052
<i>Sensitivity 4</i>	19929	18093	9196	910

Table S5: Sensitivity tests resulted in changed stages in the following percentage of cells:

Year	2006	2011	2014	2023
<i>Sensitivity 1</i>	n/a	0.05%	0.11%	0.05%
<i>Sensitivity 2</i>	n/a	0.02%	0.02%	0.03%
<i>Sensitivity 3</i>	n/a	5.07%	5.39%	1.90%
<i>Sensitivity 4</i>	35.92%	32.61%	16.57%	1.64%

4. Assigning Trajectory Classes to Stage Sequences

Table S6: Assigning descriptive trajectory labels to sequences of stages for the overlap site. Stage sequence refers to the temporal sequence of hydrogeomorphology stage assigned to each cell, with the first digit corresponding to the stage assigned in 2006 and the final digit corresponding to the 2023 stage. 0 = No Troughs Present, 1 = Undegraded, 2 = Initial Degradation, 3 = Advanced Degradation. The Stage Sequence column contains the top 35 unique stage sequences in order of abundance at the overlap site (see count for number of cells with each stage sequence, and percent for percentage of the overlap stage with that sequence). The top 35 unique stage sequences comprise 95% of all cells at the overlap site. Trajectory is assigned based on interpreting the stage sequence; Year of Change is assigned for Trajectories that change through time and indicates the year when the pixel experienced the associated change (i.e., in which year the “Advanced Degradation” stage appeared for a “Advancing” trajectory).

Stage Sequence	Count	Percent	Trajectory	Year of Change
2222	9290	16.7	Initial Degradation	n/a
1112	6869	12.4	Initiating	2023
1122	5468	9.9	Initiating	2014
3333	5224	9.4	Advanced Degradation	n/a
1222	3524	6.4	Initiating	2011
3222	3016	5.4	Drying	2011
0000	2823	5.1	No Trough	n/a
2223	1695	3.1	Advancing	2023
3233	1282	2.3	Oscillating	n/a
3223	1223	2.2	Oscillating	n/a
1113	1144	2.1	Advancing	2023
2122	1098	2.0	Oscillating	n/a
2333	920	1.7	Advancing	2011
1123	786	1.4	Advancing	2023
3122	696	1.3	Oscillating	n/a
2233	669	1.2	Advancing	2014
1223	665	1.2	Advancing	2023
3332	546	1.0	Drying	2023
3232	473	0.9	Oscillating	n/a

3322	462	0.8	Drying	2014
1212	451	0.8	Oscillating	n/a
2212	441	0.8	Oscillating	n/a
3133	437	0.8	Oscillating	n/a
1111	394	0.7	Undegraded	n/a
3323	379	0.7	Oscillating	n/a
2232	376	0.7	Oscillating	n/a
1133	356	0.6	Advancing	2014
1333	322	0.6	Advancing	2011
2322	310	0.6	Oscillating	n/a
1121	269	0.5	Oscillating	n/a
2332	261	0.5	Oscillating	n/a
2112	259	0.5	Oscillating	n/a
3112	255	0.5	Oscillating	n/a
3113	243	0.4	Oscillating	n/a
2323	238	0.4	Oscillating	n/a

5. Tracking Metrics of Ice-Wedge Degradation

Each degradation metric: (a) trough width, (b) water area, (c) the percentage of trough area that is flooded, and (d) the total trough area, showed an overall increasing trend at the study site during the 17 years of the study period (Figure 4), indicating that degradation is intensifying over time. However, analysis of individual degradation trajectories reveals that landscape change is highly heterogeneous rather than spatially uniform. As expected, the magnitude and rate of change in degradation metrics varied among trajectories; for example, the Advancing trajectory exhibited the largest increases in water area while the Drying and Undegraded trajectories consistently showed the lowest metric values and slowest rates of change, as these are areas that by definition either experienced losing water area over time or have troughs that remained under 1.5 m in width throughout the study period. While trends are variable across the trajectories, by the end of the study period nearly all trajectories have positive rates of change across all metrics, and no portion of the site shows a steady decrease in metrics which would indicate widespread recovery from ice-wedge degradation. Ice-wedge degradation is heterogeneously intensifying across this site.

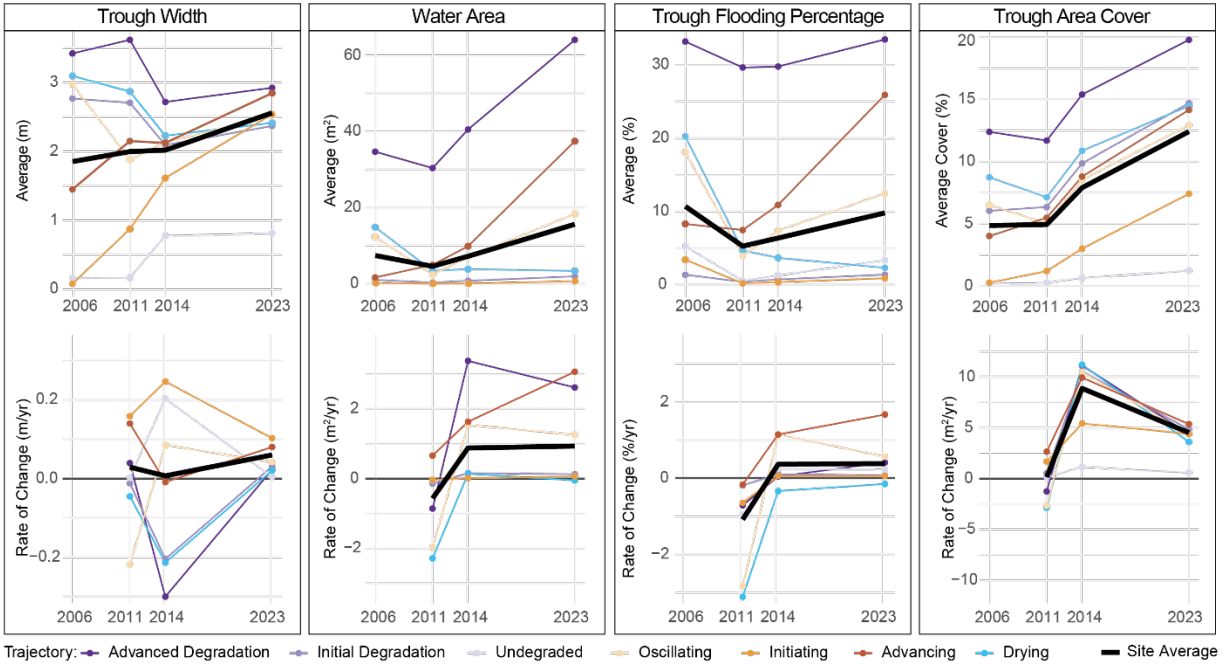


Figure S9: Ice-wedge trough degradation metrics over time, colored by degradation trajectory. Top: averaged metrics across trajectory classes with site-level average in black; Bottom: rates of change in metrics across trajectory classes with site-level average in black.

6. Landscape analyses: landforms and degradation trajectory patterns

Table S7: Contingency table for all combinations of Trajectories and Landforms, with areas in hectares for the observed class, expected class (based on an even distribution), and difference between observed and expected. Negative difference values indicate a certain class combination is underrepresented compared to an even distribution, while positive difference values indicate the class combination is overrepresented.

		Landform					
Degradation Trajectory		Aeolian	Channel	Fluvial	Lacustrine	Terrace	Grand Total
Advanced Degradation	Observed	3.6	186.9	24.7	14.7	240.3	470.2
	Expected	10.0	54.0	22.1	24.4	359.7	470.2
	Difference	-6.4	132.9	2.6	-9.7	-119.4	0.0
Advancing	Observed	6.7	100.6	29.2	35.5	418.2	590.1
	Expected	12.5	67.8	27.7	30.6	451.5	590.1
	Difference	-5.8	32.8	1.4	4.9	-33.3	0.0
Drying	Observed	3.5	48.2	4.9	11.0	294.6	362.2
	Expected	7.7	41.6	17.0	18.8	277.1	362.2
	Difference	-4.2	6.6	-12.2	-7.8	17.5	0.0
Initial Degradation	Observed	4.3	75.0	7.5	22.3	727.0	836.1
	Expected	17.7	96.1	39.3	43.3	639.7	836.1
	Difference	-13.4	-21.1	-31.8	-21.0	87.3	0.0
Initiating	Observed	53.8	35.4	65.6	35.7	1237.0	1427.5
	Expected	30.2	164.0	67.1	74.0	1092.2	1427.5
	Difference	23.6	-128.6	-1.5	-38.3	144.8	0.0
No Trajectory	Observed	4.3	12.2	27.2	37.4	154.7	235.8
	Expected	5.0	27.1	11.1	12.2	180.4	235.8
	Difference	-0.7	-14.8	16.1	25.1	-25.7	0.0
No Trough	Observed	12.7	1.4	38.8	54.0	147.2	254.1
	Expected	5.4	29.2	11.9	13.2	194.4	254.1
	Difference	7.3	-27.7	26.8	40.8	-47.2	0.0
Oscillating	Observed	15.7	113.7	32.8	44.1	576.0	782.2
	Expected	16.6	89.9	36.8	40.5	598.4	782.2
	Difference	-0.9	23.8	-4.0	3.6	-22.4	0.0
Undegraded	Observed	1.2	0.2	4.3	4.2	25.6	35.5
	Expected	0.8	4.1	1.7	1.8	27.1	35.5
	Difference	0.4	-3.9	2.7	2.4	-1.6	0.0
Grand Total:		105.75	573.66	234.81	258.84	3820.5	4993.6
Observed							

7. Topographic inversion of abandoned channel network

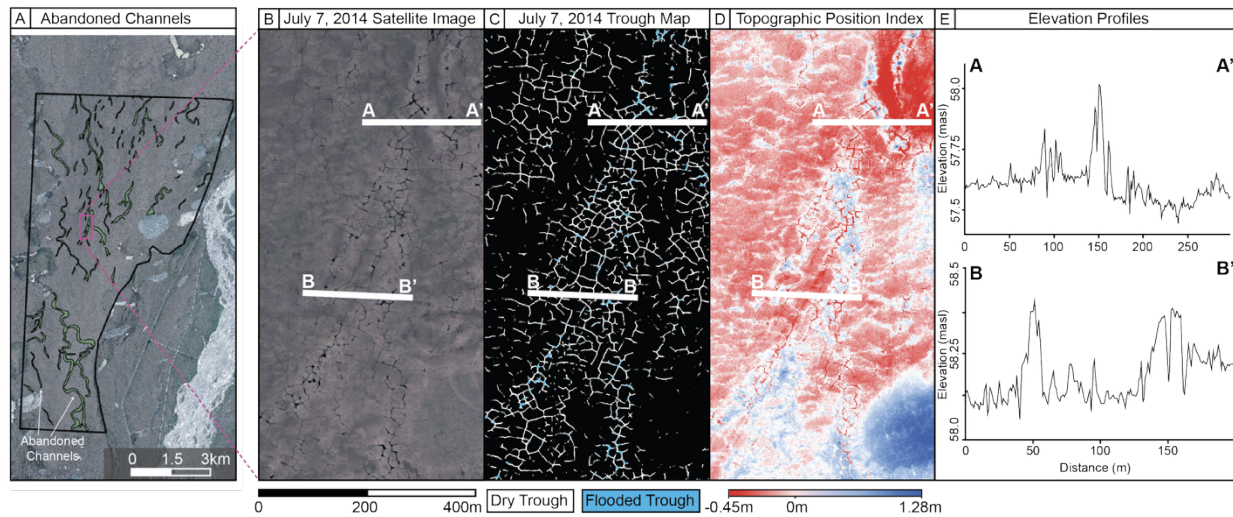


Figure S10: Relationship between degradation hot-spots, abandoned channels, and elevation. A: July 7, 2014 WorldView-2 satellite image (© Maxar) of study site with annotated locations of where degradation hot-spots form in abandoned channels. B: Satellite image from pink extent indicator at A, showing an example of degradation hot-spots (flooded troughs) arranged in linear, channel-like features. C: U-net mapped trough networks. D: Topographic position index showing higher elevation areas are associated with inverted channel-like degradation hot-spots. E: Elevation profiles along inverted channel features.

8. Improvements and remaining challenges detecting ice-wedge degradation

Accurate assessment of the trajectory of ice-wedge degradation requires confidence in the underlying maps of ice-wedge through networks. The deep learning image classifier's accuracy values (Table 1) are in line with other machine learning image classifiers in ice-wedge terrain (Abolt et al., 2019; Witharana et al., 2021; Zhang et al., 2018), and are a substantial improvement over traditional remote sensing approaches for detecting ice-wedge troughs (Braun and Andresen, 2024). Improved detection of ice-wedge troughs, particularly those that are non-flooded or captured in older imagery, enables robust classification of degradation trajectories and reveals patterns of landscape change that would otherwise remain obscured in traditional remote sensing analyses.

While trough identification is improved substantially with the U-Net image classifier used here, challenges detecting thaw remain, particularly related to the morphology of the features of interest and the inherent limitations of the imagery record. Ice-wedge troughs can be narrower than 0.5 m in width (Braun and Andresen, 2024; Dafflon et al., 2016; Kanevskiy et al., 2013; Kokelj et al., 2014), while the satellite images analyzed for this project range in resolution from 0.47-0.74 m. Narrow troughs are therefore very difficult to detect. Systematic under-classification of narrow troughs would produce a dearth in pixels classified as undegraded, which is alleviated by the undegraded stage sensitivity: pixels with no identified troughs were reassigned the undegraded stage if at another point in time we identified troughs at that location (Table S2). Changing image resolutions through time also complicates analysis of temporal trends in ice-wedge troughs. Imagery resolution increased from 0.74 m in 2011 to 0.50 m in 2014; at the same time, the maps detected a substantial increase in the amount of trough area in

1 2014 (Figure 6), and particularly in the amount of non-flooded trough area detected. This
2 increase in non-flooded trough is potentially the product of two factors: first, the initiation of
3 degradation leading to surface subsidence and trough widening without water accumulation, and
4 second, increased image resolution allowing detection of previously unidentified features.
5 Narrow and smaller troughs that were present in 2011 but undetected by the image classifier
6 could produce false increases in degradation when those troughs were detected in the higher-
7 resolution 2014 image. The substantial increase across many degradation metrics between 2011
8 and 2014 and visual inspection of the imagery and the resulting maps indicate that the trend of
9 increasing degradation measured between 2011 and 2014 is a true reflection of changing surface
10 conditions and not solely the product of increased image resolution. Field observations and other
11 studies also corroborate a major degradation event occurring during this period (Liljedahl et al.,
12 2016; Wickland et al., 2020). Finally, pixel-based comparison of trough features over time
13 remains challenging as co-registration errors between images are close in scale to the width of
14 the trough features. Thus, trends in degradation over time were assessed at an aggregated 30 m
15 resolution which alleviate most of the spatial mismatch between co-registered images. Despite
16 remaining challenges, the rapid increase in satellite image resolution and expanding
17 computational methods for analysis of those images are facilitating increasingly accurate
18 detection of fine-scale surface patterns such as detecting degradation in ice-wedge terrain.