Check for updates

JGR | Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2024JH000550

Using Texture-Based Image Segmentation and Machine Learning With High-Resolution Satellite Imagery to Assess Permafrost Degradation Landforms in the Russian High Arctic

Cornelia M. Inauen[1,2], Ingmar Nitze[1], Moritz Langer[1,3], Anne Morgenstern[1], Irena Hajnsek[4,5], and Guido Grosse[1,2]

[1]Permafrost Research Section, Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Potsdam, Germany, [2]Institute of Geosciences, University of Potsdam, Potsdam, Germany, [3]Department of Earth Sciences, Faculty of Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, [4]Department of Civil, Environmental and Geomatic Engineering, Institute of Environmental Engineering, ETH Zürich, Zürich, Switzerland, [5]Microwaves and Radar Institute, German Aerospace Center (DLR) e.V., Wessling, Germany

Abstract Amplified climate change across the Arctic causes significant permafrost thaw and an increase of permafrost degradation landforms. These landforms range from fine-scale degrading ice wedge-polygon-networks to large-scale features such as thermo-erosional gullies and reshape entire landscapes. In particular the expansion of thermo-erosional gullies could have far-reaching consequences by restructuring drainage pathways. Our study aims at finding a suitable remote sensing-based approach for quantifying landscape-scale permafrost degradation in gully-dominated Arctic landscapes. We use historical and recent high-resolution panchromatic satellite imagery allowing multi-decadal analysis of degradation trajectories. Given that degradation stages are characterized by distinct but subtle textural characteristics in satellite imagery, we tested texture-based machine learning segmentation methods including Random Forest (RF) using gray level co-occurrence matrix (GLCM) texture features and deep learning Convolutional Neural Networks (CNNs) using a UNet architecture. For CNN, we tested various framework adjustments. Our results showed that CNN outperforms RF particularly for complex texture-defined classes. CNN reached a micro mIoU of 0.71 (accuracy 83.2%) compared to 0.61 (accuracy 75.9%) for RF. Well-developed baydzherakhs, an advanced stage of ice-wedge-polygon degradation, were detected with high confidence (recall of 0.78–0.96 for CNN). Data augmentation and the use of GLCM features within CNN enhanced robustness against domain shifts. However, the most efficient way to adapt the trained model for additional sites was achieved through targeted fine-tuning. In conclusion, CNN segmentation demonstrated satisfying performance in quantifying fuzzy permafrost degradation stages. It can be expanded in space and time and therefore enables studying long-term permafrost degradation dynamics.

Plain Language Summary Climate change is particularly strong in the Arctic, causing permafrost (permanently frozen ground) to thaw. Permafrost can have high ice contents, whose melting results in localized surface subsidence as the soil collapses into the space previously occupied by the ice. This forms permafrost thaw landforms ranging from meter-scale melting ice wedge-polygon networks to features such as erosional gullies extending over hundreds of meters in length. The development of such landforms can reshape landscapes, impact ecosystems, and alter drainage pathways. On high-resolution satellite imagery, degradation structures can be identified according to their distinct patterns. In our study, we tested machine learning methods to map these structures. To enable long-term analysis of permafrost thaw, we tested these methods on historical and recent greyscale satellite imagery. The tested methods included pixel-based classical segmentation (Random Forest) using texture metrics as inputs and deep learning Convolutional Neural Networks (CNNs). Our results showed that CNNs performed best, providing good results in delineating permafrost degradation across large areas. The model can be adapted and improved for other sites by retraining it with a small amount of site-specific training data. This research is important because it enables understanding how permafrost is changing across the Arctic.

Key Points:

- Image segmentation enables landscape-scale mapping of permafrost degradation stages based on their texture in panchromatic imagery
- Convolutional Neural Networks outperform feature-based Random Forests in identifying subtle target classes with high intra-class variability
- Site specific fine-tuning is an effective way to allow transferring the model to other study sites

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

C. M. Inauen,
cornelia.inauen@awi.de

Citation:

Inauen, C. M., Nitze, I., Langer, M., Morgenstern, A., Hajnsek, I., & Grosse, G. (2025). Using texture-based image segmentation and machine learning with high-resolution satellite imagery to assess permafrost degradation landforms in the Russian High Arctic. *Journal of Geophysical Research: Machine Learning and Computation*, *2*, e2024JH000550. https://doi.org/10.1029/2024JH000550

Received 7 FEB 2025
Accepted 3 JUL 2025

Author Contributions:

Conceptualization: Cornelia M. Inauen, Guido Grosse
Data curation: Cornelia M. Inauen
Formal analysis: Cornelia M. Inauen
Methodology: Cornelia M. Inauen
Software: Cornelia M. Inauen
Supervision: Ingmar Nitze, Moritz Langer, Anne Morgenstern, Irena Hajnsek, Guido Grosse
Visualization: Cornelia M. Inauen

© 2025 The Author(s). *Journal of Geophysical Research: Machine Learning and Computation* published by Wiley Periodicals LLC on behalf of American Geophysical Union.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein, Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Writing – original draft: Cornelia M. Inauen
Writing – review & editing: Ingmar Nitze, Moritz Langer, Anne Morgenstern, Irena Hajnsek, Guido Grosse

1. Introduction

Climate change-induced permafrost thaw is significantly altering Arctic landscapes (e.g., Kokelj et al., 2021; Nitzbon et al., 2020; Nitze et al., 2018; Vincent et al., 2017). Through the combination of ground-ice melting, subsidence, and erosion, degradation landforms can form at various scales and shapes and by a combination of different processes and feedback mechanisms. The widespread emergence of such landforms can serve as indicators for the thawing of ice-rich permafrost. Fine-scale permafrost degradation landforms, for example, are formed as a result of melting ice wedges, which are arranged in typical polygonal structures. This leads to distinct microtopographical changes, forming patterns ranging from polygonal-shaped trough networks to regularly distributed thermokarst mounds (Kanevskiy et al., 2014; Liljedahl et al., 2016; Nitzbon et al., 2021; Walter Anthony et al., 2024). On the other hand, processes such as thaw slumping, ponding, and thermo-erosion can form larger individual or interconnected features such as retrogressive thaw slumps (Nesterova et al., 2024), thermokarst ponds (Abolt et al., 2024), or thermo-erosional gullies (Godin et al., 2014), which all can reach sizes of up to several hundreds of meters. Within this degradation landscape, thermo-erosional gullies play an important role. Apart from accelerating local ice loss, they also restructure drainage pathways at the landscape scale, introducing long-term changes in the hydrological regime and the ecological system, impacting sediment, nutrient, and carbon fluxes (Godin et al., 2014; Harms et al., 2014; Parmentier et al., 2024; Perreault et al., 2016; Zhang et al., 2022). Specifically not only large- but also small-scale catchments are expected to play a significant role in land-ocean fluxes (Vonk et al., 2023). Thermo-erosional gully networks are widespread in Arctic permafrost lowland regions and in some regions dense gully networks are even the dominating permafrost degradation landform (Morgenstern et al., 2021). These landscapes are highly dynamic, and as erosion and channelization are expected to increase (Chartrand et al., 2023; Liljedahl et al., 2024; Rowland, 2023), they are prone to significant changes with geophysical, hydrochemical, and biogeochemical consequences still poorly understood.

Permafrost dynamics at the landscape-scale can be studied by analyzing multi-temporal remote sensing data, allowing insights into the extent and change of different degradation features or patterns. Ice-rich permafrost degradation can be detected with precise observations of surface elevation changes associated with thaw subsidence, for example, by using repeat airborne LiDAR (Douglas et al., 2021; Rettelbach et al., 2021), InSAR (Bernhard et al., 2022), or photogrammetric techniques based on optical imagery (Huang et al., 2023; Kaiser et al., 2022). However, for remote areas such as the High Arctic, the availability of such data at very high vertical and spatial resolution is often limited to few regions and time steps. More recently, new approaches to quantify permafrost degradation employ machine learning methods on high-resolution optical remote sensing imagery. Most of these studies focus on the detection and delineation of specific landform types based on recent multi-band satellite imagery. For example, Nitze et al. (2021) developed deep learning approaches to delineate retrogressive thaw slumps based on high-resolution multispectral optical satellite imagery and elevation data. Witharana et al. (2020) investigated the fusion of high-resolution panchromatic (0.5 m) and moderate resolution (5 m) multispectral imagery to segment ice wedge polygons (IWPs) using Convolutional Neural Networks (CNNs). In the context of understanding longer-term degradation trajectories, extending these methods to historical panchromatic satellite or aerial imagery would expand observation records by several decades. Moreover, to study dynamics and process interactions across landscape units, the main targets are not necessarily individual degradation features, but the spatial extent of different degradation stages consisting of a range of landforms and process stages such as ponding areas or IWP degradation. As these stages are characterized by specific spatial patterns rather than shape, textural characteristics in imagery can provide crucial complementary information in addition to the pixel-based spectral characteristics Hall-Beyer (2017b). This can be a particular asset when working with the limited information content of single-band panchromatic imagery and potential gray tone variations across the different imagery types.

Image texture can be described as a function of pixel intensities that form repeated patterns. One of the most used measures of texture is the grey level co-occurrence matrix (GLCM) and its derived features (Ghalati et al., 2022). GLCM metrics are second order statistical measures and describe the interrelationship of neighboring pixel intensities within a region. In their initial paper Haralick et al. (1973) define 14 measures of GLCM features, which allow describing textural characteristics in a highly discriminative manner. Since then, GLCM features have been widely used in classical pixel-based image segmentation especially for biomedical applications but also remote sensing (e.g., Mboga et al., 2017). Commonly used classical supervised machine learning algorithms include Random Forest (RF) and Support Vector Machine (SVM) (Sheykhmousa et al., 2020). However, for these methods, selecting the most successful input GLCM features can be time consuming and challenging (Hall-

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein, Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Beyer, 2017b). Especially for applications where targets show high variability in intensity, texture, or shape, a large amount of input features are required to best represent each class. This leads to high dimensional data input size resulting in high computational costs both for the feature extraction and the training (Hall-Beyer, 2017b). CNNs can overcome this disadvantage due to their ability to learn spatial-contextual features directly from the data though several hierarchical non-linear layers making feature selection redundant (LeCun et al., 2015). Several studies have shown that CNNs can outperform classical image segmentation (for SAR remote sensing, e.g. Garg et al. (2021); for optical remote sensing, e.g. Korznikov et al. (2021)). A drawback of deep leaning methods, however, is the requirement for large amounts of labelled data. Creating labels from scratch is time consuming. Furthermore, different image quality, high inter-class variability, and unclear or gradual boundaries make consistent labelling difficult. Ways to overcome these issues are for example, image augmentation, where the amount of labelling data is artificially increased by creating new imagery through geometrical and spectral distortion of the original image. This approach not only provides a good tool to substantially increase otherwise sparse labelling data but can specifically be used to learn invariance, for example, to account for illumination distortions in the data source (Dosovitskiy et al., 2016). Other studies also suggested the usage of GLCM features directly as input for CNN networks as they provide additional textural information which is difficult to learn due to their non-linearity (Loebel et al., 2022). Y. Zhang et al. (2024), Tan et al. (2020) and H. Zhang et al. (2021) for example, showed that including GLCM matrices or features improve classification robustness and accuracy, while the latter two emphasized this to be particularly true for cases of high model complexity and limited availability of labelled data. However, the approach might also lead to overfitting (Loebel et al., 2022). In contrast, we did not consider other mainstream methods such as multiscale object-oriented segmentation approaches used in OBIA/GEOBIA. While these have been extensively used for and often demonstrated good performance in segmenting landscapes particularly in correctly delineating object boundaries on high resolution imagery (e.g., Kazemi Garajeh et al., 2022; Shahabi et al., 2019; Witharana et al., 2021), these methods involve a relatively high level of labor-intensive and operator-dependent manual, multistage fine tuning. Considering the gradual boundaries of our target classes, our main goal was to achieve good performance in representing areal coverages rather than exact boundaries delineation, hence we did not consider object-based methods in the context of this study.

With our study, we tested and compared classical supervised machine learning with CNN segmentation to identify the most suitable framework for extracting the extent of different permafrost degradation stages at the landscape scale. Specifically in order to unlock the treasure of abundant but currently strongly underused high-resolution historical greyscale satellite imagery allowing for decadal-scale change detection, we required an approach that provides good result with single band imagery of different sensor types. Furthermore, as our targets are not single, distinct features but rather degradation patterns with gradual transitions, we tested different texture-based segmentation frameworks. The approaches included: (a) Supervised, classical RF classification based on GLCM features, and (b) CNN segmentation. For the CNN segmentation we tested four different adjustments: (a) Different levels of augmentation in order to deal with the low amount of labeled data and to train invariance against different image properties, (b) including GLCM features for additional information content to enhance the learning of targets, which are mainly defined by their texture rather than shape, (c) using class and uncertainty weighting to mitigate class imbalance and avoid overfitting of poorly defined boundaries and uncertain labelling, and (d) testing different data normalization schemes to speed up convergence and increase robustness. We compared the different approaches in terms of segmentation quality, but also robustness concerning different image properties and transferability to other regions. We tested our segmentation approaches on three test sites located on the New Siberian Islands in the Russian High Arctic, which represent an area of dense gully networks and have been understudied even though permafrost degradation is expected to be substantial.

2. Study Area

The New Siberian Islands are located in the Russian High Arctic (Figure 1a). They are characterized by a rolling grass-moss tundra landscape incised by dense thermo-erosional gully networks (Pismeniuk et al., 2023; Wetterich et al., 2019). The islands belong to the continuous permafrost zone and are located in the Arctic climate zone exhibiting a low mean annual air temperature of −13.3°C, short summers and a mean total annual precipitation of 252.2 mm according to the 1991 to 2020 reference period of ERA5-land data (Muñoz Sabater, 2019).

Detailed geocryological studies have been conducted on the New Siberian Islands since the 1950s (e.g., Andreev et al., 2009; Kunitsky, 1998; Romanovskii, 1958a, 1958b; Schirrmeister et al., 2011). A large portion of the

10.1029/2024JH000550

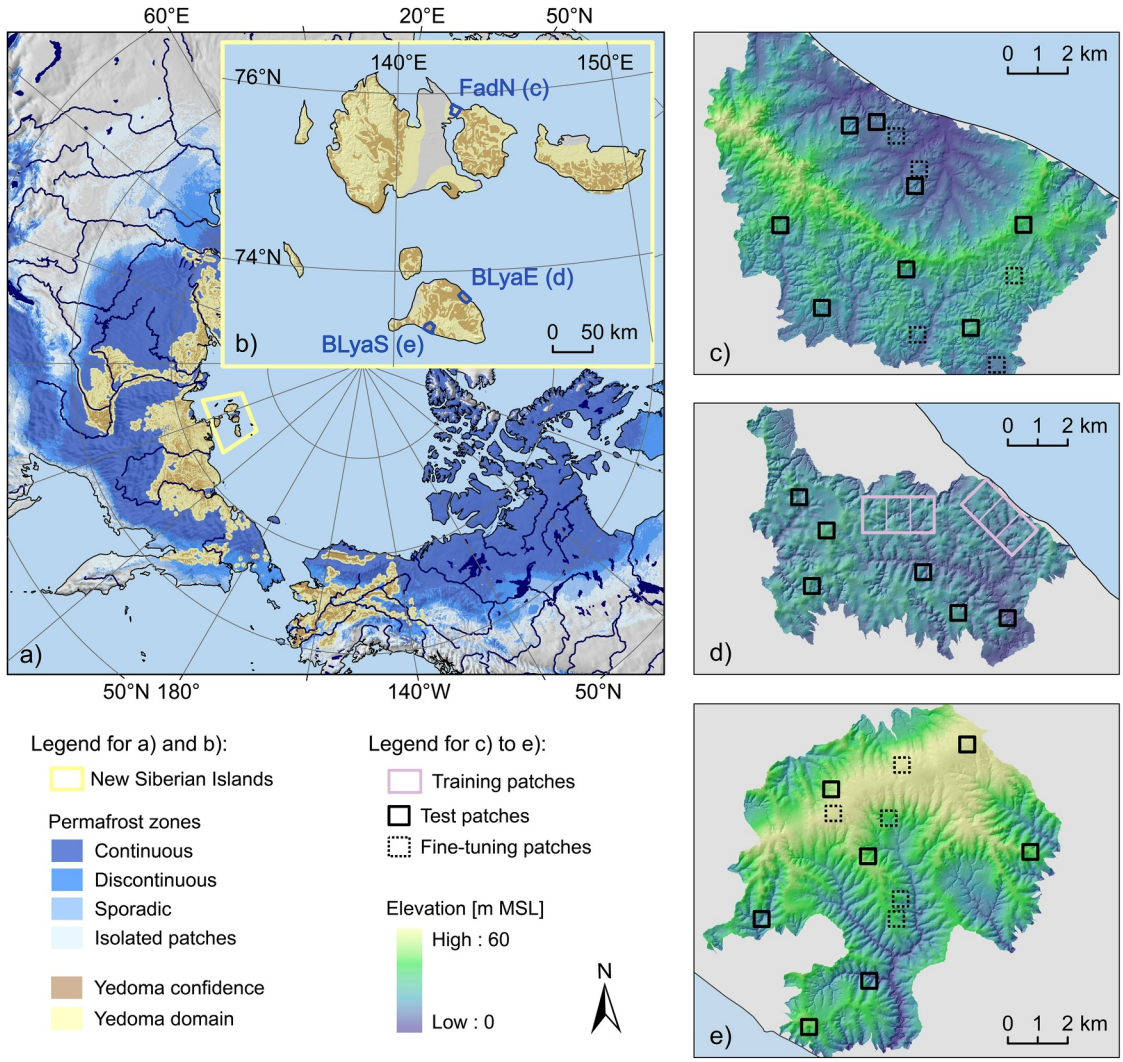

Figure 1. Location and extent of the three study areas on the New Siberian Islands. All areas are located in the Yedoma domain characterized by ice- and organic-rich permafrost deposits and represent the typical dense gully setting within gentle slopes. (Background imagery sources: Permafrost zones (Obu et al., 2018), Yedoma extent (Strauss et al., 2021; 2022), ArcticDEMv4 (Porter et al., 2023) and Natural Earth, Free vector and raster map data @naturalearthdata.com).

undulating topography on these islands is part of the Yedoma domain (Schirrmeister et al., 2011), whose extent is indicated on the map in Figures 1a and 1b. Yedoma-type ice complex deposits were accumulated in unglaciated regions of Beringia during the Pleistocene and can be several 10s of meters thick (Strauss et al., 2017). They consist of ice-rich, organic-bearing sediments, which are intersected by large ice wedges that form a typical ice wedge polygon (IWP) network structure. In our study region, ice wedges were found to be up to 6 m wide and 25 m high on the southern coast of Bol'shoy Lyakhovsky Island (BLya) (Schirrmeister et al., 2011; Wetterich et al., 2014) and up to 4 m wide and at least 5 m high on northern Faddeevsky Island (Fad) (Pismeniuk et al., 2023). Together with the high excess ice content of the sediment in-between the ice wedges, which were found to reach up to 40 to 130 wt% at south BLya, and taking into account that ice wedges can take up to about 50% volume of the Yedoma ice complex, the total ground ice contents can reach values between 65 and 90 Vol% (Schirrmeister et al., 2011). This high ice content makes the Yedoma deposits particularly susceptible to substantial thaw-subsidence and erosion, resulting in distinct permafrost degradation landforms (Strauss et al., 2017). Since the Late Glacial and throughout the Holocene, short-term warming phases and the recent enhanced climate warming, combined with neotectonic movements, have reshaped the landscape of the New Siberian Islands. This has resulted in the formation of large thermokarst depressions dating back to the end of the late Pleistocene and dense thermo-erosional valleys and gullies that dissect and shape the modern topography of Yedoma hills

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(Romanovskii et al., 2004; Schirrmeister et al., 2011). Other clear but smaller scale permafrost features visible from satellite imagery include thermokarst ponds on poorly drained ridges or basins, similar as observed by Fraser et al. (2018) in Alaska, thaw slumping along the coast (Barth et al., 2025) and most prominently the widespread distribution of baydzherakhs. Baydzherakhs represent an advanced stage of IWP degradation in Yedoma deposits and correspond to the sediment-rich polygon centers after melting of the surrounding ice wedges (Godin & Fortier, 2012; Strauss et al., 2017; Veremeeva et al., 2021) (Figure S1 with explanatory sketch in Supporting Information S1). Baydzherakhs occur in groups (massifs), often on sloping terrain and can vary in appearance (size, shape, vegetated or eroded surfaces) depending on the involved thermal denudation processes. Sumina (2020, 2023) conducted a detailed study of baydzherakh occurrences on Kotelny Island, the northeast island of the new Siberian Islands, based on field records from 1974 to 1975. She found shape ranges from small, early-stage or destroyed final-stage flat mounds of about 0.5 m high and 5–8 m in diameter to strongly developed, eroded conical mounds of up to 5 m high and diameters of up to 12 m. The latter were described as being located generally at steeper slopes (>15°), where thermal denudation is very active.

As test regions for exploring different segmentation approaches, we selected two areas on the southernmost island of Bolshoy Lyakhovsky (BLyaE and BLyaS) and one on Faddeevsky Island in the North (FadN) (Figures 1b–1e). The sites include the typical gully-dominated Yedoma landscape and cover catchments at different elevation ranges and latitudes. Low-lying plains such as large, up to several kilometres wide, drained thermokarst depressions (alasses) and steep coastal slopes were deliberately excluded as they are expected to be dominated by different degradation structures and processes.

3. Data and Methods

3.1. Remote Sensing Imagery

As input data, we selected the historical panchromatic KH-9 Hexagon panoramic camera (KH-9PC) imagery and the panchromatic imagery of SPOT-7 (Table S1 in Supporting Information S1). The restriction to panchromatic imagery even for the recent time was chosen in order to maintain direct comparably for change detection. In terms of resolution, the KH-9PC imagery has ground resolution distances (GRD) ranging from 0.6 to 1.8 m (NRO, 1972). The SPOT-7 imagery was provided at a pixel size of 1.5 × 1.5 m but, according to the metadata, has a ground sample distance (GSD) between 2.2 and 2.5 m. These resolutions are sufficient to resolve relatively fine-scale degradation features and patterns such as, for example, baydzherakhs which have a diameter of 5–10 m. The KH-9 Hexagon missions collected imagery over various regions on the globe, including the New Siberian Islands, in the 1970 and 1980s as part of US intelligence observations. The KH-9PC imagery was declassified and made available for public and scientific use through the USGS Earth Explorer portal in 2013 (USGS, 2018). It has been used for example, for urban research (Shahtahmassebi et al., 2023), tree count monitoring (Marzolff et al., 2022) or to map glacier change (Ghuffar et al., 2023). In terms of analyzing permafrost degradation, it is a data source of huge potential as it is available over areas for which historical aerial data is difficult to be obtained (i.e., Siberia) and allows analyzing landscape changes over a ˜50-year observation period at a high spatial resolution. However, a challenge is that the historical and recent satellite imagery have different radiometric qualities and overall processing levels. For example, the historical imagery is neither geometrically nor radiometrically corrected.

3.2. Image Pre-Processing

To facilitate concurrent segmentation, we applied several pre-processing steps that bring the KH-9PC and SPOT-7 imagery to a similar processing level. The different steps are summarized in Table S2 of Supporting Information S1.

To achieve geometric consistency, all files were georeferenced with respect to an already orthorectified SPOT-7 image or the ESRI basemap (for FadN) and resampled to 1.5 m pixel resolution using bilinear interpolation. Due to the strong geometrical distortion of the non-geolocated KH-9PC imagery, georeferencing was done manually in ArcGIS using a combination of polynomial transformation and triangulated irregular network interpolation techniques. With this we could achieve maximum point residuals between 2.5 and 8.2 m. Overall, we note that in particular the geometric correction of the heavily distorted Hexagon imagery in combination with a lack of clearly defined stable structures in the imagery that can be used as manual control points, is very challenging. Specifically, often stable gully junctions and more rarely re-recognizable baydzherakhs were used as control points. The high usage of gully junctions led to a good precision along the gullies but may have caused larger shifts on the

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

relatively heterogeneous ridges (please refer to Table S2 in Supporting Information S1 for parameter details and resulting precisions). The SPOT-7 imagery was already provided in an orthorectified state. For the sites BLyaE and BlyaE no further geometric adjustments were applied. For FadN, where we used the ESRI basemap as a reference, we applied an automatic global and local co-registration using the software AROSICS (Scheffler et al., 2017), providing a maximum shift residual of 8.13 m. The exact geometric stacking precision across the KH-9PC and SPOT-7 imagery is difficult to estimate and varies across the image and away from the control points. By visual comparison we estimate the accuracy to be in the 5–10 m range. This is sufficient to compare changes in areas of baydzherakh occurrence. However, individual baydzherakhs and small ponds can be shifted between the timesteps.

The raw KH-9PC and SPOT-7 imagery exhibited strong differences in gray tone distributions. In particular, the non-radiometrically corrected KH-9PC imagery was skewed toward dark colors. For comparability, we therefore harmonized the gray tone distributions across the different imagery as outlined with the parameters in Table S2 of Supporting Information S1. For the KH-9PC imagery, this involved gamma stretching to enhance details within dark gray areas, specifically on the BLyaE and BLyaS sites. Due to its strong differences, the intensities of the KH-9PC imagery from FadN were aligned to the preprocessed KH-9PC imagery from BLyaE using histogram matching. Furthermore, we harmonized all images to 8-bit quantization. While this results in a minor loss of color grading for the SPOT-7 imagery, it improves consistency for the GLCM calculation. Despite all these intensity adjustments, differences with respect to noise levels and gray tones persisted. Notably, a gray tone gradient remains across the non-radiometrically corrected KH-9PC imagery. All of these factors influence the GLCM calculation as described in Section 3.5 (e.g., Brynolfsson et al., 2017).

3.3. Classification Scheme

Based on a visual comparison of the KH-9PC and SPOT-7 imagery, we selected a suitable classification scheme. Figure 2 shows subsets of the KH-9PC and SPOT-7 imagery that illustrate the specific degradation patterns we targeted for the segmentation. On both imagery types, fine-scale permafrost degradation landforms are readily visible. Apart from the dark colored thermo-erosional gully and valley floors, baydzherakhs are visible as regularly dot-patterned areas. Thaw ponding is expressed as irregular patterned areas intersected with dark gray to black colored thermokarst ponds. Largely non-patterned areas indicate a stable stage without substantial permafrost degradation so far. On slopes, these areas are sometimes intersected by brighter stripes indicating water tracks. An additional class was defined for snow-covered areas, which sometimes were found in deep gullies even in the late summer images. Thus, we distinguished six dominant classes (Table S3 in Supporting Information S1): Stable areas, baydzherakhs, ponding areas, gully base, individual thermokarst ponds, and non-degradation features such as snow (only present on KH-9PC imagery).

3.4. Ground Truth Data Set and Labeling

As ground truth data, we labeled the entire area within two patches of 2.4 × 1.2 km for training and validation on BLyaE (purple rectangles on Figure 1d). For independent model testing, six to eight 0.5 × 0.5 km patches were additionally added at each study site (black solid squares on Figures 1c–1e). The test patches are deliberately distributed across the study sites to ensure the detection of prediction biases, caused for example, by gradual brightness change across the KH-9PC imagery. The classes snow, ponds and gully base were pre-labeled with the software ilastik/LabKit (Arzt et al., 2022; Berg et al., 2019), which provides a RF pixel classification based on intensity as well as on edge and texture measures. The label refinement and the digitization of additional classes was done manually in QGis. Our labeling strategy is outlined in Table S3 of Supporting Information S1. In general, we identified the different degradation stages on lower but class-dependent zoom levels and then refined the delineation by zooming in. For the baydzherakh areas, we choose a minimum visibility of three baydzherakhs as a threshold. Overall consistent labeling was difficult due to (a) different development stages or illumination effects affecting the visibility of baydzherakhs and (b) gradual boundaries between baydzherakhs, ponding, and stable areas. We therefore classified the baydzherakh patches into three uncertainty levels according to their clarity as outlined in Table S3 of Supporting Information S1. Furthermore, we increased the uncertainty depending on boundary proximity on a 10 m or 20 m wide strip along the boundaries of the baydzherakh and the ponding areas respectively (distance Equation 5 in Text S1 of Supporting Information S1). Both can be used to down-weight the importance of these uncertain areas during the training, as further explained in Section 3.6.

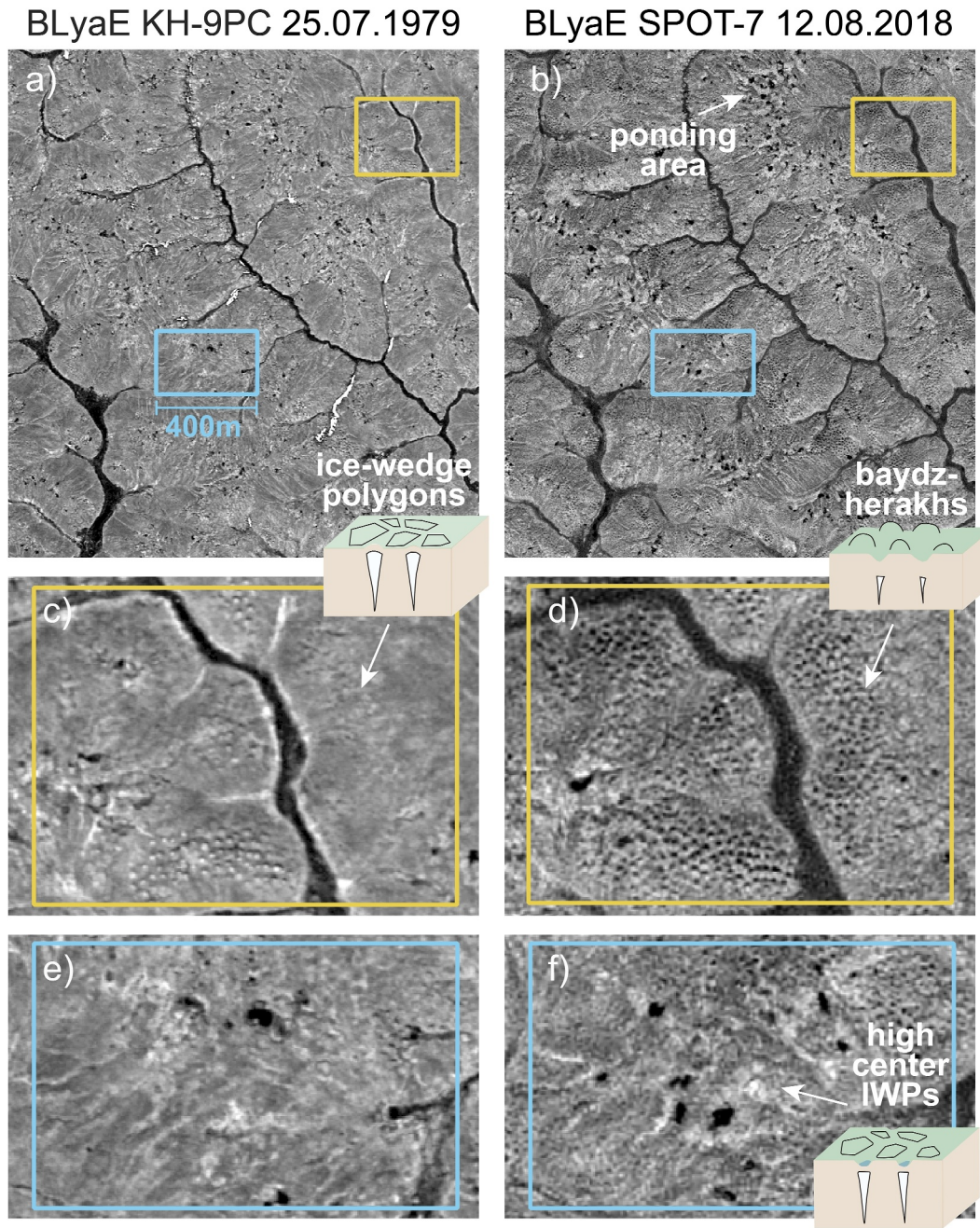

Figure 2. Imagery examples from BLyaE showing different degradation stages: (a) KH-9PC 1979 overview, (b) SPOT-7 2018 overview, (c, d) KH-9PC and SPOT-7 zoom to Baydzherakhs, (e, f) KH-9PC and SPOT-7 zoom to thermokarst ponds. (Background imagery sources: KH-9PC (USGS, 2018), SPOT-7 © Airbus DS (2018)).

Figure 3 shows a labeling example with the manually delineated classes and the certainty labels for the SPOT-7 imagery. Figure S2 in Supporting Information S1 shows an example for the KH-9PC imagery.

3.5. GLCM Calculation

For the computationally intensive calculation of the GLCM features, we used the GPU-capable Python library glcm-cupy (Faracco, 2023). This module calculates seven GLCM features according to Hall-Beyer (2017a) and as summarized in Figure S3 of Supporting Information S1. We tested different sliding window sizes of 5×5, 11×11, 21×21, 41×41 pixels with a step size of one pixel. These sizes were chosen based on the spatial scale of

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

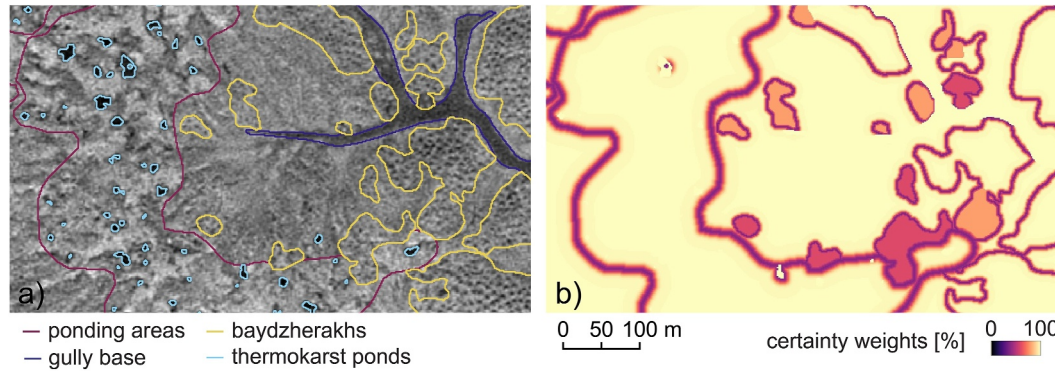

Figure 3. Ground truth data example from BLyaE. (a) Example classification of the SPOT-7 2018 imagery. (b) Certainty weights in percentage. They are used to down-weight pixels in proximity of the ponding and the baydzherakh area boundaries, to account for the uncertain, gradual boundaries. Also downweighted are patches, where the baydzherakhs are less pronounced as outlined in Table S3 of Supporting Information S1. (Background imagery source: SPOT-7 © Airbus DS (2018)).

the specific patterns of the target classes. The sensitivity of the different window sizes for each target class was analyzed using SHAP feature importance analysis (see Section 3.6.2). The texture features were calculated bidirectionally in all directions (East, South-East, South, and South-West) and averaged. Furthermore, to evaluate directionality we calculated the standard deviation between all directions. To avoid edge effects at the tile edges, we calculated the GLCM features for extended tile sizes and then clipped the results to the desired size.

An example of the calculated GLCM features as well as their sensitivity to the different degradation classes is depicted in Figure S4 in Supporting Information S1.

3.6. Semantic Segmentation

3.6.1. General Workflow

The overall goal of the image segmentation is to derive a pixel-wise classification of the imagery such that each pixel is assigned to a degradation stage. The workflow of the segmentation is sketched in Figure 4. We first evaluated different segmentation setups using a limited amount of data, including training data only from BLyaE, to reduce computation time. In the next step, we used the most successful segmentation framework to test spatial transferability to other sites (BLyaS and FadN). Furthermore, we evaluated performance improvements with site-specific model fine-tuning.

In summary, the tested frameworks comprised classical RF using GLCM features as inputs versus CNN segmentation. For the CNN segmentation, we tested further adjustments such as the application of different augmentation levels, multi-feature inputs including GLCM features, different types of normalization, and weighting schemes.

3.6.2. Random Forest Implementation

For the RF implementation, we used a combination of the Python library scikit-learn (Pedregosa et al., 2011) and the GPU capable cuML rapids library (Raschka et al., 2020).

RF classification requires several hyperparameters, which bear trade-offs in terms of accuracy versus stability versus computation time (Probst et al., 2019). Therefore, we first ran a grid search using 5-fold cross validation (CV). Due to the high computational costs, the input was limited to the first CV training set of the easternmost training patch on BLyaE (two northernmost sub-patches, Figure 1d). The tested parameters are specified in Table S4 of Supporting Information S1.

Furthermore, to identify the best combination of GLCM features and window sizes, we evaluated SHapley Additive exPlanations (SHAP) feature importance values. They describe the contribution of each specific input feature on the model's prediction and are implemented in the Python library SHAP (Lundberg & Lee, 2017). As an input for this analysis, we used all seven GLCM features averaged over all directions and calculated for window

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

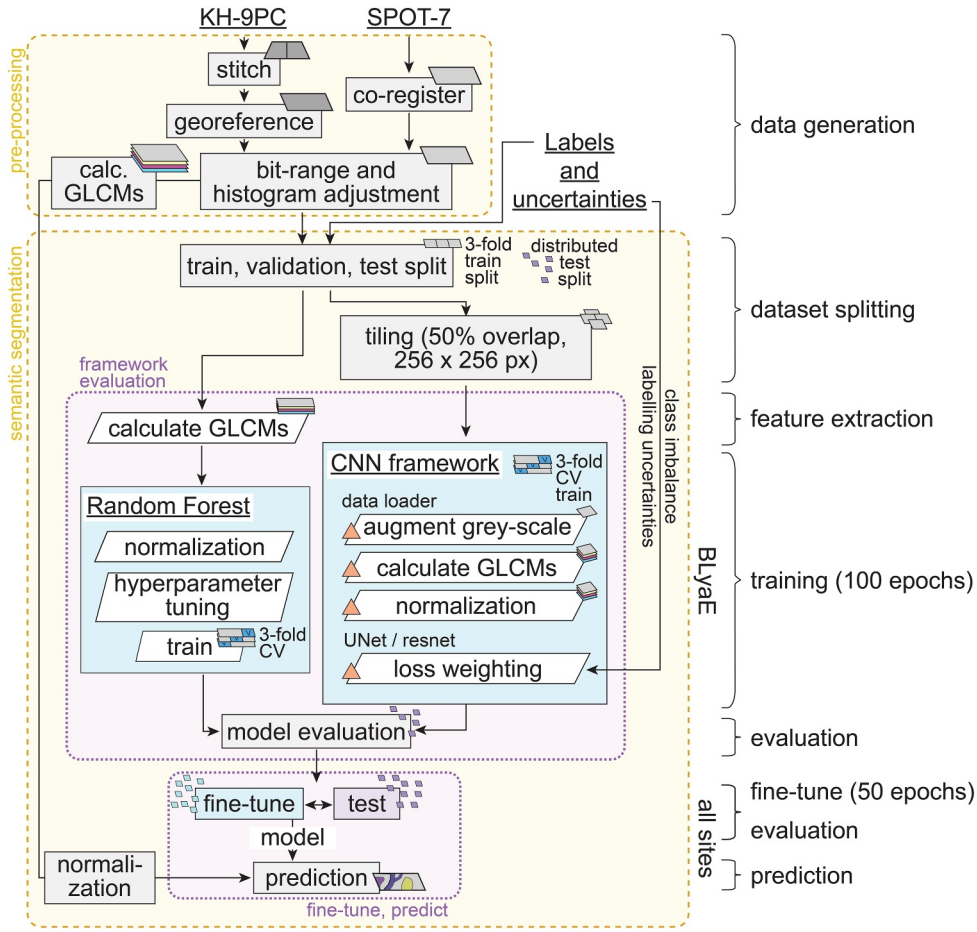

Figure 4. Full processing workflow to derive the segmentation into permafrost landscape degradation stages. The workflow can be subdivided into two steps: (1) The framework evaluation, where we tested different framework adjustments (orange triangles) using 3-fold cross-validation on BLyaE, and (2) the testing and fine-tuning with the best performing framework setup identified in (1) and involving the additional sites BLyaS and FadN.

sizes of 5 × 5, 11 × 11, 21 × 21 and 41 × 41 pixels. In addition we used the standard deviation between the different directions for window size 11 × 11 pixels. Together with the greyscale band this resulted in a total of 36 input features for hyperparameter tuning and feature importance testing. Finally, with the selected hyperparameters and the reduced amount of input features we retrained the model for the segmentation evaluation.

3.6.3. CNN Framework Implementation

The CNN framework was implemented using pytorch (Paszke et al., 2019). We choose a default UNet network architecture, which has been proven successful in many studies (e.g., Loebel et al., 2022; Ronneberger et al., 2015). In particular, we used a network implementation provided by the Python library segmentation_models_pytorch (Iakubovskii, 2019), which we set-up with a resnet34 encoder, five encoder depth layers, batch normalization and randomly initialized weights. During the training, randomized batches of size 5 were selected. To optimize the model weights, an Adam optimizer with the hyperparameter of $\beta 1 = 0.9, \beta 2 = 0.999$ and $\epsilon = 1e{-}8$, suggested by Kingma and Ba (2015), and an exponentially learning rate starting at 0.01 and decaying by a gamma of 0.95 at every epoch was applied to a cross-entropy loss function. The training was set to run for 100 epochs during initial training and 50 epochs for fine-tuning.

3.6.4. Implementation of Specific CNN Framework Adjustments

The framework adjustments as highlighted with orange triangles in Figure 4 were implemented as follows:

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein, Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Augmentation was applied to mitigate the low amount of training data as well as to train invariance against illumination effects and counteract overfitting (e.g., Dosovitskiy et al., 2016). It encompassed geometric augmentation including rotation and mirroring as well as color augmentation including contrast, brightness and gamma adjustments. It was applied to the greyscale imagery and the GLCM features were recalculated accordingly. Augmentation was applied either 2- or 4-fold at every epoch and at a probability of 50%, whereas the augmentation strength was enhanced for the 4-fold augmentation by increasing the perturbation range. All augmentation was implemented using the Python library albumentations (Buslaev et al., 2020).

Multi-feature input by including GLCM features was tested with the goal to better constrain the training process as our targets express specific textural properties. For this we included, in addition to the greyscale band, all GLCM features, which have presented the highest feature importance during the SHAP feature importance analysis based on the RF segmentation. Similarly, for the window size we used the sizes, which provided the highest SHAP values.

Normalization is expected to impact the training convergence and stability (e.g., Ranjbarzadeh et al., 2024). We expected this to be important especially with regards to our data with different illumination properties as well as strongly differing value distributions for the GLCM features. We applied standardization after calculating the GLCM features. It was done for each input feature including the greyscale imagery by subtracting the mean and dividing by the standard deviation, either with respect to the whole training area or with respect to single training tiles. Applying the standardization based on the single image patch has the advantage that it mitigates gradual gray tone changes across the KH-9PC imagery.

The *weighting scheme* was expected to assist the training process by counteracting the strong class imbalance and addressing the high labeling uncertainties (e.g., Bressan et al., 2022). For this, we implemented a scheme that weights the loss function with class-specific and/or pixel-dependent weights. Class weights were applied such that each class is trained with equal importance during training. Thus, for example, the thermokarst pond pixels were assigned larger weights in the loss function to compensate for their under-representation in the training set. High labeling uncertainties were addressed by adding pixel-specific weights that reduced the importance of pixels with high label uncertainties in the loss function. A detailed description of the applied weighting schemes can be found in Text S1 in Supporting Information S1. It was implemented following a similar approach as suggested by Bressan et al. (2022).

3.6.5. Segmentation Framework Evaluation

As a training input for the framework evaluation, we split the two training patches of BLyaE into three sub-patches (purple rectangles in Figure 1d). This allows 3-fold CV by using the two training patches concurrently. This approach maximized the use of the limited training data, resulting in three independent models for each framework setup while preventing data leakage (with two sub-patches serving as validation and the remaining four for training). The required input format for the CNN UNet segmentation are rectangular image patches. Therefore, we subdivided the training sub-patches into 256 × 256 pixel tiles with a 42% overlap in both directions. Without augmentation and by including the KH-9PC and SPOT-7 imagery, this provided us 135 tiles for training and 65 tiles for validation for the CNN segmentation. The training was carried out separately as well as combined for the KH-9PC and SPOT-7 data, allowing to evaluate the performance per data type. Finally, in order to independently evaluate the spatial model transferability, we tested each model on the distributed test patches per site and imagery (black solid squares in Figures 1c–1e). The spatial separation of the validation and the test patches from the training set can be considered as spatially blocked holdouts. They are advantageous over random cross validations since they account for spatial autocorrelation (i.e., similarity of close pixel intensities or structure, see Legendre (1993)) and therefore tend to be less prone to inflating the estimated model performance (Kattenborn et al., 2022).

The selection and evaluation of the best model were based on the validation and test data and included commonly used performance metrics, such as overall Intersection over Union (micro mIoU) and overall f1 score (micro f1 score). For class-specific performance we used IoU, precision, false discovery rate (FDR: percentage false positives with respect to predicted target area) and recall, which is the same as the true positive rate in percentage (TPR: True positives with respect to true target area). Additionally, we analyzed robustness against overfitting by observing any divergence between the training and validation IoU. To avoid strong overfitting, we selected the model at the epoch with the highest class-averaged IoU (macro mIoU), as long as the difference between the

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein, Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

macro mIoU of the training and validation sets was less than 0.1. Furthermore, we analyze the segmentation quality of the baydzherakh class in relation to its visibility (uncertainty class). We consider this important additional information, as it is a challenging target class due to its gradual visibility changes and because it is considered a crucial proxy for permafrost degradation.

3.6.6. Assessment of Model Transferability and Fine-Tuning

After evaluating the best-performing segmentation framework on BLyaE, we evaluated the model's ability to generalize to the additional sites BLyaS and FadN. Furthermore, we investigated the impact of perturbed gray tone distributions. In order to improve model transferability, we tested site specific model fine-tuning, where the model was retrained separately for the sites BLyaS and FadN.This involved five additional fine-tuning patches per site (black dashed squares in Figures 1c and 1e). They were selected deliberately in areas where the model predictions showed poor performance. While retaining the original training and validation input from BLyaE we included test patches either from BLyaS or FadN for validation and test to ensure good model performance at the specific additional site. During fine-tuning the weights of all layers in the pre-trained model were retrained (i.e., no layers were frozen).

4. Results

4.1. Segmentation Framework Evaluation

4.1.1. Random Forest Pre-Tests and Feature Selection

As a pre-test to select the hyperparameters and input features for the RF segmentation, we performed a grid search and SHAP feature importance analysis. The results of both are shown in the Supporting Information S1. The selected hyperparameter set included a tree count of 500 and a maximum tree depth of 10 (Table S4 in Supporting Information S1). This configuration achieved the best trade-off between performance and computation time (Figure S6 in Supporting Information S1).

The results of the SHAP feature importance analysis highlighted that the sensitivity of the model predictions to specific window sizes or input features varies depending on the target class (Figure S7 in Supporting Information S1). In general, large-scale features, such as ponding areas, tend to be better defined with GLCM features derived from larger windows. This contrasts with single thermokarst ponds, for which the greyscale values and the GLCM mean derived from smaller window sizes had a greater impact on their prediction. Based on this analysis, we selected the input features for the performance analysis of different segmentation frameworks on the training area of BLyaE. This included all window sizes (5×5, 11×11, 21×21, and 41×41 pixels), using all seven GLCM features and including the greyscale values while excluding the directional standard deviations, which showed no impact on the segmentation results.

The performance of four RF segmentation runs, utilizing different window sizes and input features, showed only slight variations in overall micro mIoU but exhibited significant deviations in class-specific IoUs (Figure 5). In accordance with the results from the SHAP feature analysis, the best overall performance was achieved when all window sizes and all seven GLCM features as well as the greyscale values were included (micro mIoU of 0.61, pink reference run "b"). Excluding the greyscale values and the directly related GLCM mean resulted in the lowest micro mIoU of 0.58 (run d), with notable performance decrease primarily on the thermokarst ponds and gully base classes. Their class-specific IoUs dropped by 0.45 and 0.26 respectively. Both classes contain small or thin structures that are, in the case of thermokarst ponds exclusively, and in the case of the gully base mostly, defined by darker gray tones. In contrast to this, excluding the largest window size of 41×41 pixels has led to a minor IoU decrease of 0.06 on the ponding areas, a class defined by complex large scale patterns.

Across all runs, RF achieved comparably low IoU values for the baydzherakh (run b: 0.37) and ponding area (run b: 0.28) classes. Regarding the snow class, it should be noted that due to the absence of snow in the SPOT-7 imagery as well as the uneven distribution of the snow the performance varied strongly between the CV runs as indicated by the error bars.

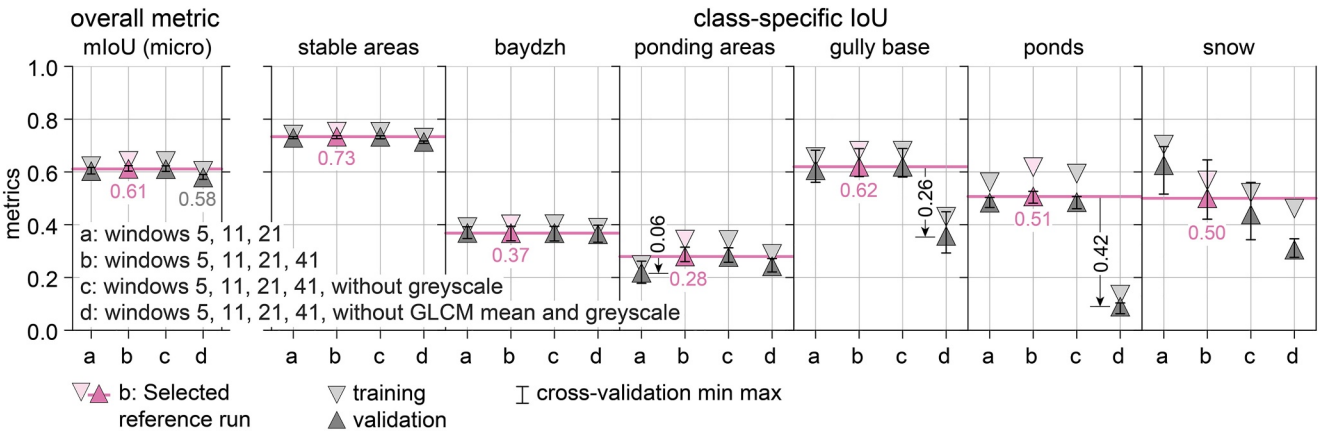

Figure 5. Performance of four random forest training runs using different input features and window sizes. Performance is given in terms of overall and class-specific IoU. All values correspond to the average across all three CVs but are displayed separately for training and validation. The error bars on the validation values indicate the minimum and maximum values of the separate CVs. The colored reference run was selected as the most suitable for further testing.

4.1.2. Impact of CNN Framework Adjustments

The framework adjustments influenced the final model performance as well as the degree of overfitting, both of which varied depending on the target class (Figure 6). The best overall performance in terms of overfitting and IoU was achieved with 2-fold augmentation, standardization with respect to the entire training area, and without the use of any weights (micro IoU of 0.71, marked as reference run in blue).

The type of *normalization* primarily influenced model convergence. The framework without any standardization exhibited a slower convergence (run a, Figure 6a), as indicated by the scattering at lower IoU values. Although the training without normalization resulted in lower performance on the snow class (with an IoU reduction of 0.18) and a slightly higher tendency for overfitting when using individual tile-based standardization, there was no significant effect on the overall micro mIoU, which remained approximately 0.71 across all normalization types.

Augmentation improved model performance across all classes (micro IoU +0.04) and substantially reduced the degree of overfitting (Figure 6b). Compared to 2-fold augmentation, the training with 4-fold augmentation exhibited a slightly higher tendency to overfit the baydzherakh and ponding area class and a slightly lower overall performance.

The application of *class weights* resulted in a strong over-prediction. This effect was particularly pronounced in the highly weighted classes (i.e., those with only a small area proportion). For these classes the false detection rate (FDR) increased by 21.7% for the gully base, 29.7% for the single pond and 36.1% for the snow class (Figure S8 in Supporting Information S1) and is reflected in an overall micro mIoU reduction of 0.07 (Figure 6c). The application of *uncertainty weights* in contrast had no notable effect on the performance.

Including GLCM features, that is, *multi-feature input*, in the CNN segmentation did not lead to an improvement of the classification that justifies the significant increase in computational overhead from recalculating the GLCM features at every epoch after augmentation. Similarly to the RF segmentation, excluding the greyscale values and the GLCM mean led to a significant reduction of the IoU on the gully base (−0.11 IoU), the thermokarst pond (−0.18 IoU) and the snow class (−0.16 IoU) (Figure 6d). For further analysis on the test patches, we therefore selected the training including all GLCM features with all window sizes as an additional reference for further evaluation (orange colored reference run "c")

4.1.3. Performance of Best RF Versus Best CNN Frameworks

Our results show that CNN clearly outperforms RF (Figure 7). On the validation patches, the best performing CNN framework (reference run "b") reached a micro mIoU of 0.71 (accuracy 83.2%) compared to 0.61 (accuracy 75.9%) for RF. The same tendency was also observed on the independent test patches (black crosses on Figure 7a). The class-specific validation IoU highlights a significant improvement especially for the baydzherakh (IoU +0.19) and the ponding area (IoU +0.23).

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein, Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10.1029/2024JH000550

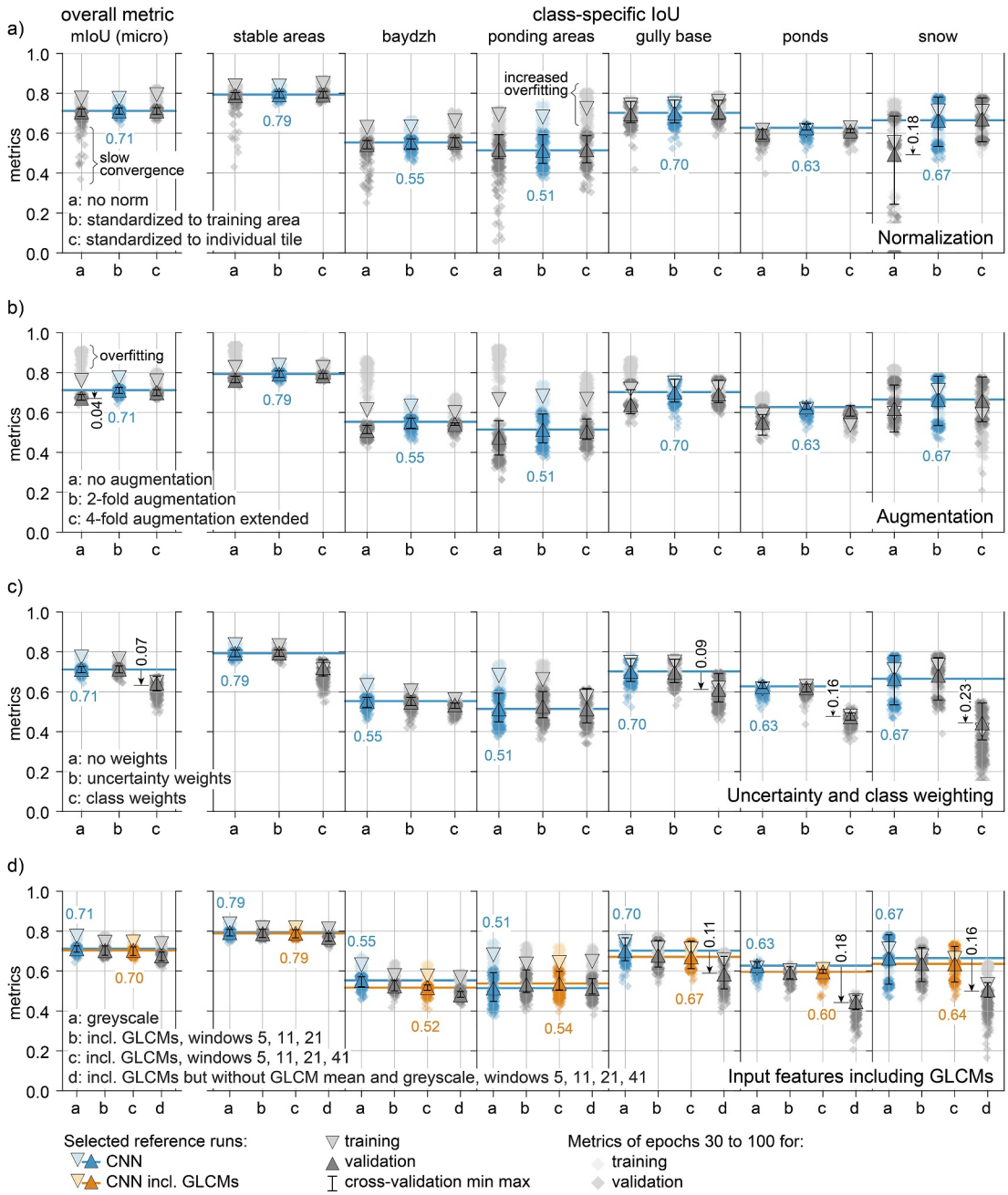

Figure 6. Performance of different segmentation frameworks per subplot in terms of overall and class-specific IoU. The solid triangles correspond to the cross-validation (CV) average of the best metric where the macro IoU of the training does not exceed the validation by more than 0.1. In addition, the background scatter displays all metrics for epoch 30 to 100 indicating potential overfitting if the training metrics strongly exceeds the validation metric. The subplots highlight the effects of the tested framework adjustments: (a) normalization, (b) augmentation (c) class and uncertainty weighting (d) multi-feature input with GLCMs. The colored reference runs were selected for further testing.

Specifically, CNN correctly identified 70.1% (+9.6% compared to RF) of the baydzherakh areas on the test patches of the KH-9PC imagery and 68.5% (+16.5%) on the SPOT-7 imagery (Figures 7d and 7e). In particular, well established baydzherakhs were detected with high confidence. For all CNN frameworks a minimum of 95.9% (L2, high certainty labels) and 79.1% (L1, medium certainty labels) of the baydzherakhs corresponding to the respective certainty level were classified correctly (Figures 7b and 7c). Moreover, CNN clearly lowered the over-prediction of baydzherakhs on the KH-9PC imagery and achieved a reduced FDR of 25.3% (−24.3%) (Figure 7d). The remaining over-prediction of baydzherakhs primarily occurred on particularly heterogeneous

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10.1029/2024JH000550

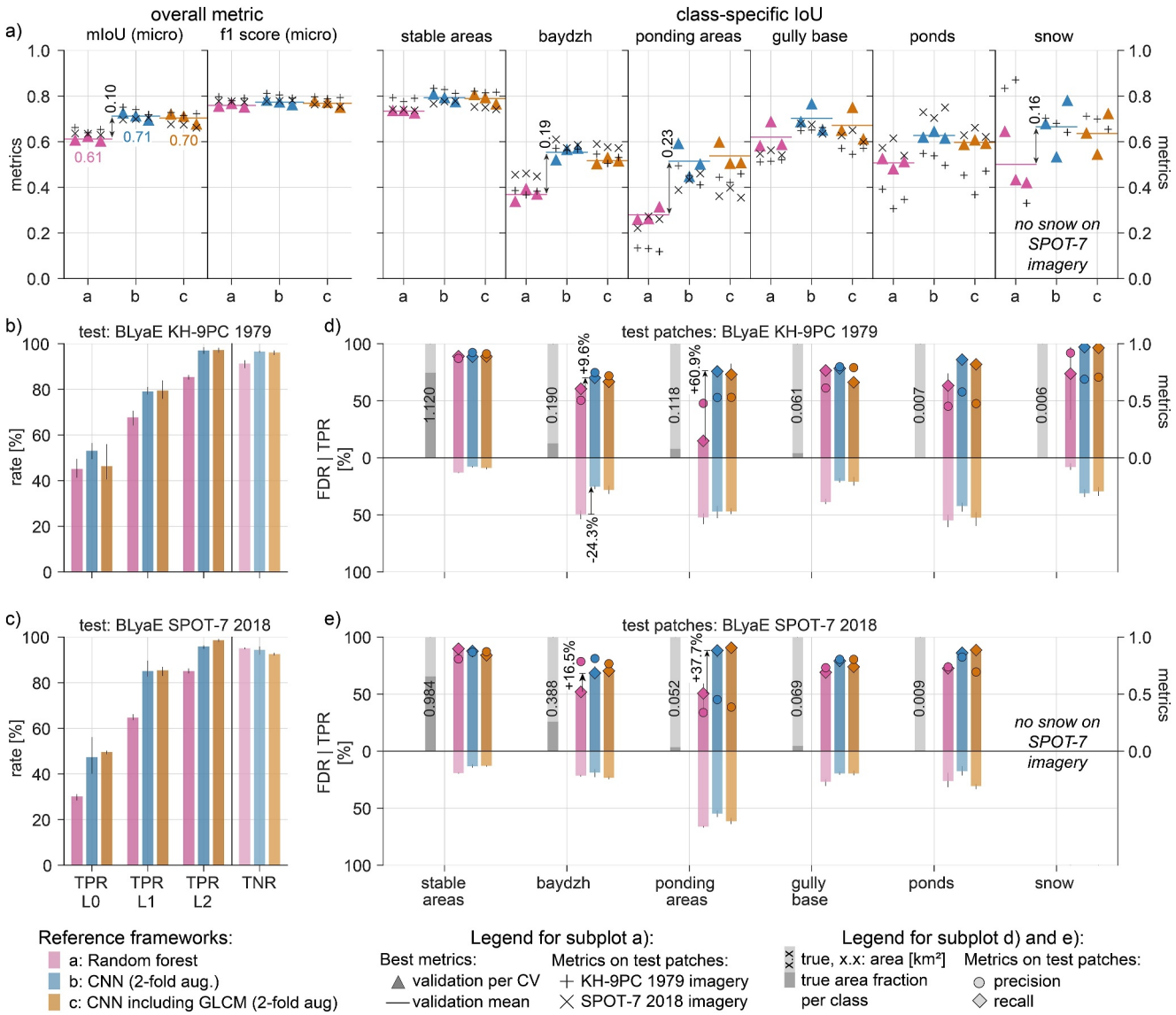

Figure 7. Performance comparison of the three reference segmentation frameworks on the validation and/or test patches of BLyaE. (a) Performance in terms of overall IoU and f1 score as well as class-specific IoU (for validation and test). (b, c) Percentage of correctly predicted baydzherakhs (true positive rate, true positive rates (TPR)) per assigned certainty wights, ranging from less pronounced (L0, ≤60%) to pronounced (L1, >60% and <90%) and to clearly pronounced (L2, ≥90%) (see Figure 3 and Table S3 in Supporting Information S1), as well as true negative rate (TNR) with respect to the non-baydzherakh area. (d, e) Left-hand axis: TPR and false detection rate per class. Right-hand axis: Precision and recall. On subplot (b–e), the displayed bar heights correspond to the cross validation mean, while the error bars indicate the minimum and maximum values.

background or along the patch boundaries, where details of the labeled boundary structures could not be resolved (Figures 8m and 8p).

On the KH-9PC imagery, RF was not able to achieve any reliable prediction for the ponding areas (e.g., Figure 8h). Instead, RF tended to create artifacts of small gully patches in areas of dark pixel values belonging for example, to ponds. Such a scattering tendency is also reflected in the size distribution of the predicted shapes, where RF peaked at too small shape sizes (Figure S10a and S10b in Supporting Information S1). CNN improved the detection of the ponding areas reaching true positive rates (TPR) of 75.8% (+60.9%) and 88.2% (+37.7%) for the KH-9PC and the SPOT-7 imagery respectively. However, a significant over-prediction remained (FDR between 47.1% and 54.8%) as CNN generally predicted the areas around thermokarst ponds as ponding areas, irrespective of the associated heterogeneous background that was used as a labeling criteria (Figure 8).

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10.1029/2024JH000550

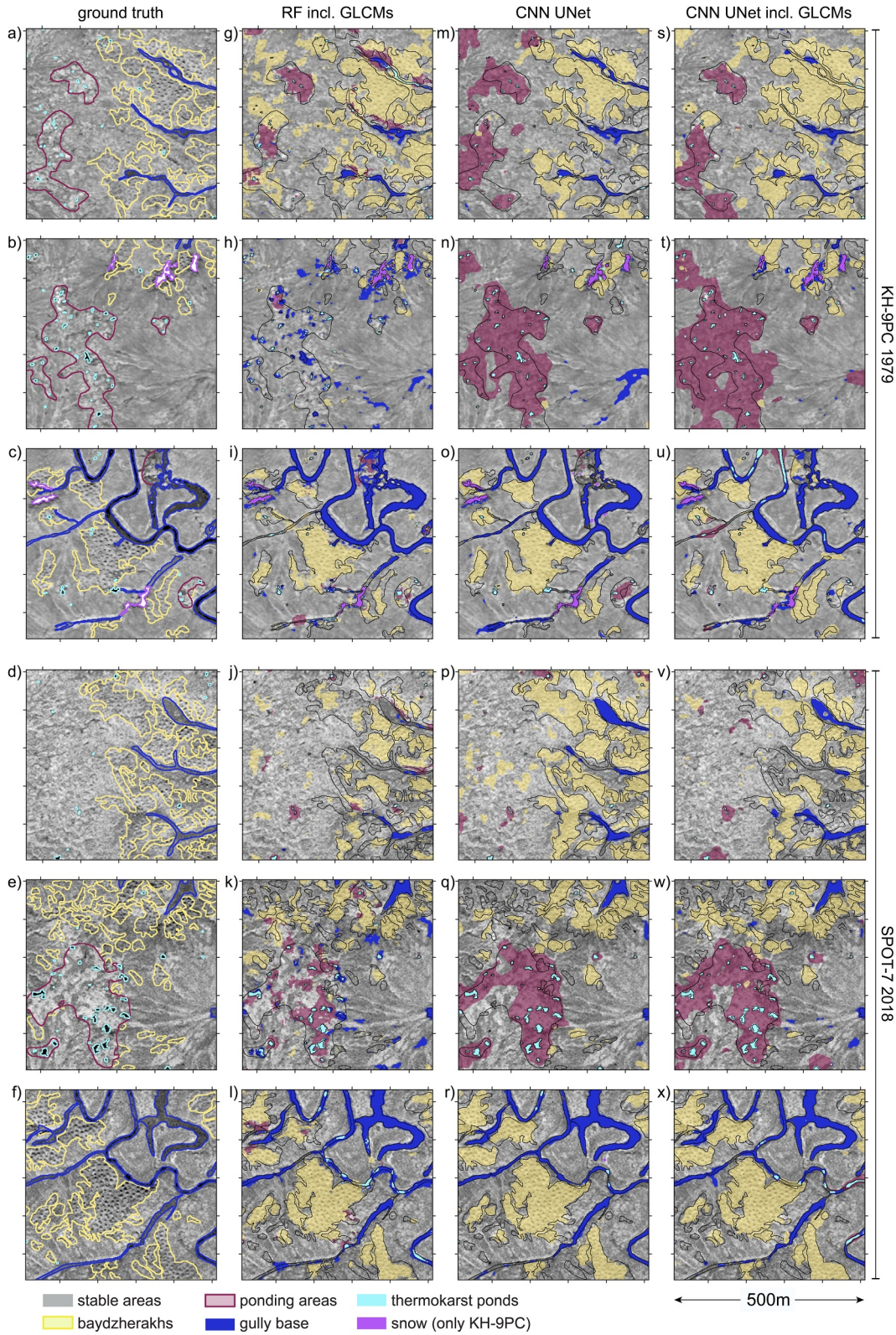

Figure 8. Exemplary predictions on test patches of BLyaE using the segmentation frameworks as displayed in Figure 7 (for the complete set of test patches see Figure S12 in Supporting Information S1). (a–f) Ground truth, (g–x) corresponding predictions. (Background imagery sources: KH-9PC (USGS, 2018) and SPOT-7 © Airbus DS (2018)).

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein, Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

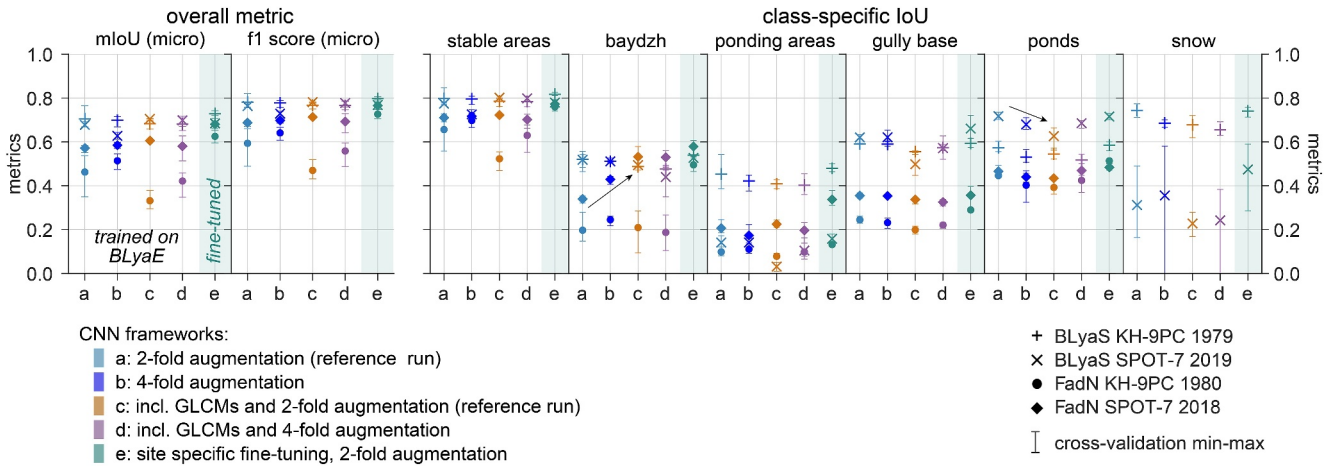

Figure 9. Performance comparison on the additional test sites BLyaS and FadN. Segmentation framework "a" to "d" were solely trained on BLyaE. Framework "e" was additionally fine-tuned with training data from FadN or BLyaS. The performance is displayed in terms of overall IoU and f1 score as well as class-specific IoU. The markers correspond to the cross validation average, while the error bars display the minimum and maximum values.

For smaller-scale structures, all CNN frameworks provided better performance with class-specific IoU values above 0.6. For the gully base, under-prediction occurred predominantly for thin gullies or where the gully base showed no strong contrast to the background (e.g., Figure 8p).

A systematic performance difference between the test patches of the KH-9PC and the SPOT-7 imagery was observed especially for the thermokarst ponds (Figure 7a, black crosses). It occurred for all frameworks and exhibited IoU values that were on average by 0.21 lower on the KH-9PC test patches. The lower performance on the KH-9PC imagery was caused by an over-prediction of primarily small ponds, as well as an overestimation of pond sizes (Figure S10 in Supporting Information S1) or can be attributed to misclassifications of shadows or gully base (Figures 8m and 8n). However, despite these differences it should be noted that the performance of the CNN training was not impacted by concurrent training. This can be seen for example, in Figures S11a and S11b in Supporting Information S1 where the performance on the test patches reaches similar or higher performance irrespective of whether the respective imagery was trained separately or concurrently with the other.

4.2. Performance in Terms of Generalization

4.2.1. Spatial Model Transferability

The performance evaluation on the test patches from the additional sites using framework setups trained solely on BLyaE (training runs "a" to "d" in Figure 9) revealed particularly low prediction reliability for FadN, while performance on BLyaS remained within a similar range to BLyaE for most classes. For FadN, the micro mIoU values were low, ranging between 0.33 and 0.61. In fact, performance across all target classes was reduced. In particular the baydzherakhs were significantly under-predicted for the reference run with 2-fold augmentation reaching mean TPR of only 21.1% and 35.0% for the KH-9PC and the SPOT-7 imagery respectively (Figures S13g and S13h in Supporting Information S1). This is also reflected on the predicted test patches in Figures 10e–10h. In particular, the baydzherakhs with bright hilltops or strong shadowing or water inundation in-between the mounds remained undetected. Furthermore, broader gully valleys without a dark-colored base were not recognized. Instead, the shadows along the edges were often wrongly classified as the gully base while single thermokarst ponds remained frequently undetected or were wrongly assigned to shadows (Figures 10f and 10h).

In contrast to the performance analysis on BLyaE, increased augmentation (4-fold) as well as including the GLCM features clearly improved the IoU of the baydzherakhs at FadN for the KH-9PC imagery (arrow in Figure 9 baydzh). However, this was not the case for the SPOT-7 imagery. Furthermore, it led to a decreased performance on the thermokarst ponds, in particular for BLyaS (arrow in Figure 9 ponds). Overall, we observed that the impacts of the different framework adjustments are complex and differ depending on class, imagery type, and study site.

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10.1029/2024JH000550

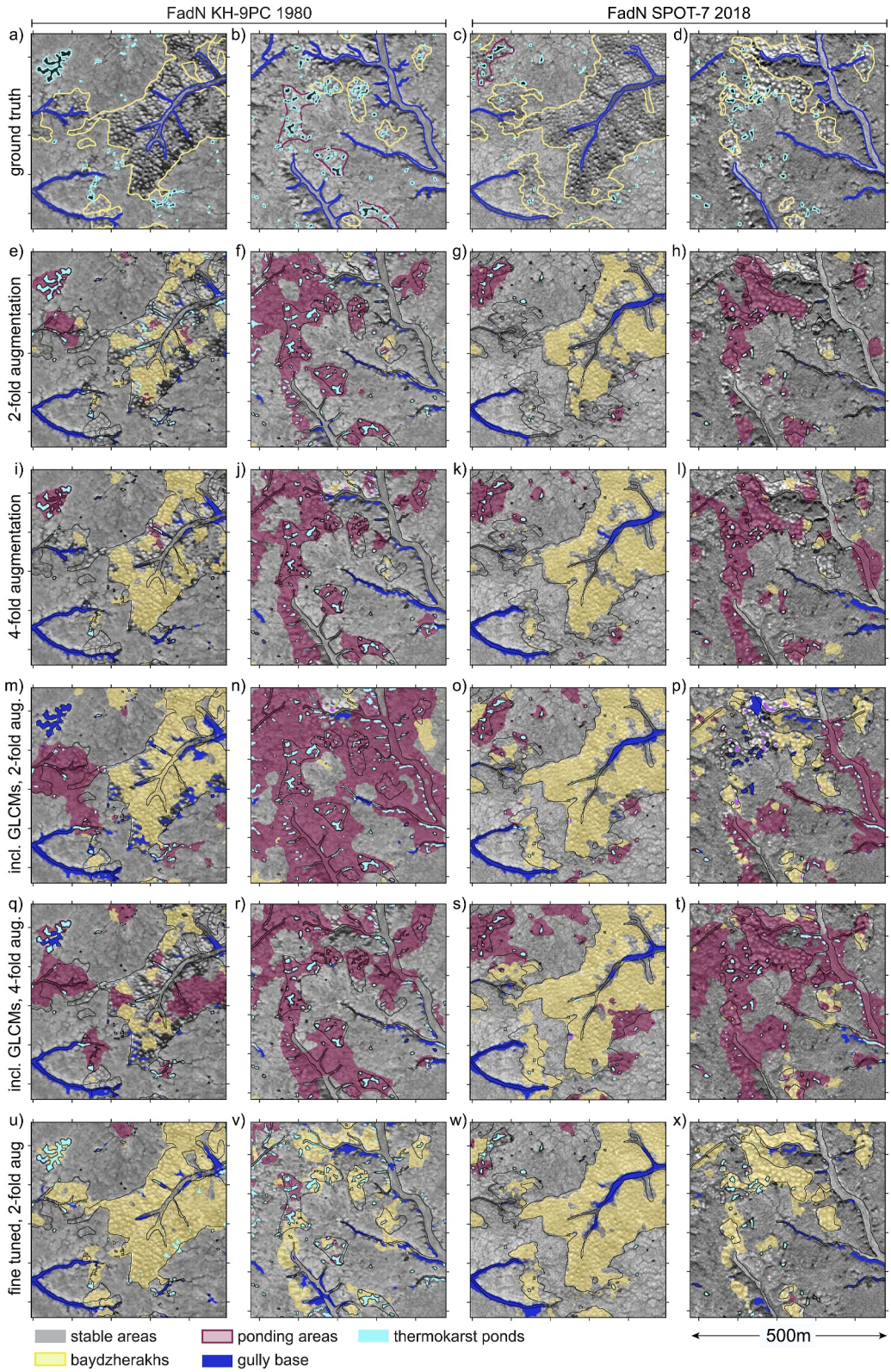

Figure 10. Exemplary predictions on test patches of the additional site FadN using the CNN segmentation frameworks as displayed in Figure 9 (for the complete set of test patches see Figures S14 to S20 in Supporting Information S1). (a–d) Ground truth, (e–t) predictions from the model trained solely on data from BLyaE and (u–x) predictions from the model fine-tuned with site-specific training patches from FadN. (Background imagery sources: KH-9PC (USGS, 2018) and SPOT-7 © Airbus DS (2018)).

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein, Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

A clear and consistent performance improvement was however achieved after separately retraining the model with five fine-tuning patches on FadN and on BLyaS, which were specifically placed at areas of poor segmentation. This particularly improved the detection of baydzherakh areas on FadN and resulted in an TPR of 64.4% (IoU: 0.50) and 67.0% (IoU: 0.58) for the KH-9PC and the SPOT-7 imagery respectively (Figure 9 and Figures S13g and S13h in Supporting Information S1). However, in some cases mislabeling of low lying ponding areas persisted (Figure S14u in Supporting Information S1). Similarly, also the mislabeling of shadows as ponds or the inability to discriminate between gully base and shadows persisted at location with abrupt topographic changes on the imagery of FadN (Figures 10v and 10x).

4.2.2. Robustness Against Gray Tone Changes

Changing the gray tone distribution of the target imagery led to a severe performance decrease, particularly for the thermokarst pond and the snow class (Figure S21 in Supporting Information S1, perturbed KH-9PC gray tone clipped to mean±std4). While the thermokarst ponds remained undetected in the reference run with 2-fold augmentation, the snow class was severely over-predicted (run d in Figure S21c and example prediction in Figure S22d in Supporting Information S1). In the case of the thermokarst ponds, 4-fold augmentation improved the detection rate to TPR 30.8% (+23.4%), while also increasing the FDR to 43.3% (+16.8%). Nevertheless, severe thermokarst pond under-prediction persisted also irrespective of the normalization type (Figures S22e and S22b in Supporting Information S1).

5. Discussion

The goal of this study was to find a suitable segmentation approach that allows quantifying the extent of permafrost degradation stages at the landscape scale and on historical and recent greyscale imagery. The assessment was done in two main steps according to which the discussion is divided. First, the best CNN framework was evaluated based on the test site BLyaE. Second, we tested the model's ability to generalize to other sites (BLyaS and FadN). Please note that in the following the stated metrics refer to the CV-averaged values.

5.1. Segmentation Framework Evaluation

5.1.1. CNN Outperforms RF for Landscape-Scale Permafrost Thaw Detection

CNN segmentation using just greyscale imagery as input feature clearly outperformed RF including GLCM texture features and provided better performance on all classes. This was reflected in a significantly higher overall mIoU of 0.71 (accuracy 83.2%) for CNN compared to 0.61 (accuracy 75.9%) for RF (Figure 7, reference framework “a” and “b”). Unlike CNNs, which can learn contextual information at multiple scales through convolutional layers (LeCun et al., 2015), RF in this study learns the contextual information solely from the GLCM features assigned to each pixel. Although incorporating multiple GLCM window sizes and input features helps to mitigate the challenges in segmenting classes with varying scales and characteristics (as shown in Figure 5 and the feature importance analysis in Figure S7 of Supporting Information S1, and similarly observed by Coburn and Roberts (2004) and Liu et al. (2023)), these features were not sufficiently discriminative to separate the target classes at a pixel level. This limitation was already indicated by the overlapping class distributions in the sensitivity analysis (Figure S4 in Supporting Information S1), and was particularly problematic for texture-defined structures such as baydzherakhs and ponding areas. In comparison to CNN, RF is more prone to scattering, often misinterpreting local spots within textured areas as distinct features based on color (e.g., thermokarst ponds or thermo-erosional gully bases). Similar findings were reported by Mboga et al. (2017), where CNN with up to five convolutional layers outperformed classical supervised classification using SVM with GLCM features for settlement detection or by Garg et al. (2021) who compared the use of the deep learning model DeepLabv3+ to RF including textural descriptors for segmenting PolSAR imagery for urban land cover mapping. Both of these studies showed a higher accuracy and smoother segmentation with less scattering for the deep learning model.

Thus in summary, our study suggests that, despite limited labeling data, CNN clearly outperforms RF in segmenting complex multi-scale and texture-defined target classes. Moreover, CNN, when used without GLCM features, is computationally advantageous, as it learns contextual features directly from greyscale imagery, eliminating the need for complex hand-engineering and feature selection.

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein, Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5.1.2. CNN Framework Adjustments Did Not Significantly Improve Model Performance

In order to improve the CNN segmentation, we tested different framework adjustments. Despite our expectation that GLCM features enhance the segmentation by adding additional contextual information, their inclusion did not provide a significant improvement of the site-specific segmentation at BLyaE (Figure 6d). Instead, it led to substantial computational overhead due to the recalculation of the GLCMs at every epoch after augmentation. Similar observations were reported by Loebel et al. (2022), who did not find a clear improvement by including GLCM features in the CNN segmentation of glacier calving fronts. This suggests that the required discriminative features could be learned by the CNN network directly from the available raw training data. However, unlike Loebel et al. (2022), we did not observe an increased tendency for overfitting. This is likely because we recalculated the GLCM features at every epoch after augmentation, rather than using offline augmentation where both the augmentation and GLCM calculation are performed outside the training loop.

In terms of the weighting schemes, the significant decrease in performance (micro mIoU reduction of 0.07, Figure 6c) due to class weighting is surprising, especially since class weighting has been successfully applied and shown to be beneficial in other studies (e.g., Ronneberger et al., 2015). However, we suggest that fine-tuning the weighting ratio might mitigate the significant over-prediction that was observed for all weighted classes (FDA >35.2% and an increase of 15.2%–36.1% compared to the segmentation without class weighting, Figure S8 in Supporting Information S1). Similarly also for uncertainty weighting, which was suggested by Bressan et al. (2022), fine-tuning the weighting factors might improve its impact.

Overall, the framework evaluation on BLyaE, apart from the superiority of augmentation, revealed no significant impact of the different framework adjustment (Figure 6). This is likely because the available raw training data sufficiently represents the target classes on BLyaE, and labeling inconsistencies and unclear class representations may have hindered further performance improvement. Figure S11c in Supporting Information S1 also supports this, showing that training on just one patch did not significantly affect performance.

5.1.3. Segmentation Quality Is Good Despite a High Label Uncertainty

During the model evaluation on site BLyaE, we achieved the best CNN segmentation performance with 2-fold augmentation, standardization to the entire training area, and no weight usage (run 'b' in Figure 7). The averaged micro IoU was 0.71, and the accuracy was 83.2% on the validation set. We consider this a good achievement, even though similar studies on panchromatic historical imagery have reached higher performance. For example, Mboga et al. (2020) achieved an overall accuracy between 87.33% and 98.83% by segmenting three classes (building, high vegetation and mixed/bare/low vegetation) using a UNet CNN. Deshpande et al. (2021) segmented five classes (barren land, built-up area, agricultural land, water-body, vegetation) and achieved an overall accuracy of 98.3%–98.99%. The lower accuracy of our segmentation can be attributed to our more granular (due to the higher imagery resolution) but subtle and less contrasting target classes with gradual boundaries and high intra-class variability. In contrast to clearly distinct classes, such as buildings, water bodies, vegetation, and barren ground, our target classes are more challenging to learn during the training process and objective ground truth labeling is difficult. The latter additionally impedes the final performance as annotation errors can lead to inaccurate evaluation metrics guiding the model to adapt to errors (Fernández-Moreno et al., 2023).

For our imagery, achieving consistent labeling was a challenge, both within individual imagery (e.g., dealing with gradual boundaries) as well as across the KH-9PC and SPOT-7 imagery. For example, the higher scattering and less smooth appearance of SPOT-7 imagery made differentiating noise from baydzherakhs more difficult. The different gray tone distribution of the imagery impacted the delineation of dark-colored thermokarst ponds. Their boundaries are gradual and follow the transition from deep-to shallow-water to wet surfaces or vegetation. Therefore, with changing gray tone distribution, it is difficult to visually define a clear greyscale threshold dividing water surfaces from ground. This was especially a challenge for the non-radiometrically corrected KH-9PC imagery, which also exhibited varying gray tones across the imagery.

These challenges are reflected in the class-specific performances, which exhibit rather large differences (Figure 7a). While the metrics evaluation on the test patches demonstrate accurate predictions for the stable areas (smallest IoU of 0.77) and the gully base (smallest IoU of 0.65), the overall metrics is reduced for the baydzherakhs (smallest IoU of 0.57) and ponding areas (smallest IoU of 0.43). This is again likely a result of the latter

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein, Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

two being less distinct due to their solely pattern-based definitions, their gradual and fuzzy boundaries, and high representation heterogeneity within the class.

Nevertheless, the uncertainty analysis shows that well-established baydzherakhs are predicted with high certainty (minimum TPR of 79.1% for the medium and 95.9% for the high certainty class; minimum true negative rate of 94.4%, Figures 7b and 7c). Thus, our segmentation provides a reliable but conservative estimate of baydzherakhs occurrence, which can serve as a proxy for significant degradation of permafrost with ice-rich polygonal ground. In addition, the segmentation might even be more consistent than manual labeling, as the model identified baydzherakhs that were missed during labeling. Moreover, we labeled the baydzherakh boundaries at high but due to their fuzziness likely subjective detail. Prediction errors along these boundaries are therefore expected but have lowered the final performance metrics.

The metrics of the ponding areas, whose boundaries were particular unclear for annotation, show strong over-fitting and over-prediction (FDR up to 54.8%). This indicates that the model was unable to learn the intended structures. It rather learned this class's connection with thermokarst ponds. Thus, the segmentation of the ponding areas should be interpreted with caution. However, it can still serve as a rough estimation about areas of increased ponding.

The segmentation performed well for the thermokarst ponds on the SPOT-7 test patches (IoU: 0.73, TPR: 86.2%, FDR: 17.5%). However, we observed a large over-prediction of thermokarst ponds on the KH-9PC imagery (IoU: 0.53, TPR: 86.1%, FDR: 42.2%). This is likely connected to the above-mentioned labeling inconsistencies. Nevertheless, Figures S10c and S10d in Supporting Information S1 shows that wrongly predicted ponds (artifacts) are predominantly smaller than 10 pixels. When comparing predictions over time, we therefore suggest removing ponds smaller than 10 pixels (22.5 m^2). Furthermore, we suggest to compare pond counts within specific size ranges rather than the absolute pond areas, as the exact pond area highly depends on the gray tone which is more variable when comparing across different historical image types. This approach allows focusing on the persistence or drainage of larger thermokarst ponds. Overall, it should be noted that small thermokarst ponds are an important source of CH_4 and CO_2 emissions (Holgerson & Raymond, 2016). Additionally, mapping small ponds is necessary to better understand hydrological dynamics, such as for example, the coalescence of smaller ponds into larger ones. Therefore, finding effective automated methods to detect small thermokarst ponds from radiometrical diverse historical imagery is a crucial but still outstanding task. It is particularly challenging, as small-scale ponds might not necessarily be thermokarstic in origin but could also be ephemeral and be formed or strongly influenced by rainfall or snow melt. A consistent interpretation of individual pond areas over time would require multi-seasonal imagery with accurate geometric alignment to understand seasonal fluctuations. While classical detection methods to extract small thermokarst pond areas include image thresholding or unsupervised k-means classification on modern high-resolution panchromatic, near-infrared, or X-band SAR satellite imagery (Muster et al., 2017), a more advanced method focused on the use of refined U-Nets on very high-resolution satellite imagery to extract pond change over time (Abolt et al., 2024). However, these methods were specifically developed using radiometrically well calibrated recent satellite imagery. Our data set did not allow for this level of analysis. Furthermore, achieving a high enough geometric, pixel-wise overlap that would allow comparing the persistence of very small ponds between the historical KH-9PC and recent satellite imagery has shown to be very challenging due to the high initial geometric distortion of the KH-9PC imagery. For recent dynamics, multispectral satellite imagery could provide improved differentiation of water bodies as well as higher geometric precision at inter-seasonal time scales as demonstrated by Freitas et al. (2024). An alternative approach is to detect wider areas affected by ponding, rather than individual thermokarst ponds as these areas are a clearer indicator of thermokarst degradation. However, as demonstrated in this study, ponding areas were, due to their heterogeneous appearance and fuzzy boundaries difficult to be learned and delineated by the segmentation model. Including additional ponding classes, such as areas with polygonal pond arrangements (e.g., Figure S14a and S14c in Supporting Information S1), which are already more prevalent at FadN, may help to better constrain different types of ponding areas.

5.2. Model Generalization

5.2.1. Spatial Transferability is Challenged Due To Different Class Representations

When transferring to the other test sites, we observed a reduced model performance on the imagery from FadN (mIoU between 0.46 and 0.57), while the model performed well on BLyaS (mIoU between 0.68 and 0.71)

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein, Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(Figure 9). This is not surprising as the KH-9PC imagery from BLyaE and BLyaS originate from the same mission set and thus have similar image properties (Table S1 in Supporting Information S1). The degradation stages on FadN also exhibit distinct characteristics, such as baydzherakhs with bright hilltops and frequently inundated troughs between mounds or valleys of steeper topography creating more shadowing (Figures 10f and 10h). Such structures were not present in the training data from BLyaE. Consequently, the segmentation struggled to accurately assign these structures. This is a common issue with deep learning networks known for their limited spatial transferability and several studies have documented this problem. For instance, Mainali et al. (2023) encountered a decline in performance on untrained wetland segmentation sites due to complex and site-specific data patterns. Similarly, Yang et al. (2021) reported lower performance when segmenting damaged buildings on untrained sites. They attributed the limited model transferability to different building structures and increased labeling challenges at specific sites. Wijesingha et al. (2024) emphasized that both spatial and temporal transferability was reduced for RF and 2D-CNN networks in agricultural land cover mapping. In our case, all imagery was acquired between the end of July to mid-August. However, seasonal weather conditions can to some extent impact the appearance of the degradation stages. For example, increased vegetation growth (though limited in the high Arctic) or extent of saturated or water-inundated areas can affect the visual appearance. Additionally, illumination conditions such as the sun angle and shadowing can impact image contrasts and for example, the visibility or appearance of the baydzherakhs. Thus, similar to spatial transferability, we expect temporal model transferability also to have some limitations.

In contrast to the tests on BLyaE, enhanced augmentation and including GLCM features improved the segmentation of the baydzherakhs on the SPOT-7 imagery of FadN (IoU increase from 0.34 to 0.43 and 0.53 respectively, Figure 9), while for the KH-9PC data the performance however remained at a low IoU between 0.2 and 0.25. The partially improved performance when including the GLCMs is likely a result of the additional texture information increasing the weight of texture characteristics in the segmentation. This aligns with the studies of Tan et al. (2020) and H. Zhang et al. (2021) which showed that including GLCM texture information can improve classification robustness in the case of sparse label availability but high intra-class variability. Overall, we conclude that the CNN framework adjustments clearly had a stronger impact on the additional sites compared to BLyaE, in particular with the augmentation and the inclusion of GLCM features. However, the impacts were difficult to predict and varied depending on class, site and imagery type. The lower impact of the framework adjustments on BLyaE can be explained by the relatively low amount of labeling data, which covers the variety and structures of BLyaE and BLyaS but is limited for good performance on FadN.

5.2.2. Gray Tone Distribution of the Target Imagery Affects Feature Detectability

To assess the effect of gray tone shifts, we examined the model's performance on an image with adjusted gray tones (Figures S21 and S22 in Supporting Information S1). The results demonstrate that classes defined by their intensity levels are significantly impacted by gray tone alterations. In our specific case, the ponds became nearly undetectable (reduction of mean TDR from 86.1% to 7.5%, reference run "d" in Figure S21 in Supporting Information S1), while the snow class was strongly over-predicted (increase of FDR from 31.1% to 58.8%). Domain shifts caused by different intensity distributions are a well-known challenge, particularly in biomedical imaging, when CNN models trained on single-band imagery are applied to data from other sensors or devices (Guan & Liu, 2022). Advanced augmentation can address domain shift problems by promoting invariance (Dosovitskiy et al., 2016). In our study, extended 4-fold augmentation improved the pond-IoU for the perturbed gray tones from 0.07 to 0.24, but slightly decreased it on the original imagery (pond-IoU from 0.53 to 0.50) (Figure S21a in Supporting Information S1). This effect was similarly observed by Güllmar et al. (2022), who analyzed the sensitivity of CNN models to input imagery with altered orientation and intensity, as well as the efficiency of augmentation. In their study, they found a significant dependence of segmentation performance on input variations and that augmentation has a positive impact primarily on perturbed test data, while it can even have a negative impact on the training population. However, on our perturbed image example, despite the improved pond detection with increased augmentation (increase of TDR from 7.5% to 30.8%), a clear under-prediction of thermokarst ponds persisted and over-prediction worsened (FDR increased from 26.5% to 43.3%). This finding suggests that histogram matching or similar imagery adaptation is required when applying our standard CNN model to imagery with gray tones strongly deviating from the training set.

Furthermore, in the case of the radiometrically non-corrected KH-9PC imagery, changes in gray tone were the most obvious differences between images from different missions. These differences could not be completely

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

eliminated by the histogram matching as exemplarily shown by not well aligned histograms of the pre-processed KH-9PC per test site (Figure S5a in Supporting information S1). Furthermore, other acquisition or sensor specific variations in imagery characteristics (resolution, contrast, sharpness, or noise) are expected to impact segmentation performance in similar ways as they lead to a difference in the marginal distribution of the feature space in the source and the target domain, also termed as co-variate shift (Ma et al., 2024). In particular for the KH-9PC imagery at FadN, such influences likely contributed to its reduced performance. This is suggested by the relatively strong deviation of its image properties in comparison to the training set even after histogram matching (local standard deviation and entropy in Figure S5c in Supporting Information 1). Sophisticated analysis methods such as for example, the domain shift analyzers as suggested by Kushol et al. (2023) or measures of area of applicability suggested by Wijesingha et al. (2024) could provide an option to detect areas of strong domain shifts requiring model fine-tuning.

Our study further illustrates the necessity of involving the different sensor types in the training process. For instance, the model trained only on the KH-9PC data reached a very poor performance on the test patches of the SPOT-7 imagery (CV averaged micro mIoU of 0.53) compared to the KH-9PC test patches (mIoU 0.75) (Figure S11b in Supporting Information S1, run "a"). This was similarly shown in the context of resolution changes by W. Zhang et al. (2020), who found a strong performance decrease on IWP delineation when applying a model trained on WorldView-2 imagery (˜0.5 m resolution) to unmanned aerial vehicle (UAV, 0.02 m resolution) data or vice versa.

5.2.3. Site Specific Fine-Tuning Improves Spatial Transferability

To improve the spatial transferability of our model to the test sites BLyaS and FadN, we used additional site specific fine-tuning patches for model retraining (Figure 1, black dashed patches). Our results confirm that this is an effective way to improve performance, as demonstrated by several other studies (e.g., Li et al., 2024; Mainali et al., 2023). At FadN, just five fine-tuning patches significantly improved performance on the baydzherakh class (IoU increase from 0.20 to 0.34 to 0.50 and 0.58 for KH-9PC and SPOT-7, respectively) achieving comparable levels to those at BLyaE (0.57 and 0.59 for KH-9PC and SPOT-7). However, mislabeling persisted in features such as thermokarst ponds and thermo-erosional gullies, due to the extended shadowing (Figure 10v) as well as in some areas with low-lying ponding areas (Figure S14u in Supporting Information S1). Additional labeling would be required to learn these challenging conditions. However, we also observed that fine-tuning the model can lead to reduced performance in specific classes, as seen in our case with the gully class at BLyaS. A potential solution to retain the strengths of each model while avoiding site-specific limitations could be to use ensemble predictions, where, in the simplest approach, the predictions of independently trained models are averaged (e.g., Roshan et al., 2024).

Furthermore, to alleviate the need for large amounts of additional data for fine-tuning, the use of synthetically generated training data for example, through Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) could be tested. Moreover, there are many recent studies demonstrating successful applications of domain adaptation methods, which aim at minimizing the domain gap (Ajith & Gopakumar, 2023; Guan & Liu, 2022) or domain generalization methods, which allow generalization to unseen domains (J. Wang et al., 2023). Such more sophisticated methods could strengthen detection as fine-tuning does not automatically lead to overall robustness and generalizability but can be prone to overfitting (Mehdipour Ghazi et al., 2017).

5.3. Applications for Mapping Landscape-Scale Thaw Impacts

Despite the challenges described in the previous section, CNN segmentation has proven to be a promising method for analyzing permafrost thaw impacts at landscape-scale in the High Arctic. It enables mapping of fuzzy landforms at unprecedented level of detail across larger regions, where manual labeling would be both too time-consuming and difficult to maintain consistency. The CNN segmentation is scalable in space and time. By fine-tuning with multiple test patches, the method becomes adaptable to other study sites. Using test patches, distributed across all study sites and covering the main image characteristics (e.g., gray tones) and class representations (e.g., different appearance of baydzherakhs), is essential to assess consistency and detect potential biases, such as those caused by domain shifts or variations in illumination conditions.

Our approach highlights the value of historical KH-9PC data also for quantitative analyses. When compared to more recent SPOT-7 data, the segmentation can be used to analyze typical long-term change patterns and

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

trajectories, whose spatial relationships can also be examined. Furthermore, factors driving permafrost degradation or feedback mechanisms can be inferred through correlation or susceptibility analyses between the segmentation and geomorphological factors (e.g., slope, aspect) or environmental variables (e.g., snow cover, vegetation). While several studies have conducted such analyses, large-scale studies often focus on single degradation landforms (e.g., Luo et al., 2024; R. Wang et al., 2023), similar to most studies applying deep learning techniques (e.g., Nitze et al., 2021; Witharana et al., 2020). However, particularly in gully dominated landscapes, we suggest that the dynamics between degradation stages can be important (e.g., the impact of changed drainage conditions due to thermo-erosional gully development). This was shown in a similar study by Jorgenson et al. (2022), who manually mapped degradation stages and found that increased drainage slowed down permafrost degradation. The ability to automatically segment several degradation stages over large areas facilitates more such studies and could also be used to inform model parametrization (e.g., similar as done in Aalstad et al., 2018). In particular, with respect to the dynamic gully landscape, a better understanding of long-term degradation dynamics and processes is essential for future prognoses and predictions.

Finally, it should be noted that using single-band panchromatic imagery made the segmentation more challenging due to its limited information content. In this study, we deliberately analyzed greyscale imagery to utilize historical satellite imagery, which allows detecting long-term permafrost degradation. However, using recent remote sensing imagery with multispectral bands would provide additional information content. For example, the NIR band is expected to significantly improve the delineation of thermokarst ponds by better differentiating between vegetation, water bodies and shadows.

6. Conclusion

In this paper, we tested various segmentation frameworks to quantify permafrost degradation at the landscape-scale using historical and recent panchromatic satellite imagery. In contrast to most other studies, which mainly focus on the detection of single-class degradation landforms and more recent satellite imagery, our goal was to achieve a multi-class segmentation that allows studying long-term dynamics in the extent of different permafrost degradation stages.

The performance tests showed the following main outcomes:

- *Segmentation framework selection:* CNN achieved superior performance over RF, even with limited labeling data. Model robustness improved with augmentation, but class weighting reduced performance. Including GLCM texture features in CNN had a weak effect on sites well-represented in training data but increased robustness for detecting complex target classes on new sites with different class representations.
- *Class-dependent performance:* The impact of framework adjustments varied depending on class specifications, imagery type, and study site. These differences can be attributed to the unique characteristics and scales of the target classes in the segmentation model. For instance, texture played a significant role in detecting baydzherakhs, while pixel intensity was crucial for identifying individual thermokarst ponds. This finding further emphasizes the need for a multi-scale segmentation approach.
- *Model transferability:* Targeted model fine-tuning using a limited amount of fine-tuning patches, which contained class representations not included in the initial training data provided an efficient way to improve model transferability. Imagery from different missions showed considerable domain shifts due to varying gray tone distributions. Histogram matching to adapt the target imagery to the training imagery was essential for our final model to detect intensity-defined classes such as the thermokarst pond and snow classes.
- *Overall performance:* When validated with distributed test patches and fine-tuned for additional sites, our segmentation approach demonstrated high confidence in delineating clearly established baydzherakhs, which can be seen as a proxy for permafrost thaw. Therefore, it provides a good but conservative estimate for strong permafrost degradation in ice-wedge polygonal landscapes of the Yedoma region. Satisfactory performance was also achieved for the gully base, thermokarst ponds larger than 10 pixels, and the snow class. However, performance was low for poorly defined and more fuzzy ponding areas and some misinterpretation of shadows as thermokarst ponds or gully base at FadN remained.

Our study demonstrates that CNN segmentation can be a powerful tool even for subtle and heterogeneous target classes on historical and recent satellite imagery of different quality. It is able to differentiate permafrost degradation stages and can be expanded in space and time to reveal typical long-term change trajectories at landscape-scale in the High Arctic. This provides a foundation for large-scale quantitative studies on spatial

dependencies, dynamics and feedback mechanisms. Particularly in gully-dominated landscapes, it addresses a critical knowledge gap within permafrost research.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The original KH-9PC imagery used in this study was provided by the United States Geological Survey Earth Resources Observation and Science (EROS) Center (USGS, 2018) and can be downloaded on the EarthExplorer (https://earthexplorer.usgs.gov/). The accompanying data including the processed KH-9PC imagery and classification results is available on Zenodo (Inauen et al., 2025a). The SPOT-7 imagery © Airbus DS was provided by the Bundesamt für Kartographie und Geodäsie (German Federal Agency for Cartography and Geodesy) and ESA (Third Party Mission programme) under a Standard Multi END-Users Licence Agreement, which restricts data dissemination. Therefore, this data cannot be shared. However, it can be acquired through Airbus DS Geo SA (https://space-solutions.airbus.com/imagery/how-to-order-imagery-and-data/). The code used for image processing and segmentation is archived on the Zenodo repository (Inauen et al., 2025b) and the development version is available at https://github.com/cinauen/segment_permafrost_texture. The code is written in Python and involves several Python libraries as cited in the methods section and listed in the code repository. The figures in the manuscript were created using a combination of Matplotlib version 3.9.1 (Hunter, 2007; The Matplotlib Development Team, 2024) available under the Matplotlib licence at https://matplotlib.org/ and seaborn version 0.13.2 (Waskom, 2021) available under the BSD-3-Clause licence at https://seaborn.pydata.org/.

Acknowledgments

The authors would like to thank the Bundesamt für Kartographie und Geodäsie (German Federal Agency for Cartography and Geodesy) (project #843370_230713_079_003) and ESA through their Third Party Mission program (project ID-54054) for providing SPOT-7 imagery, and the United States Geological Survey for providing declassified KH-9PC imagery used in this work. We acknowledge funding support by AWI INSPIRES to CI. Furthermore, CI, IN and GG were supported by the ML4EARTH project funded by the German Federal Ministry for Economic Affairs and Climate Action (Grant 50EE2201C), and IN and GG by the Permafrost Discovery Gateway project funded by Google.org. Additional thanks go to Tabea Rettelbach, Annabeth McCall, Sebastian Laboor and Alexandra Veremeeva for participating in an image labelling sprint and helping in image pre-labelling. Furthermore, we thank the anonymous reviewers for their valuable and constructive comments. Generative AI (chatGPT-4) has been used to assist language improvements and correction at the initial stage of the manuscript. The final version has been subsequently reviewed and revised by all authors. The authors acknowledge support by the Open Access publication fund of Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung. Open Access funding enabled and organized by Projekt DEAL.

References

Aalstad, K., Westermann, S., Schuler, T. V., Boike, J., & Bertino, L. (2018). Ensemble-based assimilation of fractional snow-covered area satellite retrievals to estimate the snow distribution at Arctic sites. *The Cryosphere*, *12*(1), 247–270. https://doi.org/10.5194/tc-12-247-2018

Abolt, C. J., Atchley, A. L., Harp, D. R., Jorgenson, M. T., Witharana, C., Bolton, W. R., et al. (2024). Topography controls variability in circumpolar permafrost thaw pond expansion. *Journal of Geophysical Research: Earth Surface*, *129*(9), e2024JF007675. https://doi.org/10.1029/2024JF007675

Ajith, A., & Gopakumar, G. (2023). Domain adaptation: A Survey. In M. Tistarelli, S. R. Dubey, S. K. Singh, & X. Jiang (Eds.), *Computer vision and machine intelligence* (pp. 591–602). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-7867-8_47

Andreev, A. A., Grosse, G., Schirrmeister, L., Kuznetsova, T. V., Kuzmina, S. A., Bobrov, A. A., et al. (2009). Weichselian and Holocene palaeoenvironmental history of the Bol'shoy Lyakhovsky island, new Siberian archipelago, arctic Siberia. *Boreas*, *38*(1), 72–110. Publisher: John Wiley and Sons, Ltd. https://doi.org/10.1111/j.1502-3885.2008.00039.x

Arzt, M., Deschamps, J., Schmied, C., Pietzsch, T., Schmidt, D., Tomancak, P., et al. (2022). Labkit: Labeling and segmentation toolkit for big image data. *Frontiers of Computer Science*, *4*. https://doi.org/10.3389/fcomp.2022.777728

Barth, S., Nitze, I., Juhls, B., Runge, A., & Grosse, G. (2025). Rapid changes in retrogressive thaw slump dynamics in the Russian high arctic based on very high-resolution remote sensing. *Geophysical Research Letters*, *52*(7), e2024GL113022. Publisher: John Wiley and Sons, Ltd. https://doi.org/10.1029/2024GL113022

Berg, S., Kutra, D., Kroeger, T., Straehle, C. N., Kausler, B. X., Haubold, C., et al. (2019). Ilastik: Interactive machine learning for (bio)image analysis. *Nature Methods*, *16*(12), 1226–1232. https://doi.org/10.1038/s41592-019-0582-9

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

Bressan, P. O., Junior, J. M., Correa Martins, J. A., de Melo, M. J., Gonçalves, D. N., Freitas, D. M., et al. (2022). Semantic segmentation with labeling uncertainty and class imbalance applied to vegetation mapping. *International Journal of Applied Earth Observation and Geoinformation*, *108*, 102690. https://doi.org/10.1016/j.jag.2022.102690

Brynolfsson, P., Nilsson, D., Torheim, T., Asklund, T., Karlsson, C. T., Trygg, J., et al. (2017). Haralick texture features from Apparent Diffusion Coefficient (ADC) MRI images depend on imaging and pre-processing parameters. *Scientific Reports*, *7*(1), 4041. https://doi.org/10.1038/s41598-017-04151-4

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information*, *11*(2), 125. https://doi.org/10.3390/info11020125

Chartrand, S. M., Jellinek, A. M., Kukko, A., Galofre, A. G., Osinski, G. R., & Hibbard, S. (2023). High Arctic channel incision modulated by climate change and the emergence of polygonal ground. *Nature Communications*, *14*(1), 5297. https://doi.org/10.1038/s41467-023-40795-9

Coburn, C. A., & Roberts, A. C. B. (2004). A multiscale texture analysis procedure for improved forest stand classification. *International Journal of Remote Sensing*, *25*(20), 4287–4308. https://doi.org/10.1080/0143116042000192367

Deshpande, P., Belwalkar, A., Dikshit, O., & Tripathi, S. (2021). Historical land cover classification from CORONA imagery using convolutional neural networks and geometric moments. *International Journal of Remote Sensing*, *42*(13), 5144–5171. https://doi.org/10.1080/01431161.2021.1910365

Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M., & Brox, T. (2016). Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(9), 1734–1747. https://doi.org/10.1109/TPAMI.2015.2496141

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Douglas, T. A., Hiemstra, C. A., Anderson, J. E., Barbato, R. A., Bjella, K. L., Deeb, E. J., et al. (2021). Recent degradation of interior Alaska permafrost mapped with ground surveys, geophysics, deep drilling, and repeat airborne Lidar. *The Cryosphere*, *15*(8), 3555–3575. https://doi.org/10.5194/tc-15-3555-2021

Faracco, J. (2023). GLCM-CUPY, version 0.2.1. GitHub repository [Software]. Retrieved from https://github.com/Eve-ning/glcm-cupy

Fernández-Moreno, M., Lei, B., Holm, E. A., Mesejo, P., & Moreno, R. (2023). Exploring the trade-off between performance and annotation complexity in semantic segmentation. *Engineering Applications of Artificial Intelligence*, *123*, 106299. https://doi.org/10.1016/j.engappai.2023.106299

Fraser, R., Kokelj, S., Lantz, T., McFarlane-Winchester, M., Olthof, I., & Lacelle, D. (2018). Climate sensitivity of high arctic permafrost terrain demonstrated by widespread ice-wedge thermokarst on banks island. *Remote Sensing*, *10*(6), 954. https://doi.org/10.3390/rs10060954

Freitas, P., Vieira, G., Canário, J., Vincent, W. F., Pina, P., & Mora, C. (2024). A trained Mask R-CNN model over PlanetScope imagery for very-high resolution surface water mapping in boreal forest-tundra. *Remote Sensing of Environment*, *304*, 114047. https://doi.org/10.1016/j.rse.2024.114047

Garg, R., Kumar, A., Bansal, N., Prateek, M., & Kumar, S. (2021). Semantic segmentation of PolSAR image data using advanced deep learning model. *Scientific Reports*, *11*(1), 15365. https://doi.org/10.1038/s41598-021-94422-y

Ghalati, M. K., Nunes, A., Ferreira, H., Serranho, P., & Bernardes, R. (2022). Texture analysis and its applications in biomedical imaging: A Survey. *IEEE Reviews in Biomedical Engineering*, *15*, 222–246. https://doi.org/10.1109/RBME.2021.3115703

Ghuffar, S., King, O., Guillet, G., Rupnik, E., & Bolch, T. (2023). Brief communication: Glacier mapping and change estimation using very high-resolution declassified Hexagon KH-9 panoramic stereo imagery (1971–1984). *The Cryosphere*, *17*(3), 1299–1306. https://doi.org/10.5194/tc-17-1299-2023

Godin, E., & Fortier, D. (2012). Geomorphology of a thermo-erosion gully, Bylot Island, Nunavut, Canada. *Canadian Journal of Earth Sciences*, *49*(8), 979–986. https://doi.org/10.1139/E2012-015

Godin, E., Fortier, D., & Coulombe, S. (2014). Effects of thermo-erosion gullying on hydrologic flow networks, discharge and soil loss. *Environmental Research Letters*, *9*(10), 105010. https://doi.org/10.1088/1748-9326/9/10/105010

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Proceedings of the 27th international conference on neural information processing systems* (Vol. 2, pp. 2672–2680). MIT Press.

Guan, H., & Liu, M. (2022). Domain adaptation for medical image analysis: A Survey. *IEEE Transactions on Biomedical Engineering*, *69*(3), 1173–1185. https://doi.org/10.1109/TBME.2021.3117407

Güllmar, D., Jacobsen, N., Deistung, A., Timmann, D., Ropele, S., & Reichenbach, J. R. (2022). Investigation of biases in convolutional neural networks for semantic segmentation using performance sensitivity analysis. *Zeitschrift für Medizinische Physik*, *32*(3), 346–360. https://doi.org/10.1016/j.zemedi.2021.11.004

Hall-Beyer, M. (2017a). GLCM TEXTURE: A tutorial. Retrieved from https://prism.ucalgary.ca/server/api/core/bitstreams/8f9de234-cc94-401d-b701-f08ceee6cfdf/content

Hall-Beyer, M. (2017b). Practical guidelines for choosing GLCM textures to use in landscape classification tasks over a range of moderate spatial scales. *International Journal of Remote Sensing*, *38*(5), 1312–1338. https://doi.org/10.1080/01431161.2016.1278314

Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-3*(6), 610–621. https://doi.org/10.1109/TSMC.1973.4309314

Harms, T. K., Abbott, B. W., & Jones, J. B. (2014). Thermo-erosion gullies increase nitrogen available for hydrologic export. *Biogeochemistry*, *117*(2), 299–311. https://doi.org/10.1007/s10533-013-9862-0

Holgerson, M. A., & Raymond, P. A. (2016). Large contribution to inland water CO2 and CH4 emissions from very small ponds. *Nature Geoscience*, *9*(3), 222–226. https://doi.org/10.1038/ngeo2654

Huang, L., Willis, M. J., Li, G., Lantz, T. C., Schaefer, K., Wig, E., et al. (2023). Identifying active retrogressive thaw slumps from ArcticDEM. *ISPRS Journal of Photogrammetry and Remote Sensing*, *205*, 301–316. https://doi.org/10.1016/j.isprsjprs.2023.10.008

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Iakubovskii, P. (2019). Segmentation models pytorch [Software]. *GitHub repository*. Retrieved from https://github.com/qubvel/segmentation_models.pytorch

Inauen, C., Nitze, I., Langer, M., Morgenstern, A., Hajnsek, I., & Grosse, G. (2025a). Supporting data for the paper using texture-based image segmentation and machine learning with high-resolution satellite imagery to assess permafrost degradation in the Russian high arctic [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.15325741

Inauen, C., Nitze, I., Langer, M., Morgenstern, A., Hajnsek, I., & Grosse, G. (2025b). Supporting software for the paper using texture-based image segmentation and machine learning with high-resolution satellite imagery to assess permafrost degradation in the Russian high arctic [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.15325756

Jorgenson, M., Kanevskiy, M., Jorgenson, J., Liljedahl, A., Shur, Y., Epstein, H., et al. (2022). Rapid transformation of tundra ecosystems from ice-wedge degradation. *Global and Planetary Change*, *216*, 103921. https://doi.org/10.1016/j.gloplacha.2022.103921

Kaiser, S., Boike, J., Grosse, G., & Langer, M. (2022). The potential of UAV imagery for the detection of rapid permafrost degradation: Assessing the impacts on critical arctic infrastructure. *Remote Sensing*, *14*(23), 6107. https://doi.org/10.3390/rs14236107

Kanevskiy, M., Jorgenson, T., Shur, Y., O'Donnell, J. A., Harden, J. W., Zhuang, Q., & Fortier, D. (2014). Cryostratigraphy and permafrost evolution in the lacustrine lowlands of west-Central Alaska. *Permafrost and Periglacial Processes*, *25*(1), 14–34. https://doi.org/10.1002/ppp.1800

Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M. D., & Dormann, C. F. (2022). Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, *5*, 100018. https://doi.org/10.1016/j.ophoto.2022.100018

Kazemi Garajeh, M., Li, Z., Hasanlu, S., Zare Naghadehi, S., & Hossein Haghi, V. (2022). Developing an integrated approach based on geographic object-based image analysis and convolutional neural network for volcanic and glacial landforms mapping. *Scientific Reports*, *12*(1), 21396. https://doi.org/10.1038/s41598-022-26026-z

Kingma, D., & Ba, L. (2015). Adam: A method for stochastic optimization. *Conference Paper presented at the 3rd international conference on learning representations, ICLR 2015*. May 2015). https://doi.org/10.48550/arXiv.1412.6980

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

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein, Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Korznikov, K. A., Kislov, D. E., Altman, J., Doležal, J., Vozmishcheva, A. S., & Krestov, P. V. (2021). Using U-Net-Like deep convolutional neural networks for precise tree recognition in very high resolution RGB (red, green, blue) satellite images. *Forests*, *12*(1), 66. https://doi.org/10.3390/f12010066

Kunitsky, V. (1998). Ice complex and cryoplanation terraces of Bol'shoy Lyakhovsky Island. (in Russian). In *Problemy geokriologii* (eds.). R. M. Kamensky, V. V. Kunitsky, B. A. Olovin, & V. V. Shepelev, Permafrost Institute, Yakutsk, (pp. 60–72).

Kushol, R., Wilman, A. H., Kalra, S., & Yang, Y.-H. (2023). DSMRI: Domain shift analyzer for multi-center MRI datasets. *Diagnostics*, *13*(18), 2947. https://doi.org/10.3390/diagnostics13182947

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Legendre, P. (1993). Spatial autocorrelation: Trouble or new paradigm? *Ecology*, *74*(6), 1659–1673. https://doi.org/10.2307/1939924

Li, Z., Chen, B., Wu, S., Su, M., Chen, J. M., & Xu, B. (2024). Deep learning for urban land use category classification: A review and experimental assessment. *Remote Sensing of Environment*, *311*, 114290. https://doi.org/10.1016/j.rse.2024.114290

Liljedahl, A. K., Boike, J., Daanen, R. P., Fedorov, A. N., Frost, G. V., Grosse, G., et al. (2016). Pan-Arctic ice-wedge degradation in warming permafrost and its influence on tundra hydrology. *Nature Geoscience*, *9*(4), 312–318. https://doi.org/10.1038/ngeo2674

Liljedahl, A. K., Witharana, C., & Manos, E. (2024). The capillaries of the Arctic tundra. *Nature Water*, *2*(7), 611–614. https://doi.org/10.1038/s44221-024-00276-9

Liu, J., Zhu, Y., Song, L., Su, X., Li, J., Zheng, J., et al. (2023). Optimizing window size and directional parameters of GLCM texture features for estimating rice AGB based on UAVs multispectral imagery. *Frontiers in Plant Science*, *14*, 1284235. https://doi.org/10.3389/fpls.2023.1284235

Loebel, E., Scheinert, M., Horwath, M., Heidler, K., Christmann, J., Phan, L. D., et al. (2022). Extracting Glacier calving fronts by deep learning: The benefit of multispectral, topographic, and textural input features. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–12. https://doi.org/10.1109/TGRS.2022.3208454

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777). Curran Associates Inc.

Luo, J., Yin, G.-A., Niu, F.-J., Dong, T.-C., Gao, Z.-Y., Liu, M.-H., & Yu, F. (2024). Machine learning-based predictions of current and future susceptibility to retrogressive thaw slumps across the Northern Hemisphere. *Advances in Climate Change Research*, *15*(2), 253–264. https://doi.org/10.1016/j.accre.2024.03.001

Ma, Y., Chen, S., Ermon, S., & Lobell, D. B. (2024). Transfer learning in environmental remote sensing. *Remote Sensing of Environment*, *301*, 113924. https://doi.org/10.1016/j.rse.2023.113924

Mainali, K., Evans, M., Saavedra, D., Mills, E., Madsen, B., & Minnemeyer, S. (2023). Convolutional neural network for high-resolution wetland mapping with open data: Variable selection and the challenges of a generalizable model. *Science of The Total Environment*, *861*, 160622. https://doi.org/10.1016/j.scitotenv.2022.160622

Marzolff, I., Kirchhoff, M., Stephan, R., Seeger, M., Aït Hssaine, A., & Ries, J. B. (2022). Monitoring dryland trees with remote sensing. Part A: Beyond CORONA—Historical HEXAGON satellite imagery as a new data source for mapping open-canopy woodlands on the tree level. *Frontiers in Environmental Science*, *10*. https://doi.org/10.3389/fenvs.2022.896702

Mboga, N., Grippa, T., Georganos, S., Vanhuysse, S., Smets, B., Dewitte, O., et al. (2020). Fully convolutional networks for land cover classification from historical panchromatic aerial photographs. *ISPRS Journal of Photogrammetry and Remote Sensing*, *167*, 385–395. https://doi.org/10.1016/j.isprsjprs.2020.07.005

Mboga, N., Persello, C., Bergado, J. R., & Stein, A. (2017). Detection of informal settlements from VHR images using convolutional neural networks. *Remote Sensing*, *9*(11), 1106. https://doi.org/10.3390/rs9111106

Mehdipour Ghazi, M., Yanikoglu, B., & Aptoula, E. (2017). Plant identification using deep neural networks via optimization of transfer learning parameters. *Neurocomputing*, *235*, 228–235. https://doi.org/10.1016/j.neucom.2017.01.018

Morgenstern, A., Overduin, P., Gunther, F., Stettner, S., Ramage, J., Schirrmeister, L., et al. (2021). Thermo-erosional valleys in Siberian ice-rich permafrost. *Permafrost and Periglacial Processes*, *32*(1), 59–75. https://doi.org/10.1002/ppp.2087

Muñoz Sabater, J. (2019). ERA5-Land hourly data from 1950 to present [Dataset]. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. https://doi.org/10.24381/cds.e2161bac

Muster, S., Roth, K., Langer, M., Lange, S., Cresto Aleina, F., Bartsch, A., et al. (2017). PeRL: A circum-arctic permafrost region pond and lake database. *Earth System Science Data*, *9*(1), 317–348. https://doi.org/10.5194/essd-9-317-2017

Nesterova, N., Leibman, M., Kizyakov, A., Lantuit, H., Tarasevich, I., Nitze, I., et al. (2024). Review article: Retrogressive thaw slump characteristics and terminology. *The Cryosphere*, *18*(10), 4787–4810. https://doi.org/10.5194/tc-18-4787-2024

Nitzbon, J., Langer, M., Martin, L. C. P., Westermann, S., Schneider von Deimling, T., & Boike, J. (2021). Effects of multi-scale heterogeneity on the simulated evolution of ice-rich permafrost lowlands under a warming climate. *The Cryosphere*, *15*(3), 1399–1422. https://doi.org/10.5194/tc-15-1399-2021

Nitzbon, J., Westermann, S., Langer, M., Martin, L. C. P., Strauss, J., Laboor, S., & Boike, J. (2020). Fast response of cold ice-rich permafrost in northeast Siberia to a warming climate. *Nature Communications*, *11*(1), 2201. https://doi.org/10.1038/s41467-020-15725-8

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

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

NRO. (1972). Hexagon camera user guide. *National Reconnaissance Office*. Retrieved from https://www.nro.gov/Portals/65/documents/foia/declass/ForAll/101917/F-2017-00094.pdf

Obu, J., Westermann, S., Kääb, A., & Bartsch, A. (2018). Ground temperature map, 2000-2016, Northern hemisphere permafrost [Dataset]. *Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, PANGAEA*. https://doi.org/10.1594/PANGAEA.888600

Parmentier, F.-J. W., Nilsen, L., Tømmervik, H., Meisel, O. H., Bröder, L., Vonk, J. E., et al. (2024). Rapid ice-wedge collapse and permafrost carbon loss triggered by increased snow depth and surface runoff. *Geophysical Research Letters*, *51*(11), e2023GL108020. https://doi.org/10.1029/2023GL108020

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *arXiv e-prints, arXiv:1912.01703*. (_eprint: 1912.01703). https://doi.org/10.48550/arXiv.1912.01703

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://dl.acm.org/doi/10.5555/1953048.2078195

Perreault, N., Lévesque, E., Fortier, D., & Lamarque, L. J. (2016). Thermo-erosion gullies boost the transition from wet to mesic tundra vegetation. *Biogeosciences*, *13*(4), 1237–1253. https://doi.org/10.5194/bg-13-1237-2016

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein, Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Pismeniuk, A., Semenov, P., Veremeeva, A., He, W., Kozachek, A., Malyshev, S., et al. (2023). Geochemical features of ground ice from the Faddeevsky peninsula eastern coast (Kotelny island, East Siberian Arctic) as a key to understand paleoenvironmental conditions of its formation. *Land*, *12*(2), 324. https://doi.org/10.3390/land12020324

Porter, C., Howat, I., Noh, M.-J., Husby, E., Khuvis, S., Danish, E., et al. (2023). ArcticDEM - mosaics, version 4.1 [Dataset]. *Harvard Dataverse*. https://doi.org/10.7910/DVN/3VDC4W

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, *9*(3), e1301. https://doi.org/10.1002/widm.1301

Ranjbarzadeh, R., Zarbakhsh, P., Caputo, A., Tirkolaee, E. B., & Bendechache, M. (2024). Brain tumor segmentation based on optimized convolutional neural network and improved chimp optimization algorithm. *Computers in Biology and Medicine*, *168*, 107723. https://doi.org/10.1016/j.compbiomed.2023.107723

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *Information*, *11*(4), 193. https://doi.org/10.3390/info11040193

Rettelbach, T., Langer, M., Nitze, I., Jones, B., Helm, V., Freytag, J., & Grosse, G. (2021). A quantitative graph-based approach to monitoring ice-wedge trough dynamics in polygonal permafrost landscapes. *Remote Sensing*, *13*(16), 3098. https://doi.org/10.3390/rs13163098

Romanovskii, N., Hubberten, H.-W., Gavrilov, A., Tumskoy, V., & Kholodov, A. (2004). Permafrost of the east Siberian Arctic shelf and coastal lowlands. *Quaternary Environments of the Eurasian North (QUEEN)*, *23*(11), 1359–1369. https://doi.org/10.1016/j.quascirev.2003.12.014

Romanovskii, N. N. (1958a). New data about quaternary deposits structure on the Bol'shoy Lyakhovsky Island (Novosibirskie islands). *Nauchnye Doklady Vysshei Shkoly (in Russian). Seriya geologo-geograficheskaya*, *2*, 243–248.

Romanovskii, N. N. (1958). Paleogeographic conditions of formation of the quaternary deposits on Bol'shoy Lyakhovsky island (Novosibirskie islands). In V. G. Bogorov & I. Vypysk (Eds.), *Voprosy fizicheskoi geografii polyarnykh stran* (pp. 80–88). Moscow State University. (In Russian).

Romanovskii, N. N. (1958b). Permafrost structures in Quaternary deposits. *Nauchnye Doklady Vysshei Shkoly. Seriya geologo-geograficheskaya*, *3*, 185–189.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical image computing and computer-assisted intervention – miccai 2015* (pp. 234–241). Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28

Roshan, S., Tanha, J., Zarrin, M., Babaei, A. F., Nikkhah, H., & Jafari, Z. (2024). A deep ensemble medical image segmentation with novel sampling method and loss function. *Computers in Biology and Medicine*, *172*, 108305. https://doi.org/10.1016/j.compbiomed.2024.108305

Rowland, J. C. (2023). Drainage network response to Arctic warming. *Nature Communications*, *14*(1), 5296. https://doi.org/10.1038/s41467-023-40796-8

Scheffler, D., Hollstein, A., Diedrich, H., Segl, K., & Hostert, P. (2017). Arosics: An automated and robust open-source image Co-registration software for multi-sensor satellite data. *Remote Sensing*, *9*(7), 676. https://doi.org/10.3390/rs9070676

Schirrmeister, L., Kunitsky, V., Grosse, G., Wetterich, S., Meyer, H., Schwamborn, G., et al. (2011). Sedimentary characteristics and origin of the Late Pleistocene Ice Complex on North-East Siberian Arctic coastal lowlands and islands—a review. *Quaternary International*, *241*(1), 3–25. https://doi.org/10.1016/j.quaint.2010.04.004

Shahabi, H., Jarihani, B., Tavakkoli Piralilou, S., Chittleborough, D., Avand, M., & Ghorbanzadeh, O. (2019). A semi-automated object-based gully networks detection using different machine learning models: A case study of Bowen catchment, Queensland, Australia. *Sensors*, *19*(22), 4893. https://doi.org/10.3390/s19224893

Shahtahmassebi, A. R., Liu, M., Li, L., Wu, J., Zhao, M., Chen, X., et al. (2023). De-noised and contrast enhanced KH-9 HEXAGON mapping and panoramic camera images for urban research. *Science of Remote Sensing*, *7*, 100082. https://doi.org/10.1016/j.srs.2023.100082

Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 6308–6325. https://doi.org/10.1109/JSTARS.2020.3026724

Strauss, J., Laboor, S., Schirrmeister, L., Fedorov, A. N., Fortier, D., Froese, D., et al. (2021). Circum-Arctic map of the Yedoma permafrost domain. *Frontiers in Earth Science*, *9*(1001). https://doi.org/10.3389/feart.2021.758360

Strauss, J., Laboor, S., Schirrmeister, L., Fedorov, A. N., Fortier, D., Froese, D. G., et al. (2022). Database of Ice-Rich Yedoma Permafrost Version 2 (IRYP v2) [Dataset]. *PANGAEA*. https://doi.org/10.1594/PANGAEA.940078

Strauss, J., Schirrmeister, L., Grosse, G., Fortier, D., Hugelius, G., Knoblauch, C., et al. (2017). Deep Yedoma permafrost: A synthesis of depositional characteristics and carbon vulnerability. *Earth-Science Reviews*, *172*, 75–86. https://doi.org/10.1016/j.earscirev.2017.07.007

Sumina, O. I. (2020). Classification of vegetation of baidzharakh massifs in two sites of the arctic tundra subzone in the Siberian sector of the Russian Arctic (In Russ.). *Vegetation of Russia*, *39*, 75–99. https://doi.org/10.31111/vegrus/2020.39.75

Sumina, O. I. (2023). Typology of territorial vegetation units on the example of thermokarst massifs on Kotelny Island (New Siberian Islands) in Russ. *Botanicheskii Zhurnal*, *108*(3), 210–227. https://doi.org/10.31857/S0006813623030110

Tan, J., Gao, Y., Liang, Z., Cao, W., Pomeroy, M. J., Huo, Y., et al. (2020). 3D-GLCM CNN: A 3-dimensional gray-level Co-occurrence matrix-based CNN model for polyp classification via CT Colonography. *IEEE Transactions on Medical Imaging*, *39*(6), 2013–2024. https://doi.org/10.1109/tmi.2019.2963177

The Matplotlib Development Team. (2024). Matplotlib: Visualization with Python (v3.9.1) [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.12652732

USGS. (2018). Archive—declassified data—declassified satellite imagery—3, [Dataset]. *Earth Resources Observation and Science (EROS) Center*. https://doi.org/10.5066/F7WD3Z10

Veremeeva, A., Nitze, I., Günther, F., Grosse, G., & Rivkina, E. (2021). Geomorphological and climatic drivers of thermokarst lake area increase trend (1999–2018) in the Kolyma lowland Yedoma region, North-Eastern Siberia. *Remote Sensing*, *13*(2), 178. https://doi.org/10.3390/rs13020178

Vincent, W. F., Lemay, M., & Allard, M. (2017). Arctic permafrost landscapes in transition: Towards an integrated Earth system approach. *Arctic Science*, *3*(2), 39–64. https://doi.org/10.1139/as-2016-0027

Vonk, J. E., Speetjens, N. J., & Poste, A. E. (2023). Small watersheds may play a disproportionate role in arctic land-ocean fluxes. *Nature Communications*, *14*(1), 3442. https://doi.org/10.1038/s41467-023-39209-7

Walter Anthony, K. M., Anthony, P., Hasson, N., Edgar, C., Sivan, O., Eliani-Russak, E., et al. (2024). Upland Yedoma Taliks are an unpredicted source of atmospheric methane. *Nature Communications*, *15*(1), 6056. https://doi.org/10.1038/s41467-024-50346-5

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., et al. (2023). Generalizing to unseen domains: A Survey on domain Generalization. *IEEE Transactions on Knowledge and Data Engineering*, *35*(8), 8052–8072. https://doi.org/10.1109/TKDE.2022.3178128

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Wang, R., Guo, L., Yang, Y., Zheng, H., Jia, H., Diao, B., et al. (2023). Thermokarst lake susceptibility assessment using machine learning models in permafrost landscapes of the Arctic. *Science of The Total Environment*, *900*, 165709. https://doi.org/10.1016/j.scitotenv.2023.165709

Waskom, M. L. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wetterich, S., Rudaya, N., Kuznetsov, V., Maksimov, F., Opel, T., Meyer, H., et al. (2019). Ice complex formation on Bol'shoy Lyakhovsky island (new Siberian archipelago, east Siberian arctic) since about 200 ka. *Quaternary Research*, *92*(2), 530–548. https://doi.org/10.1017/qua.2019.6

Wetterich, S., Tumskoy, V., Rudaya, N., Andreev, A. A., Opel, T., Meyer, H., et al. (2014). Ice complex formation in arctic east Siberia during the MIS3 interstadial. *Quaternary Science Reviews*, *84*, 39–55. https://doi.org/10.1016/j.quascirev.2013.11.009

Wijesingha, J., Dzene, I., & Wachendorf, M. (2024). Evaluating the spatial–temporal transferability of models for agricultural land cover mapping using Landsat archive. *ISPRS Journal of Photogrammetry and Remote Sensing*, *213*, 72–86. https://doi.org/10.1016/j.isprsjprs.2024.05.020

Witharana, C., Bhuiyan, M., Liljedahl, A., Kanevskiy, M., Epstein, H., Jones, B., et al. (2020). Understanding the synergies of deep learning and data fusion of multispectral and panchromatic high resolution commercial satellite imagery for automated ice-wedge polygon detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, *170*, 174–191. https://doi.org/10.1016/j.isprsjprs.2020.10.010

Witharana, C., Bhuiyan, M. A., Liljedahl, A. K., Kanevskiy, M., Jorgenson, T., Jones, B. M., et al. (2021). An object-based approach for mapping tundra ice-wedge polygon troughs from very high spatial resolution optical satellite imagery. *Remote Sensing*, *13*(4), 558. https://doi.org/10.3390/rs13040558

Yang, W., Zhang, X., & Luo, P. (2021). Transferability of convolutional neural network models for identifying damaged buildings due to earthquake. *Remote Sensing*, *13*(3), 504. https://doi.org/10.3390/rs13030504

Zhang, H., Wang, M., Wang, F., Yang, G., Zhang, Y., Jia, J., & Wang, S. (2021). A novel squeeze-and-excitation W-net for 2D and 3D building change detection with multi-source and multi-feature remote sensing data. *Remote Sensing*, *13*(3), 440. https://doi.org/10.3390/rs13030440

Zhang, T., Li, D., East, A. E., Walling, D. E., Lane, S., Overeem, I., et al. (2022). Warming-driven erosion and sediment transport in cold regions. *Nature Reviews Earth and Environment*, *3*(12), 832–851. https://doi.org/10.1038/s43017-022-00362-0

Zhang, W., Liljedahl, A. K., Kanevskiy, M., Epstein, H. E., Jones, B. M., Jorgenson, M. T., & Kent, K. (2020). Transferability of the deep learning mask R-CNN model for automated mapping of ice-wedge polygons in high-resolution satellite and UAV images. *Remote Sensing*, *12*(7), 1085. https://doi.org/10.3390/rs12071085

Zhang, Y., Huang, M., Chen, Y., Xiao, X., & Li, H. (2024). Land cover classification in high-resolution remote sensing: Using swin transformer deep learning with texture features. *Journal of Spatial Science*, *70*(2), 1–25. https://doi.org/10.1080/14498596.2024.2386317

29935210, 2025, 3, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000550 by Dtsch Zentrum F. Luft-U. Raum Fahrt In D. Helmholtz Gemein., Wiley Online Library on [02/12/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License