



A Sentinel-2 semantic segmentation dataset for thermokarst-associated lake and pond water surfaces in Arctic permafrost lowlands

Ze Zhang^{1,★}, Ziyu Guo^{1,★}, Qingkai Yan¹, Jinbang Zhai¹, Leonid Gagarin², Nikolai Torgovkin^{1,2}, and Andrei Zhang^{1,2}

¹College of Civil Engineering and Transportation; China-Russia International Joint Key Laboratory of Heilongjiang Province for Permafrost Science and Technology (International Cooperation), Northeast Forestry University, Harbin 150040, China

²Melnikov Permafrost Institute, Siberian Branch, Russian Academy of Sciences, 36 Merzlotnaya Street, Yakutsk 677010, Russia

★These authors contributed equally to this work.

Correspondence: Ze Zhang (zez@nefu.edu.cn)

Abstract. Lake and pond water surfaces in thermokarst-affected Arctic permafrost lowlands are important for hydrological, habitat, and carbon-cycle studies. Existing inland-water products rarely provide georeferenced image-mask pairs designed for supervised segmentation of small Arctic water bodies. We present a Sentinel-2 semantic segmentation dataset for thermokarst-associated lake and pond water surfaces across 23 selected permafrost-lowland areas of interest (AOIs) spanning Alaska, northwestern Canada, European Russia, Yamal, and Siberia. The supervised target is visible summer lake and pond water, while separate tile-level fields record thermokarst relevance and non-target-water risk. The dataset contains 5513 image-mask pairs derived from summer 2023 Sentinel-2 Level-2A surface-reflectance composites. Each image is a 256 by 256 pixel GeoTIFF with five bands (B2, B3, B4, B8, and B11), accompanied by a binary water mask and a quality-assurance (QA) mask for invalid observations and ambiguous margins. The train, validation, and test splits contain 3982, 739, and 792 tiles, respectively. The masks contain 29,141,595 positive pixels, representing 8.07 % of all pixels and a pixel-count-equivalent area of approximately 2914.2 km²; 88.35 % of target-water pixels lie north of or on the Arctic Circle at 66°34' N. The package includes 16539 image, water-mask and QA GeoTIFF files, with sample metadata, source records, processing settings, split definitions and a 150-tile low-water subset index. A geometry audit identified cross-split overlap in the legacy development split and in the original Mackenzie Delta holdout. After removing overlapping and nearby development samples and retraining with QA-aware loss, the Mackenzie Delta experiment reached a target-agreement intersection-over-union (IoU) of 0.659 (spatial-cluster bootstrap interval 0.587–0.728). The Yamal and European Russia holdout had no detected test–development footprint overlap and retained its previously reported target-agreement IoU of 0.936. These values measure agreement with the released supervision targets; independent human-reference accuracy remains to be established. The release is designed for supervised segmentation and controlled model comparison within the represented 2023 Arctic permafrost-lowland settings. The dataset is available from the National Cryosphere Desert Data Center at <https://doi.org/10.12072/ncdc.permafrost.db7645.2026> (Guo, 2026). The accompanying model-output dataset is available on Zenodo at <https://doi.org/10.5281/zenodo.22965055> (Zhang et al., 2026).



1 Introduction

Many lakes and ponds in Arctic permafrost lowlands form or evolve in landscapes affected by thermokarst, where ice-rich permafrost thaws, ground surfaces subside, and depressions retain water. Their occurrence, expansion, drainage, and seasonal state are tied to changes in surface energy balance, hydrological connectivity, ecosystem habitat, and greenhouse-gas exchange in northern landscapes (Walter Anthony et al., 2018). Because these lowlands contain dense populations of small water bodies, remote sensing is essential for spatially consistent monitoring.

Satellite-derived water products have improved substantially over the last two decades. Global and regional datasets now describe lake and reservoir storage dynamics (Hou et al., 2024), glacial lake inventories (Lesi et al., 2022), wetland and inland water classes (Lehner et al., 2025), Arctic small-water-body ice fractions (Lin et al., 2026), and independent validation samples for surface-water products (Mukherjee et al., 2024). These products are valuable, but they answer different questions from a machine-learning segmentation benchmark. Training a water-mapping model requires images paired with maps of their water pixels, together with information on uncertain observations and the samples used for evaluation. These elements are often absent from broad geophysical inventories.

Mapping small water bodies in thermokarst-prone Arctic terrain is also technically difficult. The target class includes dark, shallow, turbid, vegetated, and irregularly shaped water surfaces. Similar spectral responses may come from wet sediment, coastal flats, shadows, and dark vegetation. Conversely, deep or organic-rich water can be missed by simple thresholds. Index-based methods such as the normalized difference water index (NDWI) and its modified form (MNDWI) (McFeeters, 1996; Xu, 2006) provide useful screening, while robust supervised labels also require dark-water screening, object context, and image-quality review. Tile-based object filtering introduces an additional challenge because a single lake, river channel, or coastal water body may be split across tile boundaries.

This paper presents a documented set of georeferenced Sentinel-2 image-mask pairs for semantic segmentation of visible lake and pond water surfaces in thermokarst-affected Arctic permafrost lowlands and adjacent lake-rich settings. The release spans multiple lowland environments and includes dedicated review subsets for algorithm development, reproducible benchmarking, and controlled tests of cross-AOI and cross-region transfer.

Mapping products and segmentation training resources should be distinguished by purpose and provenance, rather than by nominal resolution alone. PeRL provides a directly relevant circum-Arctic pond and lake database (Muster et al., 2017). AL-POD maps lake and pond occurrence in Alaska from Sentinel-2 observations acquired during 2016–2021 (Levenson, 2024), overlapping several of our Alaska AOIs. The JRC Global Surface Water record provides a broader surface-water reference (Pekel et al., 2016). Our contribution is an integrated segmentation resource for selected Arctic lowland settings: georeferenced multispectral image–target pairs, a documented workflow combining spectral and object constraints with model support, spatially aligned QA layers, and explicit geographic evaluation records. These components support reuse of the training data and assessment of model transfer across the represented regions. This resource does not replace existing lake inventories or establish thermokarst origin for every mapped object.



Table 1. Comparison with related water and cryosphere data records.

Reference data record	Relevance to the present dataset
PeRL (Muster et al., 2017)	Circum-Arctic permafrost pond and lake database; a direct inventory comparison whose spatial intersection and observation dates must be considered.
ALPOD (Levenson, 2024)	Alaska Sentinel-2 lake and pond occurrence resource for 2016–2021; direct geographic overlap with the Alaska samples. The version-pinned nonzero-support comparison is reported below.
JRC Global Surface Water (Pekel et al., 2016)	Broader water-extent and occurrence reference. Spatial resolution, river inclusion and temporal definition differ from this segmentation target.
Surface-water validation dataset (Mukherjee et al., 2024)	Global high-resolution validation samples for surface-water products, designed for biome-stratified accuracy assessment.
Arctic small-water-body ice fraction (Lin et al., 2026)	Circumpolar 1 km ice-fraction estimates for small Arctic water bodies.
GloLakes (Hou et al., 2024)	Global lake and reservoir storage dynamics with documented source data, product fields, and validation.
Glacial lake dataset (Lesi et al., 2022)	A vector lake inventory built with threshold mapping, visual inspection, and uncertainty reporting.
GLWD v2 (Lehner et al., 2025)	A harmonized global class database for inland waters and wetlands.
This dataset	Georeferenced Sentinel-2 image-mask pairs for thermokarst-associated lake and pond segmentation, with fixed splits, QA/ignore masks, geomorphic interpretation fields, spatial audits and baselines.

55 2 Dataset units and relationship to existing records

An area of interest (AOI) is a selected geographic study area shown in Fig. 1 and listed in Table 2. For each AOI, a summer composite is formed by taking the median of the available cloud-screened Sentinel-2 reflectances at each pixel. We use “full scene” below to mean this exported AOI composite, rather than a single satellite acquisition.

60 A tile is a square image cut from a composite, with 256 by 256 pixels. At nominal 10 m resolution, each tile covers approximately 2.56 km by 2.56 km, or 6.5536 km². The matching binary mask records water and background, while the QA mask identifies invalid observations and uncertain margins. One image and its two matching masks form a sample.

A split assigns samples to three roles: training samples are used to fit a model, validation samples are used to select its settings or checkpoint, and test samples are used to report its performance. The standard split includes samples from the same AOIs in these roles. A geographic holdout instead excludes an entire AOI or region from training to examine transfer to that
 65 area. The spatial checks and the limitations of the original split are described in Sect. 4.8.

The present dataset complements general inland-water inventories as a machine-learning-ready segmentation resource (Table 1).



The dataset was built around five design objectives. First, each sample must be directly usable in standard semantic segmentation pipelines. Second, the target definition must separate visible lake and pond water from obvious non-target rivers, coastal ocean, saturated sediment, snow, ice, and cloud artefacts. Third, split and AOI metadata must be explicit enough for geographic holdout and source-region filtering. Fourth, source-image records and processing settings must document how the samples were produced. Fifth, the baseline model must provide a reproducible technical validation of image-mask alignment and learnability.

Recent machine-learning data records typically document the prediction task, model-ready payload, labels, fixed splits, validation protocol, and reproducibility metadata (Hu et al., 2023; Ahlswede et al., 2023; Freudenberg et al., 2025; Khaldi et al., 2022; Yuan et al., 2025; Tao et al., 2025; Meena et al., 2023; Fang et al., 2024; Betancourt et al., 2021; Ohmer et al., 2026). The present release provides these elements through matching images, water masks and QA layers, with AOI summaries and documented standard and geographic evaluation samples.

3 Study areas and source imagery

3.1 Study areas

The dataset is centred on Arctic Circle permafrost lowland AOIs across Siberia, Alaska, northwestern Canada, European Russia, and Yamal (Fig. 1). The selected regions contain abundant lake and pond surfaces in landscapes where thermokarst processes are common. For geographic reference, Fig. 1 marks the Arctic Circle at $66^{\circ}34' \text{ N}$. A pixel-level audit of the released binary masks shows that target-water pixels span $65.00\text{--}72.60^{\circ} \text{ N}$: 25,745,487 of 29,141,595 target-water pixels (88.35 %) are north of or on $66^{\circ}34' \text{ N}$, while 3,396,108 pixels (11.65 %) occur in adjacent southern permafrost lowlands, principally in Seward Peninsula, Yukon Flats, and Selawik Lowlands. These three southern areas are shown in Supplement Figs. S16, S23 and S15, respectively.

A tile-component audit gives the same overall pattern at object scale: 46,136 of 53,752 connected target-water components (85.83 %) have centroids north of or on the Arctic Circle, 7616 components (14.17 %) have centroids south of it, and 10 components cross the reference line. These are connected groups counted within tiles, not a count of unique lakes extending across tile boundaries. The dataset is therefore a thermokarst-associated water-surface resource centred on the Arctic Circle, with nearby lake-rich lowland samples retained for transfer testing and hard-negative context. The labels represent visible lake and pond water surfaces, while the accompanying geomorphic interpretation fields record thermokarst relevance and non-target-water risk at tile level.

Figure 2 shows the summer composites and released sample locations in Mackenzie Delta and Yamal, the two areas used to illustrate the geographic experiments. Equivalent maps for all 23 AOIs are provided in Supplement Figs. S1–S23. These maps locate the samples within the observed landscape and show the original training, validation and test assignments.



Figure 1. Study area map for the dataset. Colored numbered markers show the 23 selected AOIs across Arctic Circle and adjacent permafrost lowlands, with numbers corresponding to Table 2. The red dashed line marks the Arctic Circle at 66°34' N. Land outlines: Natural Earth.

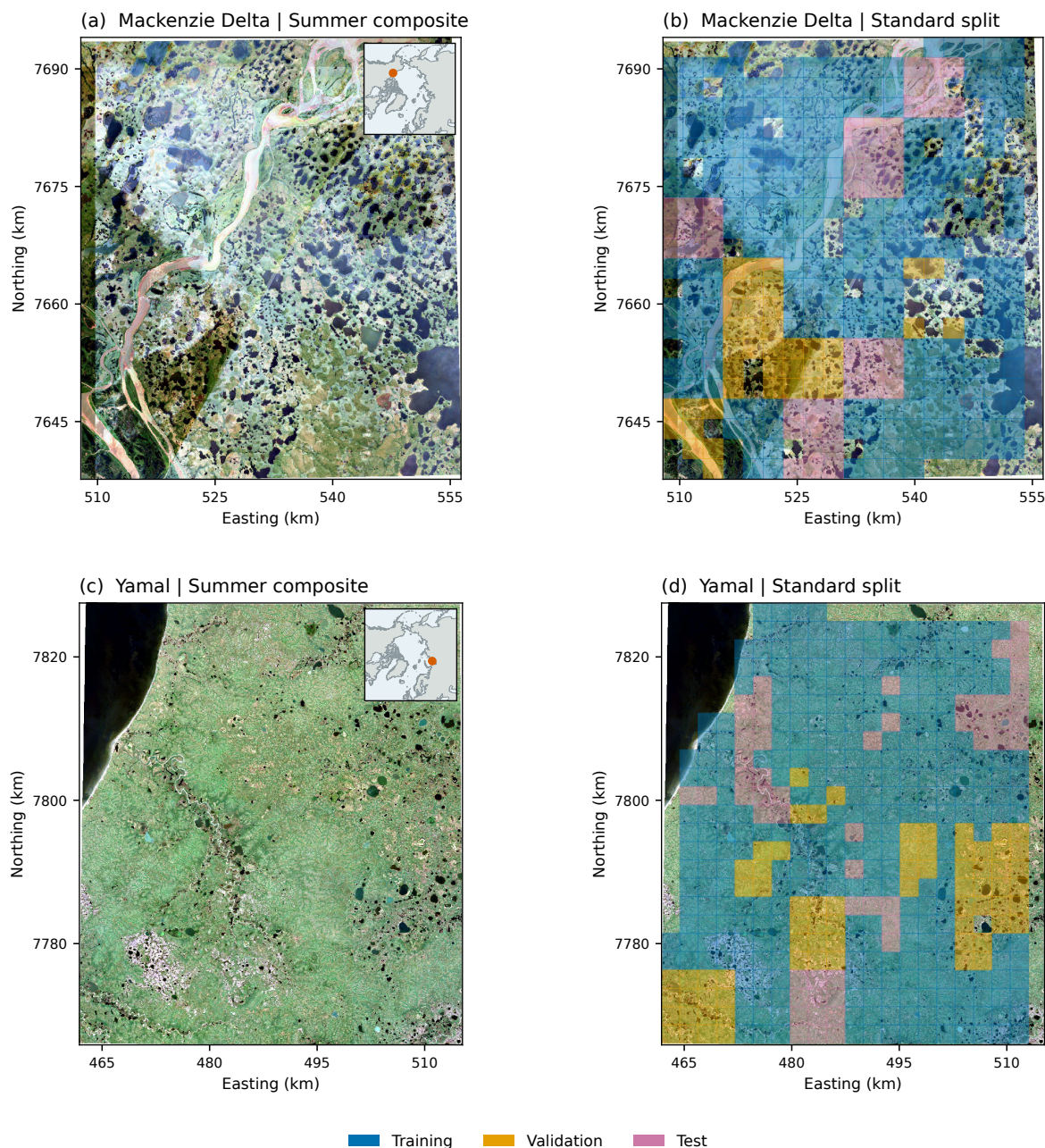


Figure 2. AOI composites and sample distributions for Mackenzie Delta (top) and Yamal (bottom). Left panels show summer Sentinel-2 red-green-blue (RGB) composites, and right panels overlay the released tile footprints coloured by the original training, validation and test assignments. Insets locate each AOI in the Arctic; coordinate axes use the local Universal Transverse Mercator (UTM) grid in kilometres. These original assignments include the spatial limitations documented in Sect. 4.8. Contains modified Copernicus Sentinel data (2023); inset land outlines: Natural Earth.



Table 2. AOI summary for the dataset. CRS denotes coordinate reference system; EPSG codes identify the UTM zones used for export.

No.	AOI	Short name	Composite window	CRS	Tiles
1	Anabar Lowland	anabar_lowland	2023-07-15–2023-08-31	EPSG:32649	40
2	Anderson Plain	anderson_plain	2023-07-15–2023-08-31	EPSG:32609	441
3	Barrow Coastal Plain	barrow_coastal_plain	2023-07-15–2023-08-31	EPSG:32604	66
4	Bykovsky	bykovsky	2023-07-15–2023-08-31	EPSG:32652	32
5	Gyda	gyda	2023-07-15–2023-08-31	EPSG:32644	416
6	Indigirka Lowland	indigirka_lowland	2023-07-15–2023-08-31	EPSG:32655	312
7	Kolyma Lowland	kolyma_lowland	2023-07-15–2023-08-31	EPSG:32657	307
8	Kurungnakh	kurungnakh	2023-07-20–2023-08-20	EPSG:32652	34
9	Lena Delta East	lena_delta_east	2023-07-15–2023-08-31	EPSG:32652	165
10	Lena Delta West	lena_delta_west	2023-07-15–2023-08-31	EPSG:32651	7
11	Mackenzie Delta	mackenzie_delta	2023-07-15–2023-08-31	EPSG:32608	324
12	Old Crow Flats	old_crow_flats	2023-07-15–2023-08-31	EPSG:32607	165
13	Pechora Lowland	pechora_lowland	2023-07-15–2023-08-31	EPSG:32639	366
14	Prudhoe Bay	prudhoe_bay	2023-07-15–2023-08-31	EPSG:32606	257
15	Selawik Lowlands	selawik_lowlands	2023-07-15–2023-08-31	EPSG:32604	178
16	Seward Peninsula	seward_peninsula	2023-07-15–2023-08-31	EPSG:32603	222
17	Taimyr Lowland	taimyr_lowland	2023-07-15–2023-08-31	EPSG:32647	312
18	Teshkepuk	teshekpuk	2023-07-15–2023-08-31	EPSG:32605	305
19	Tuktoyaktuk Peninsula	tuktoyaktuk_peninsula	2023-07-15–2023-08-31	EPSG:32608	174
20	Yamal	yamal	2023-07-15–2023-08-31	EPSG:32643	428
21	Yamal South	yamal_south	2023-07-15–2023-08-31	EPSG:32643	95
22	Yana Indigirka East	yana_indigirka_east	2023-07-15–2023-08-31	EPSG:32655	325
23	Yukon Flats	yukon_flats	2023-07-15–2023-08-31	EPSG:32606	542

3.2 Sentinel-2 input data

The source imagery is Sentinel-2 L2A Surface Reflectance Harmonized data (Drusch et al., 2012) exported through Google
100 Earth Engine (Gorelick et al., 2017) from the COPENICUS/S2_SR_HARMONIZED image collection. For each AOI, images
were filtered to the summer date range listed in Table 2 and by the configured CLOUDY_PIXEL_PERCENTAGE maximum,
usually 35–40 % depending on AOI. Summer windows were chosen to reduce snow and seasonal ice while preserving open-
water conditions.

The archived source workflow masked Sentinel-2 scene-classification (SCL) classes 0, 1, 3, 8, 9, 10, and 11 and applied
105 cloud and cirrus exclusions from the QA60 quality band. These classes flag no data, defective or saturated pixels, cloud
shadow, clouds, cirrus, and snow or ice. The revision implementation explicitly falls back to SCL where QA60 is unavailable.



Live checks on three 2023 scenes found QA60 available and identical valid-mask sums under the archived and revision policies; this result is specific to those checked assets.

110 Surface-reflectance bands were converted with the Sentinel-2 scale factor of 0.0001, and the final source composite was generated as a summer median composite. The median is calculated separately for each band and pixel, so the resulting shoreline describes a seasonal composite rather than a single observation date. The exported source composites contain B2, B3, B4, B8, B11, NDWI, and MNDWI, where $NDWI = (B3 - B8)/(B3 + B8)$ and $MNDWI = (B3 - B11)/(B3 + B11)$. The dimensionless indices compare green reflectance with near-infrared and shortwave-infrared reflectance, respectively. The final images store five bands in the fixed order B2 (blue), B3 (green), B4 (red), B8 (near-infrared), and B11 (shortwave infrared).

115 The revision metadata reconstructs the source-scene candidates for all 30 contributing AOI-dataset configurations. Applying the documented date, AOI and strictly-less-than cloud-percentage filters identifies 873 unique Sentinel-2 assets and 1283 composite-scene links. The tables record product identifiers, UTC sensing times, spacecraft, MGRS tiles and processing base-lines. These are candidate-source reconstructions from the currently accessible archive, not historical export logs or per-pixel median-source attribution. A seasonal median composite has no single acquisition timestamp. Four 32 by 32 pixel replay
120 windows from Mackenzie Delta, Yamal, Teshekpuk and Prudhoe Bay matched all five local image bands exactly; these local checks do not certify every released pixel.

All final image tiles are GeoTIFF rasters. The nominal pixel size is 10 m. Each AOI was exported directly to the UTM zone listed in Table 2 using an Earth Engine `Export.image.toDrive` request with `scale=10` m and the AOI-specific `crs`. This export defines the target analysis grid. The 20 m B11 band is sampled onto that same 10 m grid during export, and no
125 additional post-export reprojection or resampling is applied in the local tiling workflow. The tile-level GeoTIFFs retain the exported transform, projection, and georeferencing information.

4 Dataset generation

4.1 Target definition

The binary target class records visible open-water lake and pond surfaces in the summer Sentinel-2 composite. Thermokarst
130 relevance is assigned separately at tile level using AOI context and water-object morphology. The dataset separates the supervised target mask from a QA/ignore mask (Table 3) so that invalid observations and ambiguous margins are not treated as ordinary non-water background.

4.2 Initial water identification in the AOI composite

Initial water masks were generated at the full-scene scale before tiling (Fig. 3). This design avoids unstable connected-
135 component decisions when lakes, river channels, or coastal water bodies cross tile boundaries. A connected component is a group of water pixels joined along their edges or corners. A pixel enters the initial candidate mask if it exceeds the configured NDWI threshold or, where enabled, the MNDWI threshold, and also satisfies the near-infrared and shortwave-infrared



Table 3. Label and QA-mask policy. The binary target mask remains the primary supervised label; the QA mask provides optional ignore classes for training or evaluation.

Code	Class	Default use	Rule
0	Valid background	Negative class	Dry land, vegetation, bare sediment, non-water shadows, and saturated surfaces without clear standing water.
1	Target water	Positive class	Visible standing water in lake- or pond-like depressions, including dark water and shallow water where the surface expression is coherent. The paired binary mask stores this class as value 255.
2	Invalid observation	Ignore	Residual no-data, strong cloud/snow/ice-like contamination, or image-border artefacts inferred during QA-mask construction.
3	Ambiguous margin	Ignore or sensitivity class	Mixed shoreline pixels, vegetated margins, and wet mud-flats where the evidence for standing water is uncertain.

upper limits. The default index thresholds are 0.05 and reflectance limits are typically 0.15–0.17; Supplement Table S2 lists the settings for each contributing configuration.

140 The code removes groups smaller than the configured minimum, fills holes below the same size threshold and connects small gaps with a one-pixel-radius closing operation. It then identifies water objects by edge-or-corner connectivity. An object is rejected if its area exceeds the configured maximum. Otherwise, it is rejected as elongated only if its area, major-to-minor-axis ratio and eccentricity all reach their respective thresholds. The ratio is measured along the fitted object axes, rather than the bounding box. Supplement Table S2 gives the thresholds. Applying these tests before tiling allows each object to be screened
 145 over the complete composite. The initial screening alone does not establish whether a retained object is a lake, river or coastal water body.

Simple index thresholds can miss dark water and include wet sediment or shallow flats. The refinement in Sect. 4.4 therefore combines the initial mask with additional reflectance tests and pilot-model support. The QA layer is calculated separately afterwards, as described in Sect. 4.6.

150 The water-index baselines reported below use NDWI and MNDWI as defined above. A standard automated water extraction index (AWEI) implementation requires both shortwave-infrared bands, SWIR1 and SWIR2 (Feyisa et al., 2014). With B12 absent from the released five-band stack, the additional AWEI-style baseline uses the single-SWIR diagnostic approximation $4(B3 - B11) - [0.25B8 + 2.75B11]$.



4.3 Cutting and selecting image tiles

155 Each AOI was tiled into non-overlapping 256 by 256 pixel patches with stride 256. At nominal 10 m resolution, each tile covers approximately 2.56 km by 2.56 km, or 6.5536 km². Samples from different processing configurations can nevertheless overlap geographically. The initial selection retains tiles with at least 90 % valid pixels, at least 100 water pixels in the initial mask and a water fraction no greater than 0.30.

The sampling audit reconstructs 7767 initial tiled-scene records across the 30 processed AOI-dataset configurations, of which
 160 7051 passed this initial retention rule and 5513 were included in the public release. The initial-to-release fraction is therefore 0.710, with AOI-level values provided in the release metadata. These retention rules create a water-enriched benchmark with class frequencies distinct from the area-wide natural background distribution.

After final binary-mask and QA-mask reconciliation, 104 tiles have empty final binary masks and 150 tiles have fewer than 100 positive pixels. The initial selection precedes final-mask refinement, so the 100-pixel threshold is not imposed again on
 165 the final mask. The low-water subset manifest indexes these 150 existing core tiles for false-positive analysis without adding duplicate raster samples.

4.4 Combining image evidence to produce the final water mask

The initial water mask provides a first estimate of water extent. Refinement is then applied to each tile using its reflectance values and, where available, a pilot model. This model assigns a number between 0 and 1 to each pixel, with higher values
 170 indicating stronger predicted support for water. Its predictions provide an additional input to mask construction; reflectance and object-shape rules constrain how they are used.

The pilot model was a three-level U-Net trained on the initial 31-tile Kurungnakh subset (23 training, 4 validation, and 4 test tiles; seed 42). Kurungnakh is located in Fig. 1; its samples are shown in Supplement Fig. S8. The original checkpoint, input image–target pairs and fixed split are included in the provenance supplement. None of these 31 tile footprints overlaps
 175 the Mackenzie or Yamal/European Russia test footprints. A probability raster was available during mask assembly for 5482 of the 5513 released tiles; the 31 pilot tiles were assembled without this layer. For these 31 tiles, the processing code supplied zeros in place of the probability image.

At each pixel, refinement first tests for two types of image evidence. Index-based support requires NDWI or MNDWI greater than 0.03 and both B8 and B11 reflectance below 0.17. Dark-water support requires B2 below 0.055, B3 and B4 below 0.075,
 180 B8 below 0.060, and B11 below 0.070. The latter test can retain dark lake interiors that do not pass an index threshold. All reflectance thresholds use scaled, dimensionless surface reflectance.

The code combines the initial mask, these two image tests and the pilot probability using three alternatives. A pixel is included if it satisfies any one of the following conditions:

1. It is water in the initial mask and has index-based support, dark-water support, or a pilot probability of at least 0.35.
- 185 2. Its pilot probability is at least 0.62, both B8 and B11 are below 0.18, and it also has index-based support, dark-water support, or membership in the initial mask.

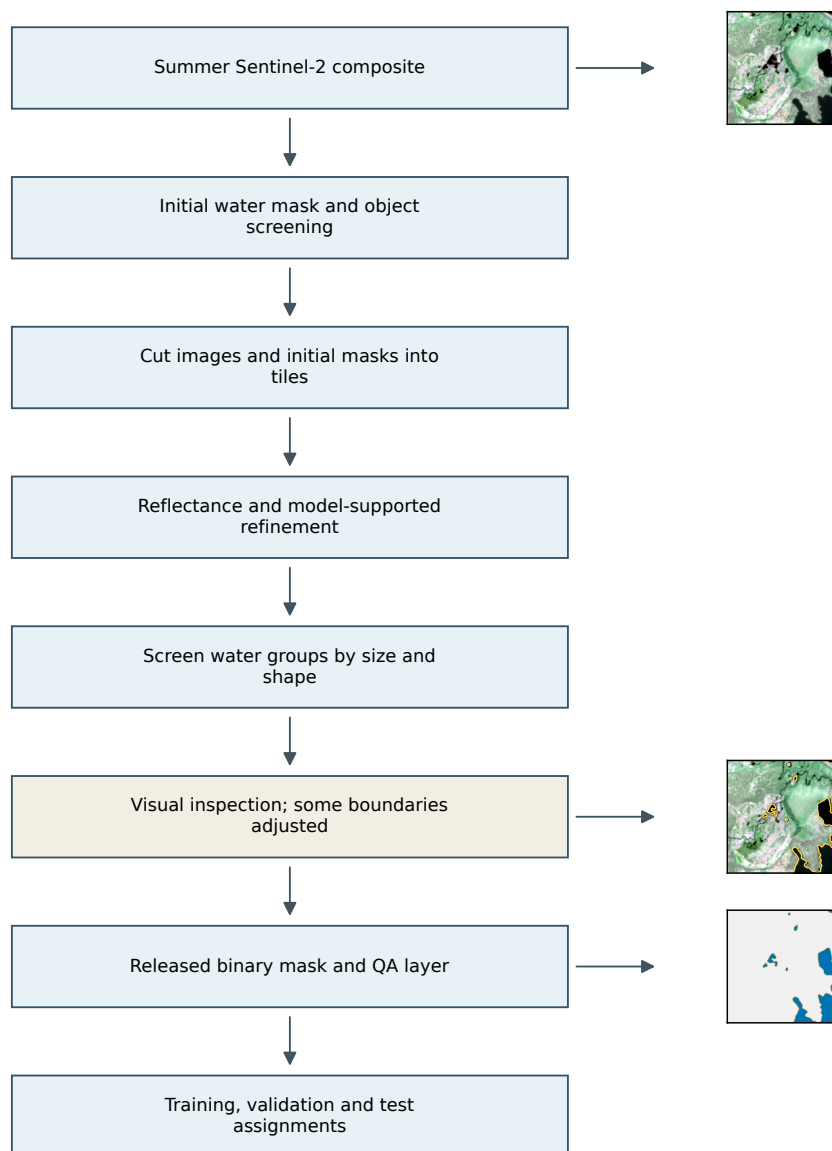


Figure 3. Water-mask production and author review. Index thresholds and initial object screening operate on the AOI composite. Tiled images and masks then undergo automatic refinement and object screening, followed by visual inspection and boundary adjustment in some masks. The released binary mask and QA layer share the image grid. Right-hand panels show an image, a water-boundary overlay and its QA layer. Detailed product comparisons and the QA legend are given in Fig. 5. Contains modified Copernicus Sentinel data (2023).



3. Its pilot probability is at least 0.78 and both B8 and B11 are below 0.22.

Before the tile-level tests, non-finite image values are replaced with zero and all-zero band vectors are excluded. The QA step separately identifies invalid observations from the original image (Sect. 4.6). A closing operation with a one-pixel-radius disk connects small gaps. Enclosed holes smaller than 20 pixels are filled, and water groups smaller than 16 pixels are removed. These small-object and hole filters use four edge-sharing neighbours; the subsequent object-screening step groups pixels sharing either an edge or a corner. The result is a candidate mask for object screening.

Object screening examines each connected group rather than each pixel separately. A group is removed if it meets any one of the following conditions:

1. Its area exceeds 34 % of the tile.
2. Its median B8 or median B11 reflectance exceeds 0.24.
3. Its eccentricity exceeds 0.992, its major-to-minor-axis ratio exceeds 9, and its area exceeds 120 pixels.
4. It touches a tile edge, occupies more than 18 % of the tile, and has an axis ratio above 4.
5. Its solidity is below 0.18, its area exceeds 200 pixels, and its median pilot probability is below 0.65.

All conditions joined by “and” within an item must be met together. Eccentricity measures how closely an object approaches a line, and solidity is its area divided by its convex-hull area. These are screening rules for large, elongated or fragmented candidates, rather than a geomorphic identification of rivers.

A group that passes these rejection rules must retain at least one of the following sources of support:

1. Positive median NDWI or MNDWI, with both median B8 and median B11 reflectance below 0.18.
2. Dark-water reflectance, with median B8 below 0.065 and median B11 below 0.075.
3. A median pilot probability of at least 0.50.
4. Initial-mask coverage of at least 25 % of the group area.

A group with less than 25 % initial-mask coverage is additionally rejected if its median pilot probability is below 0.42 and it lacks the dark-water support just defined. The same gap-closing, hole-filling and minimum-size operations are repeated after screening to produce the automatically refined binary mask. Initial object filtering therefore occurs on the AOI composite, while the final automatic checks operate on tiles.

The automatic stage produces a binary water mask for visual inspection. The 5513 resulting masks are stored as supervised targets. The manifest records the split, AOI, year, source configuration, QA-mask path, label quality, processing fields, and package-management fields for each tile.

The manifest supports both standard model development and stricter geographic evaluation. Users can construct AOI-, region-, or source-configuration holdout splits without changing the data payload. The supplied geographic-membership table records the training, validation and test roles of the two reported holdout experiments.



Table 4. Target construction and uncertainty information.

Item	Dataset record
Target class	Visible lake and pond water-surface masks, with rivers, coastal water, invalid observations, and ambiguous margins handled through the target definition and QA policy in Table 3.
Construction inputs	Sentinel-2 RGB composites, water-index layers, dark-water screening, full-scene object context, pilot-model probability support, and image-derived contextual diagnostics.
Mask production and review	Image-based rules and pilot-model support were followed by image-by-image visual inspection in QGIS, including boundary adjustments in some masks. Pilot-model probabilities supported mask assembly for 5482 tiles; 31 pilot tiles did not use this layer.
Mask-correction targets	Dark-water infill, wet-sediment rejection, edge-artifact removal, channel/coastal-water filtering, and shoreline-margin classification.
QA and interpretation metadata	QA codes identify invalid observations and uncertain margins from the imagery and final mask. Geomorphic-context fields support subsequent tile-level interpretation after water-mask and QA generation (Sect. 4.7).

4.5 Visual inspection and manual boundary adjustments

The authors used QGIS to inspect image–mask pairs individually. The reference for these adjustments is the visible water surface in the summer composite, following the target definition in Sect. 4.1. B4–B3–B2 RGB composites and enlarged shoreline views were used to compare mask boundaries with visible standing water, adjoining vegetation and exposed sediment. Selected boundaries were adjusted to follow the visible water extent. The output mask retains the image CRS, extent, 10 m pixel spacing and 256 by 256 dimensions, with 255 for water and 0 for background.

Invalid observations and ambiguous margins are handled separately by the QA procedure in Sect. 4.6. This boundary-review step is distinct from the automatic thermokarst-relevance classification in Sect. 4.7. Figure 4 illustrates the image, boundary overlay and enlarged view used to explain this inspection.

4.6 Quality-assurance mask and uncertain margins

The QA mask is generated after the final binary mask on the same grid to identify invalid observations and uncertain background margins. Table 3 defines its four stored codes.

First, invalid observations are identified where any band is non-finite, all five bands have absolute reflectance at most 10^{-8} , or mean B2–B4 reflectance exceeds 0.35 together with B8 above 0.22. The brightness test flags residual high-albedo contamination; it is not a separate cloud, snow or ice classification.

Second, potentially uncertain background pixels are selected in two ways. The first is the one-pixel-wide outer neighbour band of the binary water mask, constructed with the four edge-sharing neighbours. The second is background with B8 and

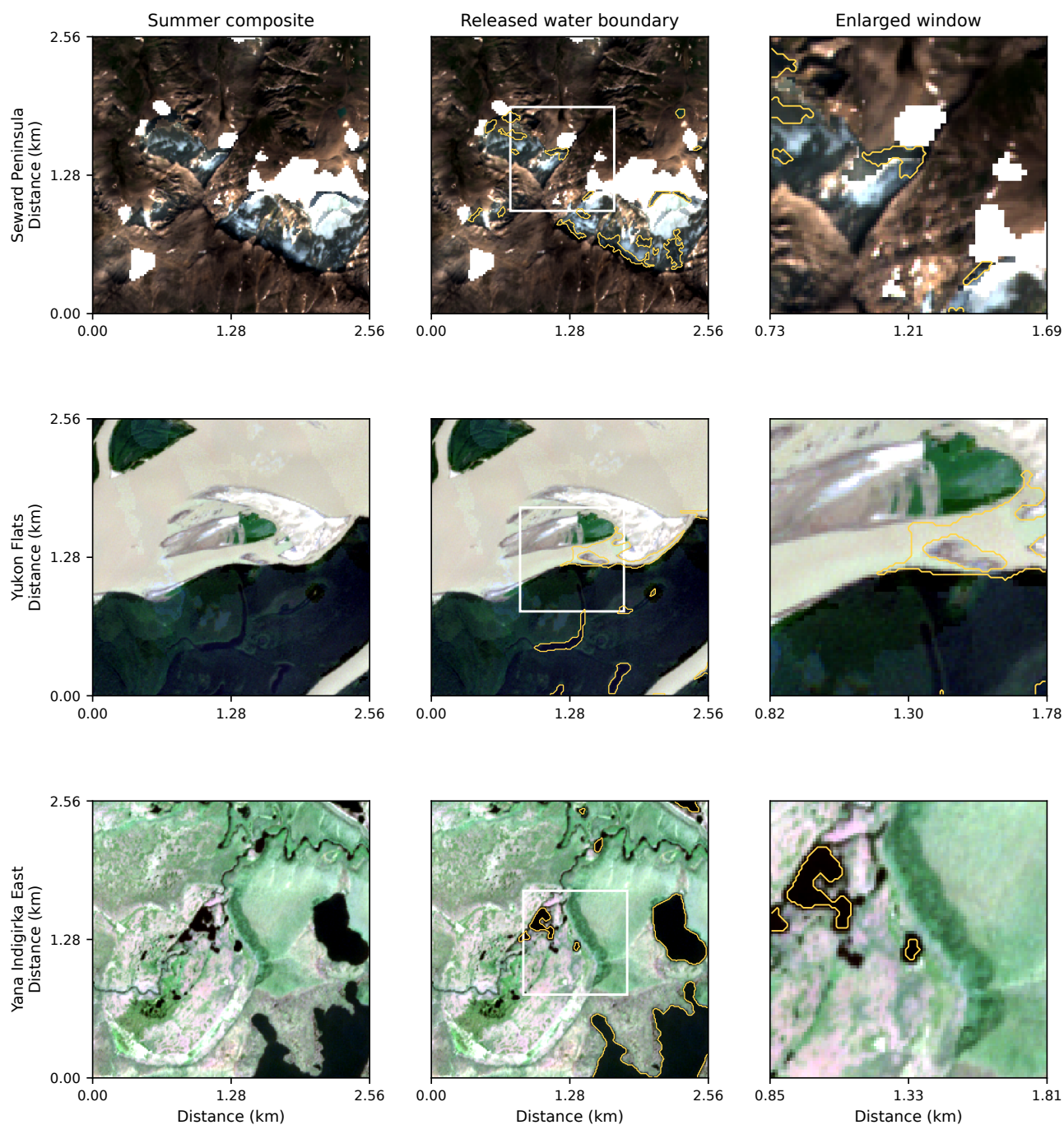


Figure 4. Image–mask correspondence for visual inspection. Each row shows an RGB tile, its water-mask boundary, and an enlarged view of the outlined window. Yellow lines identify the water boundary; coordinates show distances in kilometres. These views illustrate the spatial correspondence examined during review. Contains modified Copernicus Sentinel data (2023).



235 B11 below 0.25 and either NDWI or MNDWI strictly between -0.03 and 0.08 . This second test marks locations with a weak water-index response, including possible wet or vegetated margins away from the immediate shoreline.

The QA array begins with code 0 for background. The union of the two uncertainty tests receives code 3, binary-mask water then receives code 1, and invalid observations finally receive code 2. This assignment order means that invalid observations override every other class and binary water overrides the uncertainty flag. Strict evaluation excludes code 2 only; QA-aware
240 evaluation also excludes code 3. Differences between these evaluations therefore arise from the pixels included in the comparison, rather than from a different binary target mask.

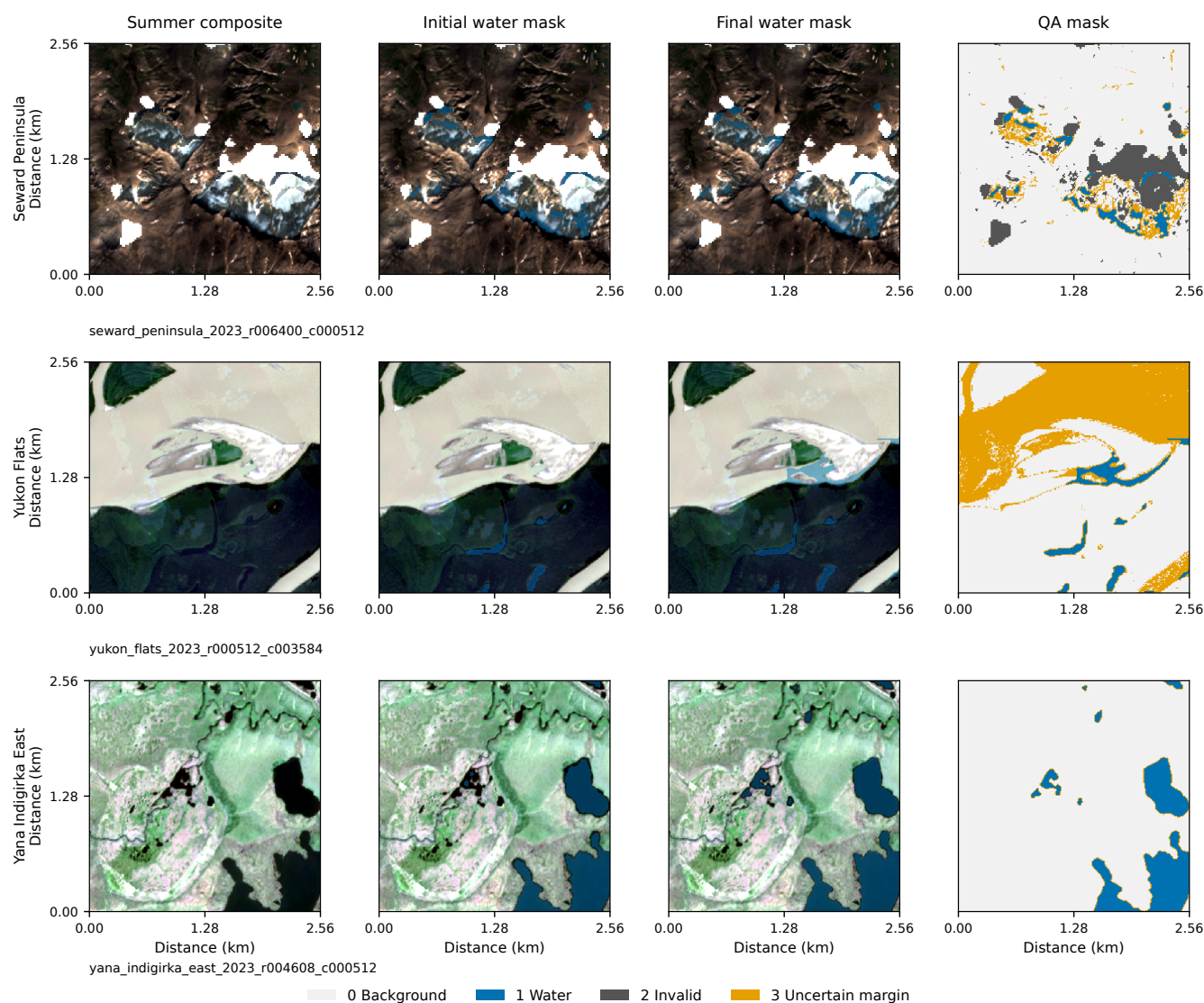


Figure 5. Stored processing products for three sample tiles. Each row shows the same location in the summer RGB composite, archived initial water mask, released final water mask and QA layer. Blue overlays identify water in the initial and final masks; QA codes use the legend below the panels. Changes between the two water masks result from the combination and screening rules in Sect. 4.4. The QA layer separately flags invalid observations and uncertain background margins under Sect. 4.6. Tile identifiers are shown beneath the image panels, and axes give distances in kilometres. Enlarged comparisons are provided in Supplement Figs. S24–S26. Contains modified Copernicus Sentinel data (2023).



4.7 Rule-based thermokarst-relevance classification

To make the dataset more directly usable for thermokarst lake and pond studies, each tile is assigned a tile-level geomorphic interpretation in the release metadata. This interpretation is separate from the supervised water mask. Thermokarst relevance is calculated after the water and QA masks are available. The classification is automated using the rules below and the AOI categories listed in Supplement Table S1.

The first step measures connected water groups using edge-or-corner connectivity. A group is classified as elongated if it has at least 100 pixels and either a bounding-box aspect ratio of at least 5 or eccentricity of at least 0.98. A compact group has at least 10 pixels, is not elongated and does not touch a tile edge. This definition differs from the axis-ratio rejection test used during final-mask construction. Edge-contact fraction is the fraction of all water pixels belonging to any edge-touching group, rather than the fraction of shoreline lying on the edge. Elongated-water fraction is defined analogously. Largest-group fraction uses the whole tile as its denominator.

The second step adds one risk point for each satisfied item below:

1. The AOI belongs to a coastal or deltaic/channelized category in Supplement Table S1.
2. Binary water occupies more than 0.18 of the tile.
3. The largest water group occupies more than 0.12 of the tile.
4. The edge-contact fraction exceeds 0.35 and QA target water is present.
5. The elongated-water fraction exceeds 0.35.
6. QA code 3 occupies more than 0.05 of the tile.
7. The AOI is Bykovsky, Kurungnakh or Lena Delta West, the configured small-sample set.

Risk is low at 0–1 points, moderate at 2–3 points and high at 4 or more points. Coastal and deltaic/channelized AOI categories are mutually exclusive, so AOI context contributes at most one point.

The final step applies the conditions in Table 5 in order and stops at the first match. The pixel count used for the first condition is QA code 1, whereas water fraction and shape measurements come from the binary mask. Across the 5513 released tiles, 2374 tiles are classified as high-confidence thermokarst-like lake or pond contexts, 2554 as probable thermokarst-related contexts, 335 as origin-unassigned or low-confidence, and 250 as likely non-target or ambiguous water contexts. These assignments are rule-based screening metadata rather than object-level confirmation of lake origin.

These thresholds are operational screening choices: elongated, edge-touching and locally dominant water groups warrant more caution than isolated compact groups. They are not calibrated probabilities of thermokarst origin, and the term “high-confidence” in the stored class name refers to satisfying these screening criteria.



Table 5. Ordered decisions used to assign tile-level thermokarst relevance. Conditions are tested from top to bottom.

Code	Assigned context	First matching condition
0	Unassigned or near-empty	Fewer than 100 QA target-water pixels.
−1	Likely non-target or ambiguous	High risk and either largest-group fraction > 0.12 , elongated-water fraction > 0.35 , or edge-contact fraction > 0.50 .
2	High-confidence thermokarst-like	At least two compact groups and low risk.
1	Probable thermokarst-related	At least one compact group and low or moderate risk.
1	Probable thermokarst-related	At least two groups, binary water fraction ≤ 0.18 , and moderate risk.
0	Unassigned or low confidence	No preceding condition is met.

Table 6. Split definitions distributed with the dataset package.

Split field or file	Primary use	Interpretation
<code>split</code> ; standard-grid file	Standard model development	Within-AOI train/validation/test benchmark; geographic transfer is evaluated with the dedicated holdout files.
<code>aoi_holdout_fold</code> ; AOI file	Leave-one-AOI-out testing	All tiles from one AOI can be withheld to test geographic transfer to that AOI.
<code>region_holdout_fold</code> ; regional file	Leave-one-region-group-out testing	AOIs are grouped into Western Arctic, Yamal and European Russia, and East Siberia and Lena–Kolyma region groups.

4.8 Training, validation and test assignments

The dataset is split into train, validation, and test directories. Split assignment was built from AOI-level processing configurations using 6 by 6 spatial grid cells, validation and test fractions of 0.15 each, and random seed 42. Figure 2 and Supplement Figs. S1–S23 show the resulting sample assignments geographically. A spatial audit of this standard split found 1183 cross-split adjacent tile pairs, including 530 pairs where water pixels touch the shared tile edge. The standard split therefore serves as the within-AOI development benchmark.

The original AOI-name holdouts were subsequently checked against raster footprints. The audit found 202 positive-area overlapping tile pairs, including 99 crossing the legacy standard splits. In the original Mackenzie Delta experiment, 154 test–development tile pairs overlapped despite having different AOI identifiers. The Yamal and European Russia holdout had no detected test–development footprint overlap. We therefore retain the legacy results as development diagnostics and provide a newly purged Mackenzie experiment below. The manifest also includes explicit AOI-holdout and regional-holdout fields (Table 6) so that users can construct additional geographic experiments without redefining the data payload.



5 Data records

5.1 Images and matching masks

285 The model-ready dataset has a simple directory structure:

```
arctic_lake_pond_segmentation_dataset/  
  manifest.csv  
  README_CN.md  README_EN.md  
  summary/  provenance/  spatial/  
290 splits/geographic_membership.csv  
  pilot_model/  hard_negative/subset_manifest.csv  
  train/images/  train/masks/  train/qa_masks/  
  val/images/  val/masks/  val/qa_masks/  
  test/images/  test/masks/  test/qa_masks/
```

295 Each image tile has one paired binary mask tile and one paired QA mask tile with the same file stem. Each binary mask is a single-band GeoTIFF with values 0 for background and 255 for target water. Values greater than zero can be converted to the positive class in machine-learning code. Each QA mask is a single-band GeoTIFF using the codes in Table 3; QA codes 2 and 3 can be excluded from loss functions and metrics for ignore-aware evaluation.

5.2 Package metadata

300 The package pairs the core rasters with a tile manifest, source-scene tables, processing configurations, spatial audits, geographic split membership and derived statistics. Supplementary provenance includes the pilot checkpoint and its original 31-tile input dataset. Table 7 lists the principal records. The downloadable ZIP archives group multiple GeoTIFF files for transfer; they are packaging units, whereas a tile is an individual image sample. Checksum tables allow users to verify file integrity after download.

305 5.3 Dataset statistics

Table 8 summarizes the train-validation-test split. The complete AOI and split summaries are provided in the package metadata. Figure 6 shows representative image-mask overlays, and Fig. 7 summarizes AOI tile counts and positive-mask fractions.

The positive class is intentionally sparse relative to the background class. Figure 8 shows the per-tile target-water fraction distribution for the supervised masks. The dataset contains 104 empty-mask tiles, and these are retained intentionally as
 310 negative or near-negative samples.

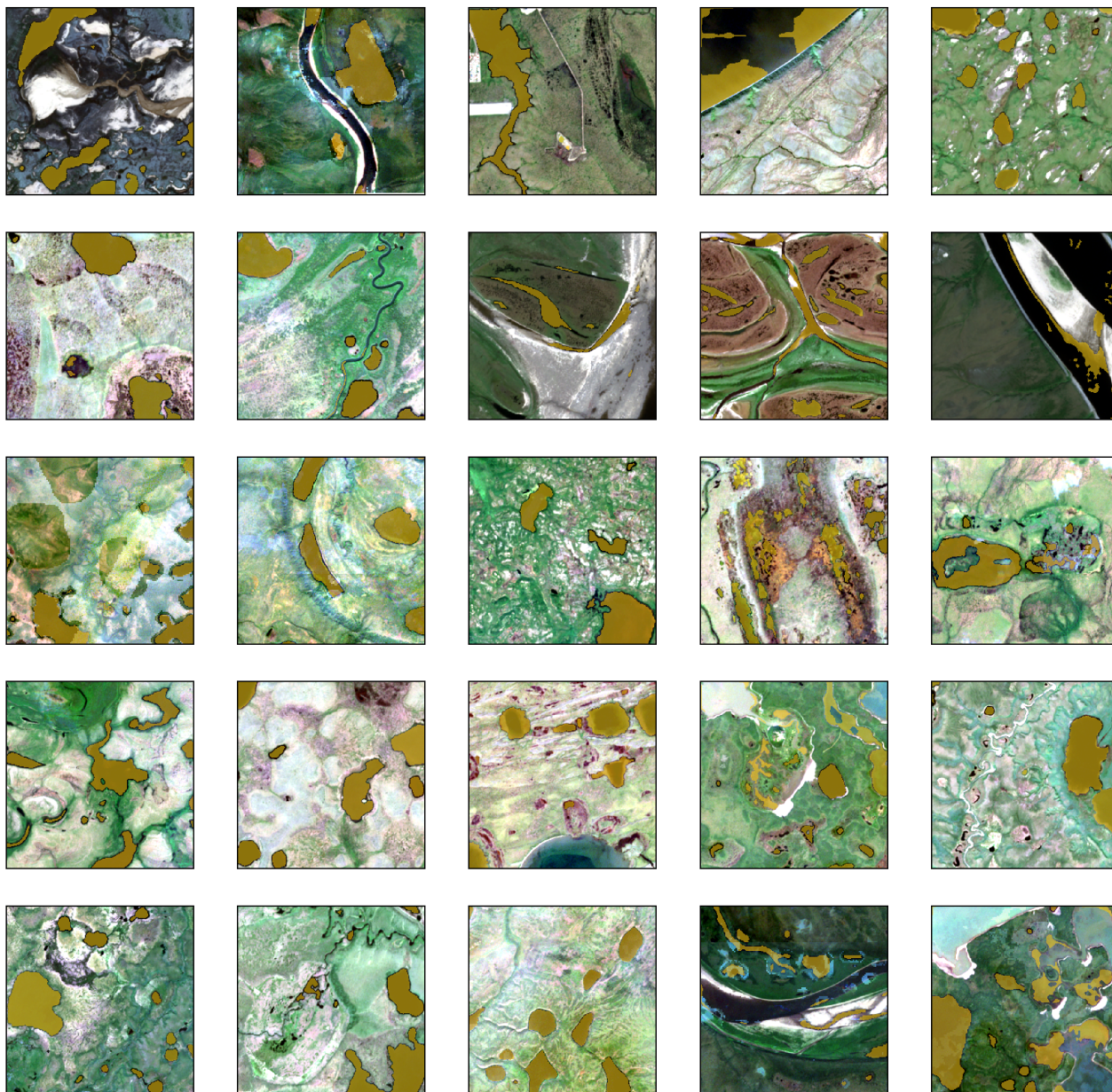


Figure 6. Example image-mask overlays from the study AOIs. RGB composites use Sentinel-2 bands B4, B3, and B2; yellow overlay indicates positive target-water mask pixels. Contains modified Copernicus Sentinel data (2023).



Table 7. Core data records and metadata files.

Record	Description
Image, target and QA tiles	5513 georeferenced triplets; five-band imagery, binary supervision targets and multi-class QA.
<code>manifest.csv</code>	Tile identifiers, relative raster paths, source configuration, split, QA counts and geographic context.
<code>provenance/</code> and <code>configs/</code>	Candidate scene identifiers and acquisition times, composite associations and processing parameters.
<code>spatial/</code> and <code>splits/</code>	Actual tile footprints, overlap audit and geographic experiment membership.
<code>pilot_model/</code>	Original pilot checkpoint, 31 source image/target pairs, fixed split, logs and geographic-separation audit.
<code>summary/</code> and <code>legacy_baseline/</code>	Derived sample statistics and historical target-agreement diagnostics.
<code>hard_negative/</code>	Index of 150 existing core tiles containing fewer than 100 final target-water pixels.
<code>aoi_summary.csv</code>	AOI settings and sample counts.
<code>data_dictionary.csv</code> , JSON schemas	Field definitions, dataset description and raster encoding.
Download checksums	SHA256 values for verifying downloaded ZIP archives and metadata files.

Table 8. Split-level summary. Target-water pixels are mask pixels with value greater than zero.

Split	Tiles	Empty masks	Target-water pixels	Target-water fraction
Train	3982	78	20,797,879	0.0797
Validation	739	12	4,104,551	0.0848
Test	792	14	4,239,165	0.0817
Total	5513	104	29,141,595	0.0807

6 Technical validation

Validation addresses three questions: whether the files and grids are consistent; whether documented processing can reproduce selected stored products; and how model agreement varies between the sampled landscapes. Pixel intersection-over-union (IoU) is the area shared by predicted and target water divided by their combined area. Dice is twice their shared area divided by the sum of their areas. Precision reports the fraction of predicted water that agrees with target water, while recall reports the fraction of target water recovered. Boundary F1 summarises precision and recall of boundary pixels within a stated distance tolerance. These quantities describe agreement with the released masks.



Table 9. AOI-level target-water statistics.

AOI	Tiles	Empty masks	Target-water fraction
Anabar Lowland	40	0	0.0283
Anderson Plain	441	0	0.0522
Barrow Coastal Plain	66	0	0.1030
Bykovsky	32	0	0.0457
Gyda	416	0	0.0761
Indigirka Lowland	312	7	0.0959
Kolyma Lowland	307	14	0.1034
Kurungnakh	34	4	0.0117
Lena Delta East	165	0	0.1333
Lena Delta West	7	0	0.0108
Mackenzie Delta	324	4	0.1232
Old Crow Flats	165	0	0.0549
Pechora Lowland	366	0	0.0626
Prudhoe Bay	257	0	0.1395
Selawik Lowlands	178	0	0.0757
Seward Peninsula	222	0	0.0613
Taimyr Lowland	312	0	0.0599
Teshekpuk	305	0	0.0999
Tuktoyaktuk Peninsula	174	0	0.1435
Yamal	428	75	0.0345
Yamal South	95	0	0.1172
Yana Indigirka East	325	0	0.0915
Yukon Flats	542	0	0.0589

6.1 Structural validation

The dataset contains 5513 images, 5513 binary water masks and 5513 QA masks, giving 16539 GeoTIFF files. The accompanying tables record sample assignments, geomorphic interpretation, spatial audits and baseline metrics. Each row in `manifest.csv` pairs one image tile, one binary mask tile, and one QA mask tile, and the manifest records relative paths, split assignments, AOI identifiers, source settings, QA-pixel counts and thermokarst-relevance interpretation fields. Summary tables were computed from the masks and QA masks.

Across all tiles, the QA layer marks 713,875 pixels (0.20 %) as invalid observations and 5,204,429 pixels (1.44 %) as ambiguous margins. A complete raster audit found 705,255 non-finite image pixels in 226 tiles; every non-finite pixel coincides with QA code 2, and no infinite values were found. The QA target-water count is 29,141,409 pixels, which differs slightly from

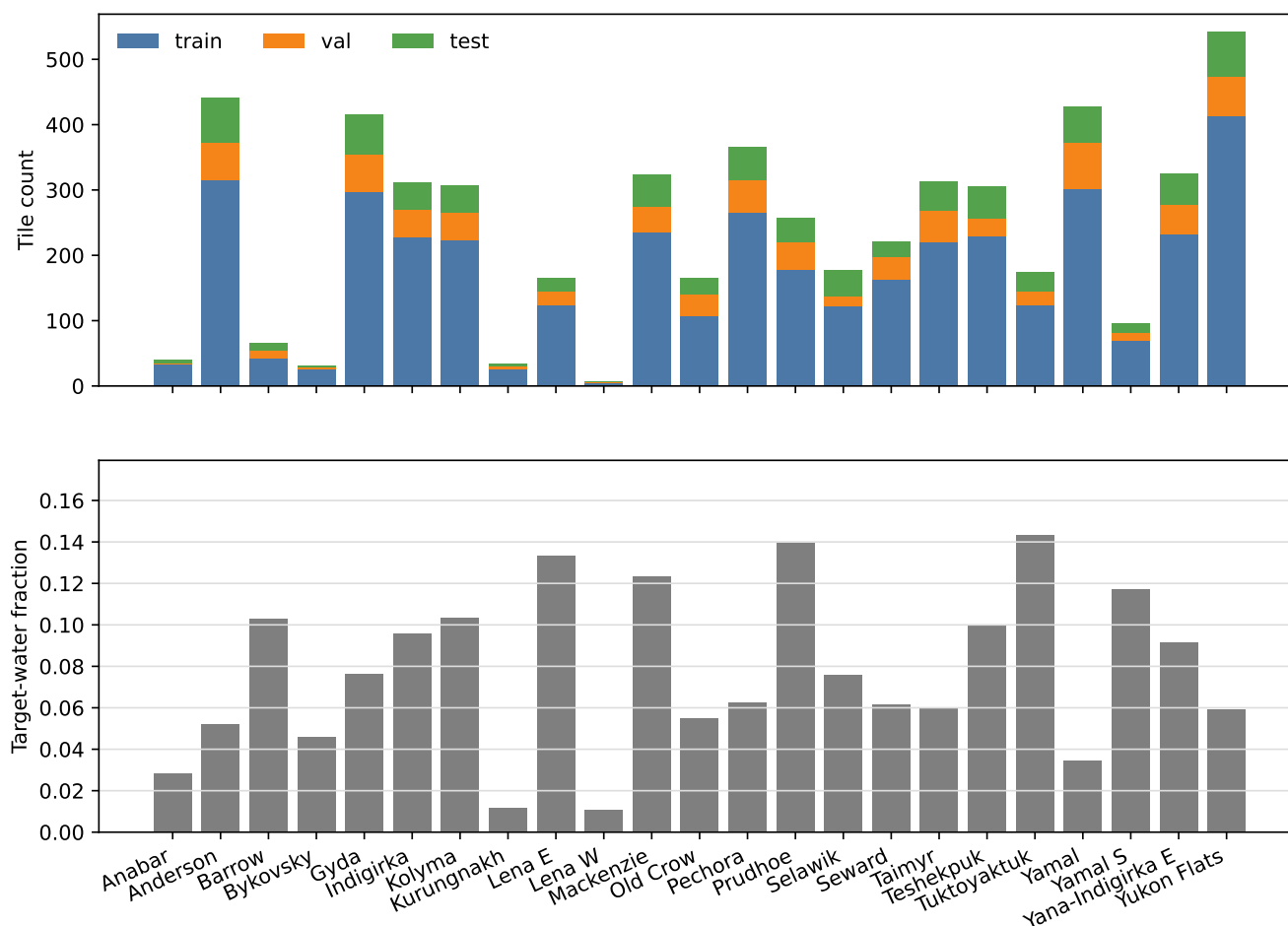


Figure 7. Dataset statistics by AOI. Upper panel: tile counts by split. Lower panel: target-water fraction by AOI.

the binary-mask positive-pixel count because invalid-observation screening can override target pixels in the QA layer. The QA layer does not contain an active river, ocean, or excluded-water class in the present package.

6.2 Illustrated checks of masks and QA

330 Figure 5 presents three stored initial masks alongside the released final masks and QA layers. In the Seward Peninsula example, refinement removes 38 initially positive pixels; the QA layer separately flags 6938 invalid-observation pixels. In the Yukon Flats example, refinement adds 1366 pixels and removes one, while QA marks 23649 uncertain-margin pixels. In the Yana Indigirka East example, refinement adds 748 pixels and QA marks 726 uncertain-margin pixels. Thus, QA can flag uncertain background beyond the immediate shoreline.

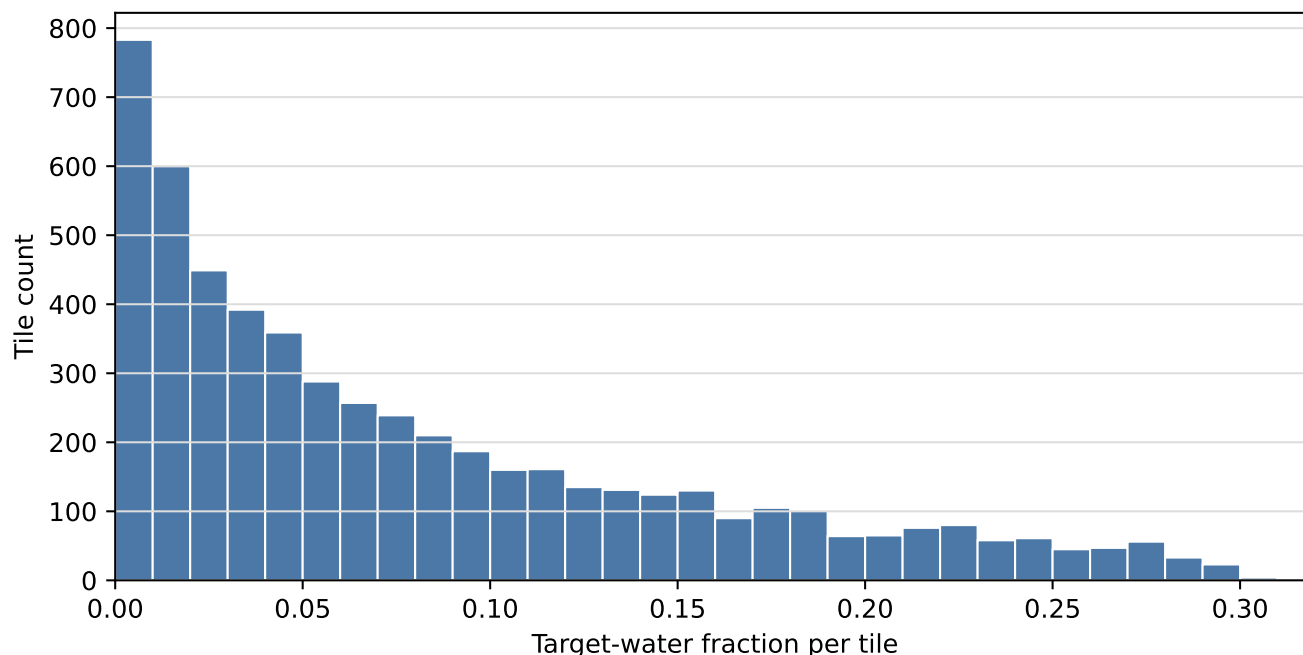


Figure 8. Distribution of target-water fraction per tile. The right-skewed distribution reflects the sparse positive class and the tile-retention upper bound of 0.30 described in the Methods section.

335 We checked each illustrated image against the corresponding window in the original AOI composite, then reran automatic mask assembly using the stored initial mask and pilot-probability image. The image values, automatically assembled binary mask and recalculated QA layer matched their released counterparts exactly in all three cases. Supplement Table S4 records the intermediate water-pixel counts.

6.3 Geometry-audited Mackenzie experiment

340 The corrected experiment removes development tiles intersecting the held-out Mackenzie footprints or their 2560 m buffer, and additionally purges training tiles against buffered validation footprints. Buffer construction uses the Mackenzie UTM coordinate system; the distance is an operational separation rule, not a universal spatial-correlation threshold. The resulting split contains 3145 training, 695 validation and 324 test tiles. A subsequent check against the full raster-footprint overlap list found no positive-area cross-role overlaps. Split counts for 0, 1280 and 2560 m buffers are supplied for sensitivity inspection.

345 The U-Net was freshly trained for 10 epochs with seed 20260919, batch size 8, AdamW, learning rate 0.001, positive-class weight 10, and masked binary cross-entropy (BCE) plus Dice loss. Both training and validation excluded QA codes 2 and 3, and checkpoint selection used pooled validation IoU. This differs from the legacy training implementation, which loaded binary targets without the QA ignore layer. Consequently, the new-versus-old score difference is not a controlled estimate of the effect of leakage alone.



350 Against the released supervision targets, strict IoU was 0.622 and QA-aware IoU was 0.659, with QA-aware Dice 0.794, precision 0.665 and recall 0.986. A 1000-replicate bootstrap over 30 spatial clusters on a 10.24 km grid gave a QA-aware IoU interval of 0.587–0.728. These intervals describe target agreement within the represented sample, not independently verified shoreline accuracy.

6.4 Overlap with an existing Alaska product

355 We compared all 628 released tiles from Barrow Coastal Plain, Teshekpuk and Prudhoe Bay with the version-pinned ALPOD Zenodo rasters (DOI 10.5281/zenodo.14199365). Downloaded files were checked against their published MD5 checksums and reprojected onto the corresponding image grids. The downloaded rasters have positive native values exceeding 100 despite a percentage description and do not supply an explicit scale factor. We therefore report the overlap of nonzero mapped support. ALPOD zero/nodata is interpreted as absence of positive mapped support, not as independently verified dry land.

360 Under the QA-aware policy, nonzero-support IoU was 0.808 for Barrow Coastal Plain, 0.249 for Teshekpuk and 0.851 for Prudhoe Bay. The much larger discrepancy at Teshekpuk identifies a priority for independent review. These are cross-product overlap diagnostics, not accuracy estimates: the 2016–2021 occurrence product and our 2023 seasonal-composite targets differ in temporal support and construction. The published sample is water-enriched, and these results cannot be extrapolated directly to area-wide lake abundance.

365 6.5 Legacy threshold, model, and geographic baselines

A U-Net baseline (Ronneberger et al., 2015) was trained using the provided train-validation-test split and two geographic holdout splits. The model uses the five reflectance bands to predict a water probability for each pixel. Supplement Table S3 gives the network, input preparation, training settings and hardware; the buffered Mackenzie rerun is described separately above.

370 Threshold baselines and the U-Net baseline are evaluated under two policies: a strict policy that excludes only invalid-observation pixels (QA code 2), and a QA-aware policy that excludes invalid-observation and ambiguous-margin pixels (QA codes 2 and 3). The standard train/validation/test split serves as the within-AOI technical benchmark. Geographic transfer is evaluated separately with a leave-one-AOI-out Mackenzie Delta experiment and a leave-one-region-out Yamal and European Russia experiment, where the held-out AOI or region is absent from training.

375 Simple index thresholds were substantially weaker than the trained segmentation baseline on the standard split (Table 10). The best threshold-style method, the AOI-specific spectral rules used during mask construction, reached an IoU of 0.5895 and Dice score of 0.7418 on the QA-aware test pixels. The U-Net baseline reached an IoU of 0.8489, Dice score of 0.9183, precision of 0.8565, and recall of 0.9897 under the same QA-aware policy. These aggregate pixel metrics indicate that the image and mask pairs are aligned and learnable. Because pilot-model probabilities contributed to most target masks, these
 380 model scores measure agreement with the released targets; external shoreline accuracy remains a separate validation task.

The QA-aware policy removes 769,803 pixels (1.48 %) from the standard test split, 494,209 pixels (2.33 %) from the Mackenzie Delta holdout, and 824,884 pixels (0.96 %) from the Yamal and European Russia holdout. Compared with the strict



Table 10. QA-aware standard test-set baseline comparison. The QA-aware policy excludes pixels with QA codes 2 and 3 from metric computation. The AWEI-style row is a single-SWIR diagnostic approximation because the released image stack does not include B12.

Method	Input or rule		IoU	Dice	Precision	Recall
AOI-specific spectral rules	NDWI/MNDWI	+	0.5895	0.7418	0.6353	0.8911
	NIR/SWIR constraints					
AWEI-style single-SWIR > 0	B3, B8, B11		0.4527	0.6232	0.5773	0.6770
MNDWI > 0	B3, B11		0.5663	0.7231	0.6266	0.8549
MNDWI > 0.05	B3, B11		0.5499	0.7096	0.6197	0.8300
NDWI > 0	B3, B8		0.5467	0.7069	0.6208	0.8208
NDWI > 0.05	B3, B8		0.5151	0.6800	0.6067	0.7734
U-Net baseline	B2, B3, B4, B8, B11		0.8489	0.9183	0.8565	0.9897

Table 11. Legacy strict and QA-aware U-Net target-agreement results. The standard split and original Mackenzie holdout contain footprint overlap; these rows are retained for provenance, not as independent evaluation. Strict evaluation excludes only QA code 2; QA-aware evaluation excludes QA codes 2 and 3. Boundary F1 is reported with a two-pixel tolerance.

Experiment	Tiles	Strict excl. %	QA excl. %	Strict IoU	QA IoU	QA Dice	BF1 _{2px}	Obj. P	Obj. R	FP obj. km ⁻²
Standard within-AOI test	792	0.04	1.48	0.809	0.849	0.918	0.904	0.311	0.997	3.72
Leave-one-AOI-out: Mackenzie Delta	324	0.38	2.33	0.733	0.757	0.862	0.786	0.337	0.990	3.12
Leave-one-region-out: Yamal and European Russia	1305	0.13	0.96	0.921	0.936	0.967	0.967	0.591	0.997	0.90

policy, excluding ambiguous margins increases pixel IoU in all three experiments, but it also changes object-level precision because fragmented predictions around ignored shoreline pixels can be counted as separate false-positive objects. Object and boundary metrics are therefore reported alongside pixel recall.

AOI-balanced summaries are reported in addition to micro-averaged pixel metrics. For the standard test split, the QA-aware macro-average across the 19 AOIs with at least 10 test tiles is IoU 0.895 (standard deviation 0.117, median 0.937, interquartile range 0.895–0.972), Dice 0.940, precision 0.902, recall 0.991, object precision 0.371, object recall 0.997, and Boundary F1 at two pixels 0.906. The principal macro-average comprises these 19 AOIs; four smaller AOIs are reported separately as diagnostic cases. For the Yamal and European Russia regional holdout, the QA-aware macro-average over the four held-out AOIs is IoU 0.927 (standard deviation 0.055, median 0.949, interquartile range 0.914–0.962), object precision 0.610, object recall 0.996, and Boundary F1 at two pixels 0.963.

The low-water subset indexed from the core dataset contains 150 tiles with fewer than 100 target-water pixels in the final mask, including 104 empty-target tiles and 46 near-empty-target tiles. Nineteen of these tiles belong to the standard test split. On this hard-negative test subset, the U-Net baseline has a strict pixel false-positive rate of 0.00154 and 0.249 false-positive objects km⁻²; under the QA-aware policy, the corresponding values are 0.00147 and 0.289 false-positive objects km⁻².



Table 12. Strict and QA-aware U-Net performance by AOI on the standard test split. Status B marks AOIs included in the principal macro-average; status D marks AOIs with fewer than 10 test tiles and diagnostic interpretation. Confidence intervals are omitted because cluster-aware spatial resampling was not applied to the AOI rows.

AOI	Test tiles	Strict IoU	QA IoU	Obj. P	Obj. R	FP obj. km ⁻²	Status
Kurungnakh	4	0.097	0.105	0.098	0.952	8.97	D
Bykovsky	3	0.136	0.140	0.136	1.000	5.88	D
Kolyma Lowland	42	0.588	0.627	0.115	1.000	12.19	B
Teshkepuk	49	0.603	0.632	0.212	1.000	9.68	B
Mackenzie Delta	49	0.697	0.722	0.328	0.998	3.04	B
Indigirka Lowland	42	0.742	0.773	0.151	1.000	4.79	B
Anderson Plain	68	0.870	0.894	0.621	1.000	0.86	B
Lena Delta East	20	0.823	0.897	0.213	0.975	10.71	B
Yamal South	14	0.868	0.901	0.467	1.000	2.71	B
Tuktoyaktuk Peninsula	29	0.889	0.916	0.391	0.997	2.78	B
Old Crow Flats	25	0.897	0.928	0.328	1.000	1.51	B
Yukon Flats	69	0.829	0.937	0.225	0.990	4.27	B
Yamal	55	0.921	0.963	0.286	0.996	1.69	B
Anabar Lowland	5	0.901	0.964	0.312	1.000	1.97	D
Selawik Lowlands	40	0.916	0.965	0.423	1.000	2.40	B
Prudhoe Bay	37	0.924	0.965	0.456	0.999	6.98	B
Seward Peninsula	24	0.926	0.972	0.442	0.990	3.46	B
Barrow Coastal Plain	12	0.935	0.973	0.326	1.000	3.93	B
Yana Indigirka East	48	0.938	0.973	0.324	1.000	2.36	B
Taimyr Lowland	44	0.945	0.984	0.450	1.000	1.24	B
Gyda	61	0.963	0.986	0.712	0.995	0.72	B
Pechora Lowland	51	0.954	0.991	0.577	0.997	0.83	B
Lena Delta West	1	0.983	0.995	1.000	1.000	0.00	D

Per-AOI results show stronger heterogeneity than the aggregate metric alone (Table 12). Several AOIs, including Gyda, Pechora Lowland, Taimyr Lowland, Barrow Coastal Plain, Prudhoe Bay, and Yana-Indigirka East, show high pixel-level IoU, while Bykovsky and Kurungnakh have low per-AOI scores in the present standard test split. AOIs with fewer than 10 test tiles are reported as diagnostic cases. Kolyma Lowland, Mackenzie Delta, and Teshkepuk also have lower scores than the aggregate metric despite larger test counts. These cases identify priorities for targeted audit, broader hard-negative sampling, and AOI-balanced validation. Figure 9 pairs dataset coverage with target-water fraction, test-tile counts, pixel IoU, and Dice score for each AOI.

6.6 Limitations and uncertainty

The release captures visible summer water surfaces in selected 2023 Sentinel-2 composites. Its labels describe the observed water surface and tile-level thermokarst relevance; they do not resolve seasonal dynamics or establish object-level thermokarst origin. The QA layer is image-derived, tile retention creates a water-enriched sample distribution, and formal inter-annotator agreement is not included. The standard split is a within-AOI benchmark, while the reported geographic holdouts provide initial

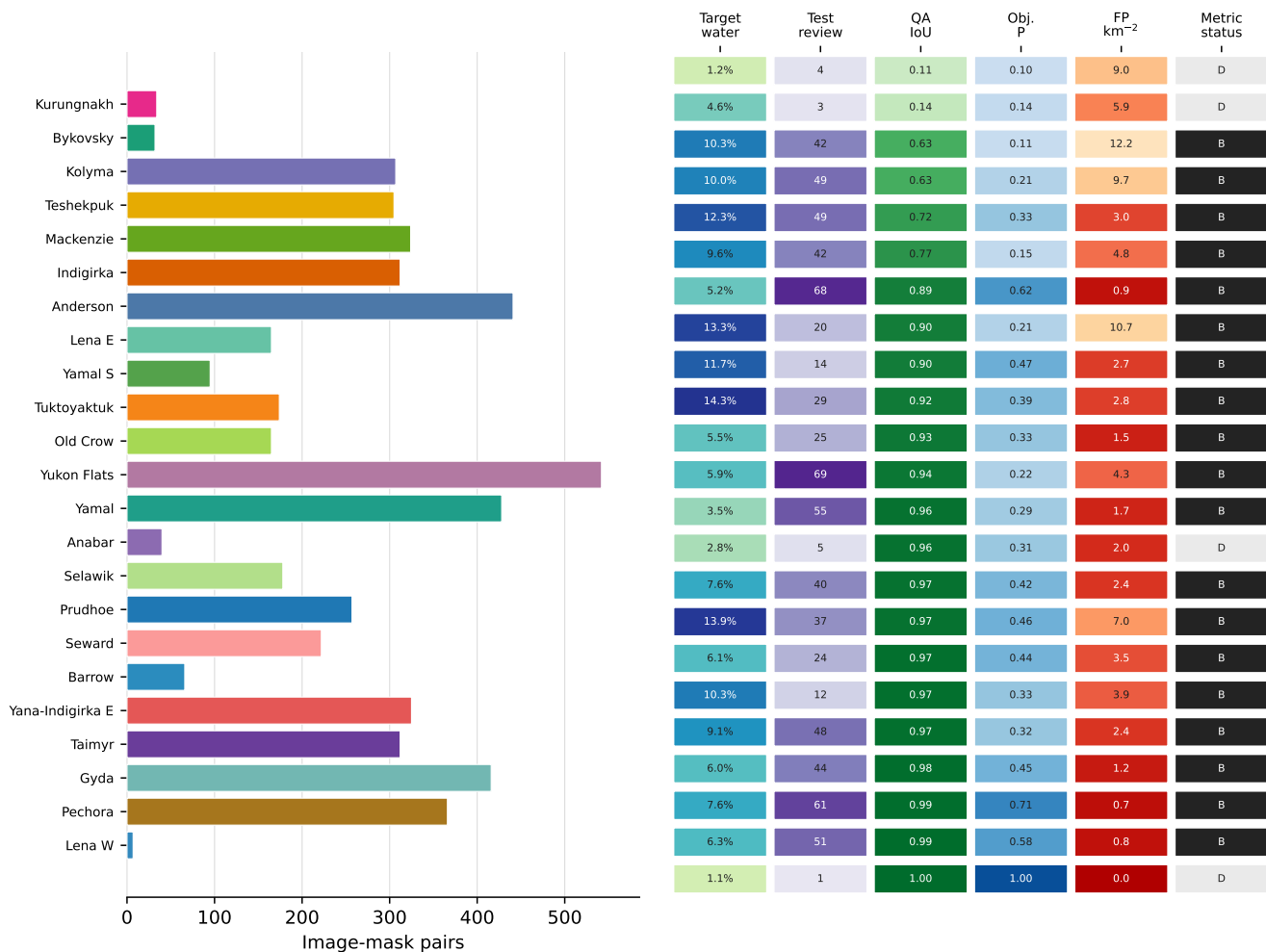


Figure 9. AOI-level dataset coverage and technical validation matrix. The left panel shows the number of image-mask pairs per AOI. The right panel reports target-water fraction, test-tile count, QA-aware pixel IoU, object precision, false-positive objects per square kilometre, and metric-status class. Rows are ordered by QA-aware pixel IoU to emphasize regional variation; exact values are listed in Table 12 and the rigorous-evaluation CSV files.

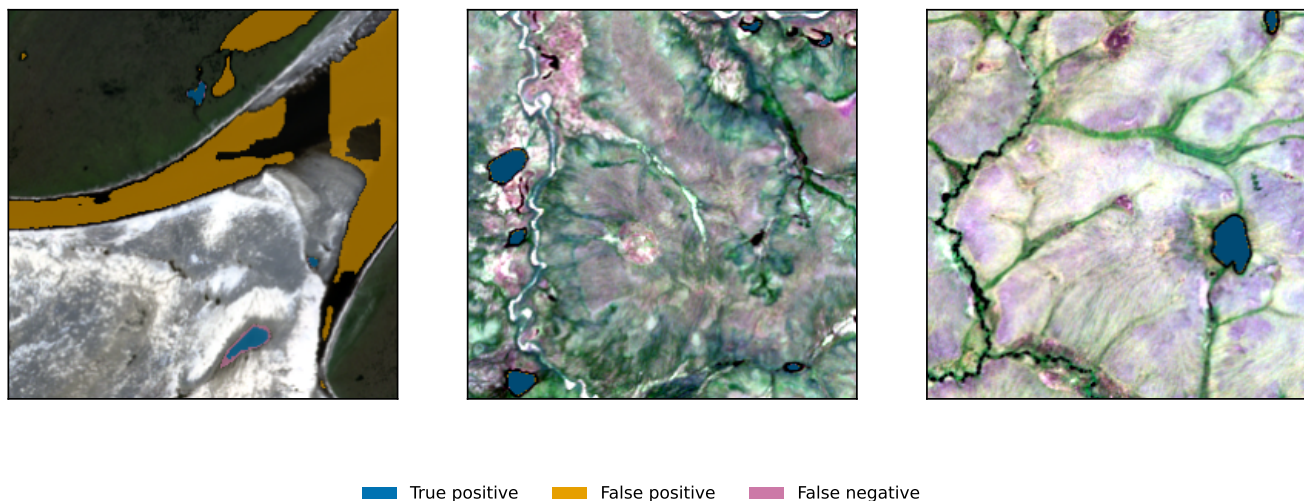


Figure 10. Baseline U-Net prediction review examples from the held-out test split. Panels show lower-, median-, and higher-IoU cases from left to right; blue marks true positives, orange marks false positives, and magenta marks false negatives. Exact tile identifiers and metrics are provided in the review manifest and baseline CSV files. Contains modified Copernicus Sentinel data (2023).

transfer diagnostics. Low-sample and lower-performing AOIs remain diagnostic cases. These limits define the benchmark scope
 410 and motivate extension across additional years, background conditions, and independent reference audits.

7 Usage notes

The dataset can be used directly in semantic segmentation workflows. A typical workflow is to read `manifest.csv`, select rows by `split`, load each image GeoTIFF in the documented five-band order, replace non-finite image values with a finite fill value such as zero, load the paired mask GeoTIFF, and convert mask values greater than zero to the positive class. Users
 415 should load the paired QA mask and exclude QA code 2 from loss functions and evaluation metrics; QA code 3 may also be excluded under the QA-aware policy. The original split assignments reproduce the legacy within-AOI benchmark and retain the documented spatial overlaps. New spatial-generalisation experiments should use geometry-audited assignments such as the supplied buffered Mackenzie split. The greater-than-zero mask rule applies to the binary targets, not to the multiclass QA masks.

420 Users should avoid mixing tiles across splits unless they construct and document a new split. Geospatial generalisation experiments must inspect actual raster footprints and document overlap removal and buffer rules; different AOI names alone do not establish spatial separation. Label-quality studies can stratify by `label_quality`, QA-code counts, or region group. External-accuracy studies should audit the validation and test masks against an independent reference before final model comparison.



425 The sequence of summer Sentinel-2 compositing, initial full-scene screening, tiling, tile-level target assembly and package
validation provides a basis for extending the dataset to additional AOIs.

8 Conclusions

We present a Sentinel-2 semantic segmentation dataset for thermokarst-associated lake and pond water surfaces across 23
selected Arctic permafrost-lowland AOIs. The package contains 5513 image-mask pairs, one QA/ignore mask per tile, tile-level
430 thermokarst-relevance fields, documented benchmark and geographic-holdout splits, an index of 150 low-water core tiles, AOI
summaries, processing records and technical-validation baselines. The geometry audit documents overlap in the original splits;
the buffered experiment and existing-product comparison provide complementary target-agreement assessments. Independent
shoreline accuracy has not been established. The dataset supports supervised segmentation and controlled model comparison
in the represented 2023 Arctic lowland settings. Pan-Arctic inventory applications will require broader temporal coverage and
435 independent object-level geomorphic and shoreline validation.

9 Code and data availability

The training resource and processing code are openly available from the National Cryosphere Desert Data Center under DOI
<https://doi.org/10.12072/ncdc.permafrost.db7645.2026> (Guo, 2026). The English-language record and repository downloads
are available at <https://www.ncdc.ac.cn/portal/metadata/af618102-22ef-4618-b1a8-66d55a360fd9?lang=en>. The accompany-
440 ing model-output dataset is a separate resource openly available on Zenodo at <https://doi.org/10.5281/zenodo.22965055> (Zhang
et al., 2026).

Author contributions. ZZ was responsible for supervision, project administration, funding acquisition, and writing – review and editing. ZG
was responsible for conceptualization, methodology, software, data curation, formal analysis, validation, visualization, and writing – original
draft. QY and JZ contributed to data curation, validation, and writing – review and editing. LG, NT, and AZ contributed to investigation,
445 resources, validation, and writing – review and editing.

Competing interests. The authors declare that they have no conflict of interest.

Disclaimer. The views expressed in this manuscript are those of the authors.



Financial support. The authors gratefully acknowledge support from the National Key R&D Program of China (2025YFE0112500) and budget-funded projects of the Ministry of Science and Higher Education of the Russian Federation (Project No. 126020616798-4 and No. 126020516709-1).

Acknowledgements. This work used Sentinel-2 data provided by the European Space Agency and Google Earth Engine for source-image preparation. Contains modified Copernicus Sentinel data (2023).



References

- Ahlswede, S., Schulz, C., Gava, C., Helber, P., Bischke, B., Förster, M., Arias, F., Hees, J., Demir, B., and Kleinschmit, B.: TreeSatAI
 455 Benchmark Archive: a multi-sensor, multi-label dataset for tree species classification in remote sensing, *Earth System Science Data*, 15,
 681–695, <https://doi.org/10.5194/essd-15-681-2023>, 2023.
- Betancourt, C., Stomberg, T., Roscher, R., Schultz, M. G., and Stadtler, S.: AQ-Bench: a benchmark dataset for machine learning on global
 air quality metrics, *Earth System Science Data*, 13, 3013–3033, <https://doi.org/10.5194/essd-13-3013-2021>, 2021.
- Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A.,
 460 Spoto, F., Sy, O., Marchese, F., and Bargellini, P.: Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services,
Remote Sensing of Environment, 120, 25–36, <https://doi.org/10.1016/j.rse.2011.11.026>, 2012.
- Fang, C., Fan, X., Wang, X., Nava, L., Zhong, H., Dong, X., Qi, J., and Catani, F.: A globally distributed dataset of coseismic landslide map-
 ping via multi-source high-resolution remote sensing images, *Earth System Science Data*, 16, 4817–4842, <https://doi.org/10.5194/essd-16-4817-2024>, 2024.
- 465 Feyisa, G. L., Meilby, H., Fensholt, R., and Proud, S. R.: Automated Water Extraction Index: A new technique for surface water mapping
 using Landsat imagery, *Remote Sensing of Environment*, 140, 23–35, <https://doi.org/10.1016/j.rse.2013.08.029>, 2014.
- Freudenberg, M., Schnell, S., and Magdon, P.: A Sentinel-2 machine learning dataset for tree species classification in Germany, *Earth System
 Science Data*, 17, 351–367, <https://doi.org/10.5194/essd-17-351-2025>, 2025.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R.: Google Earth Engine: Planetary-scale geospatial analysis
 470 for everyone, *Remote Sensing of Environment*, 202, 18–27, <https://doi.org/10.1016/j.rse.2017.06.031>, 2017.
- Guo, Z.: 10 m Sentinel-2 Semantic Segmentation Dataset for Thermokarst-Associated Lake and Pond Water Surfaces in Arctic Permafrost
 Lowlands (2023), <https://doi.org/10.12072/ncdc.permafrost.db7645.2026>, data set, version 1.0, 2026.
- Hou, J., Van Dijk, A. I. J. M., Renzullo, L. J., and Larraondo, P. R.: GloLakes: water storage dynamics for 27 000 lakes globally from 1984
 to present derived from satellite altimetry and optical imaging, *Earth System Science Data*, 16, 201–218, <https://doi.org/10.5194/essd-16-201-2024>, 2024.
- 475 Hu, J., Liu, R., Hong, D., Camero, A., Yao, J., Schneider, M., Kurz, F., Segl, K., and Zhu, X. X.: MDAS: a new multimodal benchmark
 dataset for remote sensing, *Earth System Science Data*, 15, 113–131, <https://doi.org/10.5194/essd-15-113-2023>, 2023.
- Khaldi, R., Alcaraz-Segura, D., Guirado, E., Benhammou, Y., El Afia, A., Herrera, F., and Tabik, S.: TimeSpec4LULC: a global multi-
 spectral time series database for training LULC mapping models with machine learning, *Earth System Science Data*, 14, 1377–1411,
 480 <https://doi.org/10.5194/essd-14-1377-2022>, 2022.
- Lehner, B., Anand, M., Fluet-Chouinard, E., Tan, F., Aires, F., Allen, G. H., Bousquet, P., Canadell, J. G., Davidson, N., Finlayson, C. M.,
 et al.: Mapping the world’s inland surface waters: an upgrade to the Global Lakes and Wetlands Database (GLWD v2), *Earth System
 Science Data*, 17, 2277–2329, <https://doi.org/10.5194/essd-17-2277-2025>, 2025.
- Lesi, M., Nie, Y., Shugar, D. H., Wang, J., Deng, Q., Chen, H., and Fan, J.: Landsat- and Sentinel-derived glacial lake dataset in the China–
 485 Pakistan Economic Corridor from 1990 to 2020, *Earth System Science Data*, 14, 5489–5512, <https://doi.org/10.5194/essd-14-5489-2022>,
 2022.
- Levenson, E.: Alaska Lake and Pond Occurrence Dataset, <https://doi.org/10.5281/zenodo.14199365>, 2024.



- Lin, H., Du, J., Kimball, J. S., Cheng, X., Donnelly, J. P., Watts, J. D., and Bartsch, A.: A satellite-based ice fraction record for small water bodies of the Arctic Coastal Plain (2017 to 2023), *Earth System Science Data*, 18, 535–549, <https://doi.org/10.5194/essd-18-535-2026>, 2026.
- McFeeters, S. K.: The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features, *International Journal of Remote Sensing*, 17, 1425–1432, <https://doi.org/10.1080/01431169608948714>, 1996.
- Meena, S. R., Nava, L., Bhuyan, K., Puliero, S., Soares, L. P., Dias, H. C., Floris, M., and Catani, F.: HR-GLDD: a globally distributed dataset using generalized deep learning for rapid landslide mapping on high-resolution satellite imagery, *Earth System Science Data*, 15, 3283–3298, <https://doi.org/10.5194/essd-15-3283-2023>, 2023.
- Mukherjee, R., Policelli, F., Wang, R., Arellano-Thompson, E., Tellman, B., Sharma, P., Zhang, Z., and Giezendanner, J.: A globally sampled high-resolution hand-labeled validation dataset for evaluating surface water extent maps, *Earth System Science Data*, 16, 4311–4323, <https://doi.org/10.5194/essd-16-4311-2024>, 2024.
- Muster, S., Roth, K., Langer, M., Lange, S., Cresto Aleina, F., Bartsch, A., Morgenstern, A., Grosse, G., Jones, B., Sannel, A. B. K., Sjöberg, Y., Günther, F., Andresen, C., Veremeeva, A., Lindgren, P. R., Bouchard, F., Lara, M. J., Fortier, D., Charbonneau, S., Virtanen, T. A., Hugelius, G., Palmtag, J., Siewert, M. B., Riley, W. J., Koven, C. D., and Boike, J.: PeRL: a circum-Arctic Permafrost Region Pond and Lake database, *Earth System Science Data*, <https://doi.org/10.5194/essd-9-317-2017>, 2017.
- Ohmer, M., Liesch, T., Habbel, B., Heudorfer, B., Gomez, M., Clos, P., Nölscher, M., and Broda, S.: GEMS-GER: a machine learning benchmark dataset of long-term groundwater levels in Germany with meteorological forcings and site-specific environmental features, *Earth System Science Data*, 18, 77–95, <https://doi.org/10.5194/essd-18-77-2026>, 2026.
- Pekel, J.-F., Cottam, A., Gorelick, N., and Belward, A. S.: High-resolution mapping of global surface water and its long-term changes, *Nature*, <https://doi.org/10.1038/nature20584>, 2016.
- Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351 of *Lecture Notes in Computer Science*, pp. 234–241, Springer, https://doi.org/10.1007/978-3-319-24574-4_28, 2015.
- Tao, C., Zhong, D., Mu, W., Du, Z., and Wu, H.: A large-scale image–text dataset benchmark for farmland segmentation, *Earth System Science Data*, 17, 4835–4864, <https://doi.org/10.5194/essd-17-4835-2025>, 2025.
- Walter Anthony, K., von Deimling, T. S., Nitze, I., Frolking, S., Emond, A., Daanen, R., Anthony, P., Lindgren, P., Jones, B., and Grosse, G.: 21st-century modeled permafrost carbon emissions accelerated by abrupt thaw beneath lakes, *Nature Communications*, 9, 3262, <https://doi.org/10.1038/s41467-018-05738-9>, 2018.
- Xu, H.: Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery, *International Journal of Remote Sensing*, 27, 3025–3033, <https://doi.org/10.1080/01431160600589179>, 2006.
- Yuan, Z., Xiong, Z., Mou, L., and Zhu, X. X.: ChatEarthNet: a global-scale image–text dataset empowering vision–language geo-foundation models, *Earth System Science Data*, 17, 1245–1263, <https://doi.org/10.5194/essd-17-1245-2025>, 2025.
- Zhang, Z., Guo, Z., Yan, Q., Zhai, J., Gagarin, L., Torgovkin, N., and Zhang, A.: U-Net geographic-holdout predictions for Arctic lake and pond water surfaces from Sentinel-2 imagery (2023), <https://doi.org/10.5281/zenodo.22965055>, data set, 2026.