https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Advancing Methods to Monitor Greenland Ice-Marginal Lake Area Fluctuations and Outburst Floods Using Machine Learning and Multispectral Optical Imagery

Ethan Carr[1,2], Alison F. Banwell[3,2], Mark Serreze[1,2,4], Melisa Diaz[5], Waleed Abdalati[1,2], Keith N. Musselman[1,6], J. Rachel Carr[7], Jonathan L. Carrivick[8], Mads Dømgaard[9], Penelope How[10], Aman KC[11], Kristian Kjeldsen[10], Jonas Köhler[10], Natalie Lützow[12], Joseph Mallalieu[13], Daniel McGrath[14], Michela Savignano[1,2], Dan H. Shugar[15]

[1]Department of Geography, University of Colorado Boulder, Boulder, USA
[2]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, USA
[3]Centre for Polar Observation and Modelling (CPOM), School of Geography and Natural Sciences, Northumbria University, Newcastle Upon Tyne, UK
[4]National Snow and Ice Data Center (NSIDC), Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, USA
[5]School of Earth Sciences, The Ohio State University, Columbus, USA
[6]Institute of Arctic and Alpine Research, University of Colorado Boulder, Boulder, USA
[7]School of Geography, Politics and Sociology, Newcastle University, Newcastle Upon Tyne, UK
[8]School of Geography and water@leeds, University of Leeds, Leeds, UK
[9]Department of Geoscience and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark
[10]Department of Glaciology and Climate, Geological Survey of Denmark and Greenland, Copenhagen, Denmark
[11]Department of Geosciences, Boise State University, Boise, USA
[12]Institute of Environmental Science and Geography, University of Potsdam, Potsdam-Golm, Germany
[13]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, UK
[14]Department of Geosciences, Colorado State University, Fort Collins, USA
[15]Department of Earth, Energy, and Environment, University of Calgary, Calgary, Canada

Correspondence to: Ethan Carr (ethan.carr@colorado.edu)

Abstract. Ice-marginal lakes (IMLs) are increasing in number and size around the Greenland Ice Sheet and its peripheral glaciers. Yet consistent long-term mapping of IML surface area and detection of sudden drainages, possibly as glacial lake outburst floods (GLOFs), remain unrealized. This study evaluated 10 spectral indices and 28 machine-learning models to identify an optimal method for extracting IML surface area in Greenland. Performance was assessed using expert-elicited pixel classifications and stratified random samples from six known GLOF-producing IMLs. Spectral index performance varied substantially: the Water Ratio Index (WRI) and Icy Lakes Index (ILI) performed well across highly weighted metrics and produced the highest weighted rank-sum values (0.8990 and 0.8987). The widely used Normalized Difference Water Index (NDWI) performed poorly, scoring the third-lowest weighted rank-sum value (0.7126). Machine-learning models, developed using a clustering-to-random-forest methodology, generally outperformed spectral indices, with 10 models exceeding the best spectral index scores. The top performer, Model_1i, trained on Landsat 8's seven surface reflectance bands and including lake ice pixels in the training data, achieved a weighted rank-sum value of 0.9327. An expert intercomparison dataset of 300 commonly interpreted pixels showed a mean agreement of 72% among experts, underscoring the need for rigorously trained and tested automated approaches. These combined results demonstrate that machine-learning methods can provide the most consistent, scalable, and expert-aligned approach for tracking Greenland IML areas and their potential drainages. Future work will apply the optimal model (Model_1i) to produce a Greenland-wide, multi-decadal database of IML dynamics that will, notably, be used for GLOF detection.

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

1 Introduction

The Greenland Ice Sheet (GrIS) and its peripheral glaciers and ice caps have become the largest freshwater contributors to global sea-level rise, a trend expected to continue through the end of the century (AMAP, 2021; Box et al., 2022; Hanna et al., 2026; Park et al., 2023; Stokes, 2025). Although large-scale climate models assume that surface meltwater drains directly into the oceans, some of it drains into ice-marginal lakes (IMLs), which globally hold 0.43 mm of sea-level equivalent (How et al., 2025; Shugar et al., 2020). Around the GrIS, 3,347 (± 8%) IMLs larger than 0.05 km² occupy ~10 % of the ice sheet and ~5 % of the peripheral glaciers and ice cap margins (Carrivick et al., 2022; How et al., 2021). These IMLs exert a measurable influence on ice dynamics as calving at ice-lake boundaries produces ~4.9 Gt yr^{-1} of GrIS mass loss, and lake-terminating glacier margins flow ~25 % faster than their land-terminating counterparts (Carrivick et al., 2022; Mallalieu et al., 2021; Harpur et al., 2025). As the GrIS continues to lose mass, the number and area of IMLs is expected to further increase as newly exposed overdeepenings in the proglacial region fill with meltwater (Carrivick & Quincey, 2014; Carrivick et al., 2022).

While the number of IMLs in Greenland has increased in recent decades (How et al., 2021), a phenomenon associated with IMLs, glacial lake outburst floods (GLOFs), remain largely understudied, especially when compared to GLOFs in other glaciated regions worldwide (Zhang et al., 2024). GLOFs are sudden drainage events that can occur from IMLs. GLOFs in Greenland can modify ice dynamics (Kjeldsen et al., 2017; Stevens et al., 2016), alter fjord circulation and salinity (Kjeldsen et al., 2014), reshape proglacial geomorphology (Carrivick et al., 2013; Carrivick & Tweed, 2019; Dømgaard et al., 2022), and affect local ecosystems (Schiøtt et al., 2022). Despite Greenland hosting ~34% of the world's IMLs (Song et al., 2025), only 153 of the 3,151 observed GLOFs (~5 %) listed in the 2023 Glacier Lake Outburst Flood Database are from Greenland (Lützow et al., 2023; Lützow & Veh, 2023). Recent studies indicate that these are severe underestimates: a satellite lidar altimetry-based assessment identified 541 GLOFs from 1,100 IMLs between 2008 and 2022 (Dømgaard et al., 2024), and manual delineation identified 552 GLOFs from 47 IMLs between 1985 and 2022 (Carr, 2024). These findings highlight the need for better identification of drainage events, including GLOFs, across all of Greenland's 3,347 IMLs.

Commonly used approaches to identify IML drainages, including GLOFs, face many limitations in Greenland. Stream gauges are sparse and may fail to capture events that drain directly into fjords. Satellite altimetry, while effective, favors larger IMLs, and manual delineation of IML area outlines is subjective and neither feasible nor repeatable for >3,000 IMLs. Spectral index approaches offer scalable automation but vary in performance across regions and complex glaciated terrain. For instance, the widely used Normalized Difference Water Index (NDWI; Gao, 1996; McFeeters, 1996) performs inconsistently due to spectral similarities between turbid IMLs and debris-covered ice, as well as large solar zenith angles and their corresponding shadows. Large solar zenith angles also reduce reflectance and can result in misclassification of remnant lake ice (Feyisa et al., 2014; Hengst et al., 2021; Maciel et al., 2023; Veh et al., 2022; Zhao et al., 2025). To better understand Greenland IML dynamics, including GLOF behavior, an improved automated approach that uses multispectral optical satellite observations is needed.

To this end, the present study evaluates 10 spectral indices and 28 optical satellite image band combinations within a machine learning-based clustering-to-random-forest framework to identify the optimal method for mapping IMLs. The methods are trained on Landsat 8 imagery from southwest Greenland and tested on six known GLOF-producing IMLs (Carrivick & Tweed, 2019) using expert-elicited pixel classification and stratified random sampling. We conclude by recommending the top-performing method for IML area change detection, which, in future work, will be scaled to all 3,347

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

(±8%) IMLs in Greenland to ultimately establish a comprehensive database of drainages and GLOFs across the entire Landsat satellite era (1984 - present).

2 Data and methods

2.1 Study area

The methods developed in this study were trained across the proglacial region in southwestern Greenland and tested over six known GLOF-producing ice-dammed IMLs (Fig. 1). Four of these have known names - Tininnilik, Russell, Iluliallup Tasersua and Isvand. The two remaining lakes are unnamed according to Oqaasileriffik (the Language Secretariat of Greenland) placename database (Oqaasileriffik, 2026)but have been referred to in previous work as North Midternæs and Base Camp (Carrivick and Tweed, 2019). These alternative names will be adopted for the purpose of this study. Selection of these IMLs was informed by the Greenland-wide review of these features by Carrivick & Tweed (2019) and consideration of three additional criteria: (i) inclusion of the two most extensively studied IMLs in the region, Tininnilik and Russell; (ii) broad spatial dispersion across southwestern Greenland; and (iii) representation of three lake-area classes (~30 km², ~5 km², ~0.5 km²) discussed further in Section 2.5.2.

2.2 Scene Selection

Landsat 8 scenes for both the training and testing datasets were selected and pre-processed using consistent criteria and methods (Table S1). For both datasets, Landsat 8 scenes located within the area shown in Figure 1a and with less than 25% cloud cover were selected for each full melt season (1 May to 31 October) from 2013-2024. From these images (171 in total), Google Earth Engine selected 20 scenes at random, producing a training dataset covering southwestern Greenland (Fig. 1a), a region including roughly 1,250 IMLs (How et al. 2021). For the test dataset, an additional Landsat 8 image for each of the six IMLs (Fig. 1b-g) was selected using the same criteria. This approach ensures that classifiers and spectral indices are trained and tested across a wide range of environmental and spectral characteristics.

2.3 Spectral Indices and Thresholding

2.3.1 Spectral Indices

We tested the performance of 10 spectral indices (Table 1), which have been utilized in previous studies to calculate the area of water bodies, snow, or lake ice, as well as, in particular, IMLs and/or to identify GLOFs.

2.3.2 Optimal Image Segmentation using Otsu Thresholding Method

We use the algorithm developed by Otsu (1979) to divide selected images into spectrally distinct non-overlapping regions. This is a widely used, computationally efficient thresholding technique for image segmentation across different land-cover classifications (Yu et al., 2023). It identifies an optimal threshold for segmenting water pixels from the rest of the image by analyzing a histogram of spectral index values and maximizing the separation between the classes.

The glacial margin and proglacial environment of southwest GrIS are primarily characterized by IMLs (with and without suspended sediment), land, snow, debris-covered ice, and glacial ice, which pose greater challenges for image segmentation due to overlapping spectral properties. Otsu's method was chosen because its optimization criterion minimizes within-class variance and maximizes between-class variance, aligning well with spectral indices designed to distinguish water from other land-cover types, and because it performs robustly under variable illumination, which is an important consideration in

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

high-latitude regions with large reflectance variability (Chen et al., 2024; Rosser et al., 2017). Figure 2 illustrates how the Otsu threshold separates water from other land-cover types and provides the thresholds for each spectral index.

For the spectral indices, 30,000 pixels were sampled from each of the 20 training images—10,000 pixels per land-cover class (water, land, and glacial ice)—yielding a total of 600,000 pixels. Otsu's algorithm was applied to determine the optimal threshold for each index. When an index produced multimodal histograms, or yielded non-optimal thresholds, the algorithm was applied to subsets of land-cover classes. In a small number of cases, manual adjustment was required to achieve accurate segmentation. Once thresholds were identified, the spectral indices were applied to each of the six test images, and the optimized threshold for each index was used to extract the water body area. The area of the water body was then calculated by multiplying the number of pixels by their spatial resolution (900 m^2 for Landsat 8).

2.4 Clustering-to-Random-Forest

Training data generation and classifier training followed the general approach of Dell et al. (2022, 2024) and Halberstadt et al. (2020), hereafter referred to as the clustering-to-random-forest method. Training data were produced using a supervised k-means clustering algorithm in Google Earth Engine (Arthur & Vassilvitskii, 2007), applied to spectrally distinct pixels across 20 Landsat 8 training scenes.

The k-means algorithm sampled 100,000 pixels from each of the 20 training images. The wekaXMeans clustering algorithm (Arthur & Vassilvitskii, 2007), which automatically identifies the optimal number of clusters within a specified range, was then applied to the 2,000,000 sampled training pixels. Following Dell et al. (2022), we used wekaXMeans with a minimum of 5 and a maximum of 70 clusters; increasing the upper bound beyond 70 did not alter results, as the algorithm consistently produced a maximum of 30–40 clusters. The original Dell et al. (2022) approach was designed to identify and differentiate between supraglacial meltwater and slush in Antarctica, which used an NDWI mask to remove rock or land areas. Since such a mask would have removed land pixels from our testing data, we instead used an additional cluster that identified land pixels, preserved them, and allowed their inclusion in the training data.

Using the 14 band combinations listed in Table 2, 28 models were generated by systematically varying the input bands to assess their performance. Model_1 used the seven Landsat 8 surface-reflectance bands, following Dell et al. (2022). Model_2 started with all surface reflectance bands and included all spectral indices from Table 1 as bands. From Model_2, we derived Models 3–7 by iteratively removing the least useful predictors based on feature-importance scores, producing progressively leaner models. Models 8–10 were selected using a combination of feature importance and permutation-importance metrics to identify the most informative bands, and Models 11–14 also used feature importance and permutation-importance but excluded indices entirely to evaluate how well band-only models perform.

Dell et al. (2022, 2024) demonstrated that multiple clustering iterations often yield "imperfect" or overlapping clusters, for example, water and ice occasionally being grouped together. This behavior is not unexpected in spectrally complex proglacial environments such as southwest Greenland and is incorporated into our clustering-to-random-forest workflow shown in Figure 3, which describes the parameters used in both our clustering and random forest models.

2.5 Accuracy Assessment

The performance of the spectral indices and the clustering-to-random-forest classifiers was assessed using a dataset of six test scenes, each consisting of one image per test IML. Two data collection methods were used to assess all methods: expert elicitation and stratified random sampling.

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

2.5.1 Expert Elicitation

"Expert elicitation" (Dell et al., 2022; Bamber & Aspinall, 2013) involved scientists selected for their expertise in analyzing glacial hydrologic features, proglacial regions, and various forms of glacial lakes using medium-resolution (30 m) optical satellite data. Twelve experts were asked to manually interpret a total of 100 pixels for each of the six test images, classifying the pixel as 'water', 'lake ice', 'glacier ice', 'land', or 'other'. Experts viewed each pixel within the surrounding environment and could zoom in and out of the image to gain context. For each pixel, experts also assigned a confidence score reflecting the certainty of their manual interpretation: 'high', 'medium', or 'low'. The experts were not involved in developing the different methods or training the classifiers. Experts were not given explicit instructions (as provided in the Supplementary Information) on how to interpret the pixels, so their interpretations were not biased by the individual who developed the various methods.

Of the 100 pixels per image that each expert interpreted, the first 50 for each image were identical between experts. These 300 pixels were used as an 'expert intercomparison dataset' to compare expert opinions. The first author also interpreted the 300 pixels from the expert intercomparison dataset, comprising 3,900 interpreted pixels. The second 50 pixels per image were unique to each expert and formed the 'expert validation dataset', comprising 3,600 interpreted pixels.

2.5.2 Stratified Proportional Random Sampling

The Stratified Random Sampling (SRS) method (Chaaban et al., 2022; Shetty et al., 2021) was used to evaluate each technique's ability to identify pixels labeled as water using samples from manually delineated IML areas. To ensure area-proportional sampling, we used a density of 100 pixels per km². This produced sample counts of 3,000 pixels for the ~30 km² IMLs (Tininnilik and Iluliallup Tasersua), 500 pixels for the ~5 km² IMLs (Isvand and North Midternæs), and 50 pixels for the ~0.5 km² IMLs (Russell and Base Camp); 7,100 pixels in total. SRS provides probability-based, class-specific accuracy estimates, allows standard-error calculation, and is flexible in sample sizing (Chaaban et al., 2022; Stehman, 2009).

2.6 Comparing Performance of Spectral Indices and Clustering-to-Random-Forest Methods

We applied a weighted rank-sum approach to assess the performance of each spectral index and clustering-to-random-forest model to establish the top overall method. The 10 evaluation metrics were grouped into five priority tiers: highest (5), high (4), medium-high (3), medium (2), and low (1). Tier scores were normalized and used to compute a weighted rank-sum for each method, enabling a standardized multi-metric comparison.

The highest priority was assigned to the high-confidence Kappa coefficient (Kappa_HC) and high-confidence omission error (OE_HC). Kappa_HC captures agreement beyond chance across water and non-water pixels, while OE_HC directly reflects the ability to avoid false negatives, critical for reliable water detection. High priority was given to high-confidence commission error (CE_HC), which reflects the ability to avoid overestimating the IML area. CE_HC was ranked below OE_HC because minimizing missed water pixels was considered more important than minimizing false positives. Medium-high priority metrics were the pixel-weighted accuracy of the stratified random sample (PWA_SRS) and the Kappa coefficient for the full expert validation dataset (Kappa). PWA_SRS provides an unbiased, area-weighted assessment, while Kappa and Kappa_HC ensure agreement beyond chance. Medium priority was assigned to high-confidence overall accuracy (OA_HC), which reflects agreement among the most reliable expert-classified pixels. Finally, low-priority metrics, full expert validation dataset overall accuracy (OA), stratified-sample overall accuracy (OA_SRS), and full expert validation dataset omission and commission errors (OE, CE), were included primarily to be used as tiebreakers. These metrics are informative but are considered less important for distinguishing water from non-water in complex proglacial environments. The normalized weights and reasoning for all metrics follow in Table 3.

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

2.7 Expert Intercomparison Dataset

The expert intercomparison dataset is intended to highlight the subjectivity of pixel identification and land-cover classification in satellite imagery from topographically complex environments. This dataset was not used directly to identify the top-performing method; rather, it was used to assess the levels of agreement, or lack thereof, among experts. Additionally, the first author of this study contributed to the expert intercomparison dataset as an additional expert, supplementing the other 12 experts, resulting in 3,900 interpretations for this 300-pixel dataset.

Agreement between experts was assessed in two ways: "surface type agreement" is defined as all experts interpreting a pixel as belonging to the same land-cover class, i.e., all experts chose 'water' for a given pixel. "Binary surface type agreement" classifies water and lake ice as 'water,' and land and glacier ice as 'non-water.' The binary surface type agreement assessment evaluated expert interpretation in the same way that the models are expected to perform; by classifying pixels as 'water' or 'non-water,' rather than assigning one of the four land-cover classes. Additionally, agreement was assessed for both the "surface type agreement" and "binary surface type agreement" analyses using only pixels assigned a high confidence level by the experts.

3 Results

3.1 Spectral Indices

Across all 10 assessment metrics and the six test IMLs, the spectral indices produced a structured hierarchy of performance, with WRI and ILI emerging as the two strongest indices (Fig. 4a). Although their weighted rank-sum values were nearly identical (0.8990 and 0.8987), the two indices reached these high rankings through strong performance in the majority of metrics, unlike NDWI, which performed more poorly for most metrics (Fig. 4b).

ILI had the strongest agreement with expert classifications, particularly within the high-confidence subset (Kappa_HC = 0.8878; OA_HC = 0.9570) (Fig. 4b; Fig. S1). Its exceptionally low high-confidence commission error (CE_HC = 0.0132) (Fig. 4b) indicates that it rarely misclassified non-water pixels as water, aligning closely with expert interpretations of IML boundaries. This precision makes ILI highly reliable for delineating spectrally ambiguous margins where water, ice, and debris-covered ice share overlapping reflectance characteristics. However, ILI's more conservative behavior is reflected in its omission errors (OE_HC = 0.1237; OE = 0.2076), showing that it occasionally missed true water pixels (Fig. 5b-c), particularly in low-reflectance or shadowed regions. Thus, ILI excels in boundary accuracy but sacrifices complete water detection.

WRI, by contrast, achieved the highest water-focused stratified-sample metrics (PWA_SRS = 0.9939; OA_SRS = 0.9938) (Fig. 4b; Fig. S1) and the lowest omission errors (OE_HC = 0.0372; OE = 0.0896), demonstrating an exceptional ability to detect true water pixels (Fig. 5b-c). This made WRI particularly effective at capturing the full IML extent in instances of low surface reflectivity or ice-covered water bodies, which other indices frequently missed. However, this strength came at the cost of higher commission errors (CE_HC = 0.1422; CE = 0.1166), consistent with a tendency to overestimate water along spectrally ambiguous boundaries (Fig. 5b). Pixel-level examples (Fig. 5b-c) illustrate these behaviors as WRI expands water labels into marginal glaciated areas, while ILI rarely misclassifies ice or land as water but misses some true water pixels.

Beyond the top two indices, the AWEI variants were the next highest performers (Fig. 4a). AWEI_NSH achieved relatively high overall accuracy (OA_HC = 0.9438; OA = 0.9004) (Fig. 4b; Fig. S1) and strong stratified-sample performance (PWA_SRS = 0.9155), though it tended to overclassify non-water pixels (CE_HC = 0.0660; CE = 0.0932). AWEI_SH had similar overall accuracy (OA_HC = 0.9327; OA = 0.8860) and even higher pixel-weighted SRS performance (PWA_SRS = 0.9330), but with larger commission errors (CE_HC = 0.1023; CE = 0.1424), suggesting greater boundary overreach. ANDWI performed comparably in overall accuracy (OA_HC = 0.9385; OA = 0.8944) and maintained very low commission errors

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

(CE_HC = 0.0187; CE = 0.0156), yet its omission errors were substantially higher (OE_HC = 0.1727; OE = 0.2630), indicating a conservative bias similar to ILI but with less favorable trade-offs.

Indices such as SWI and MBWI produced moderate performance across most metrics. SWI showed balanced but unreliable accuracy (Kappa_HC = 0.7402; OA_HC = 0.8888) (Fig. 4b; Fig. S1) and elevated omission and commission errors (OE_HC = 0.1725; CE_HC = 0.1485), indicating a tendency to both miss water and overclassify non-water, an unfavorable combination. MBWI achieved a high pixel-weighted SRS score (PWA_SRS = 0.9582) but exhibited high commission errors (CE_HC = 0.2160; CE = 0.2024), frequently overestimating the water extent and indicating poor boundary precision, particularly along glacial margins.

At the lower end of the performance spectrum, NDWI, ANDSI, and MNDWI demonstrated poorer detection capability than the previously-mentioned indices (Fig. 4a). Of particular note, NDWI, despite its widespread use in global IML and GLOF studies, performed poorly under Greenland's optically challenging surface conditions, yielding a weighted rank-sum value of only 0.7126 (Fig. 4a-b). It produced very high omission errors (OE_HC = 0.4058; OE = 0.4480) (Fig. 5b-c; Fig. S1), weak agreement with expert labels (Kappa_HC = 0.6222; Kappa = 0.5524), and low water-focused accuracy (PWA_SRS = 0.7817; OA_SRS = 0.5877). These deficiencies likely stem from low solar angles, shadows, and debris-covered ice, which confounded separating water from dark or contaminated glacial surfaces. ANDSI and MNDWI performed even worse than NDWI (Fig. 4a), with low overall accuracy (ANDSI OA_HC = 0.7800; MNDWI OA_HC = 0.7028) and extremely high omission errors (MNDWI OE_HC = 0.6502; OE = 0.6508).

In summary, the weighted rank-sum results (Fig. 4a) clarify the complementary strengths of the top-performing indices. WRI provides the most complete water detection and minimizes false negatives, but tends to overestimate IML area due to poor boundary precision (Fig. 5b). ILI offers the highest boundary-precision accuracy with minimal false positives but at the cost of missing some true water pixels (Fig. 5c). Both indices are effective for IML monitoring. The choice between them should depend on whether completeness (WRI) or precision (ILI) is prioritized under similar geographic and spectral conditions. Meanwhile, NDWI, the field-standard spectral index, proved ineffective for Greenlandic IMLs, regularly underestimating IML extent (Fig. 5b-c) and failing to capture drainage-related area changes.

3.2 Clustering-to-Random-Forest

The 28 clustering-to-random-forest models exhibited a wide range of performance. Ten models produced weighted rank-sum values greater than the best performing spectral indices of WRI and ILI (Fig. 6a), while a few performed particularly poorly in rank-sum values, comparable to MNDWI (Fig. 6b). We grouped the models into three performance groups based on the weighted rank-sum values; top (weighted rank-sum > 0.9100), middle (0.9100 > weighted rank-sum > 0.8700), and lowest-performing methods (0.8700 > weighted rank-sum) (Fig. 6b-c). The 10 top-performing models (Fig. 6a-c) consistently achieved high agreement, with balanced error profiles, and strong accuracy across both the expert-validation dataset and the stratified random samples. At the top of this distribution, Model_1i (weighted rank-sum = 0.9327) (Fig. 6d) delivered the strongest overall performance (Kappa_HC = 0.9067; OA_HC = 0.9615) (Fig. 6c; Fig. S2), paired with excellent stratified-sample results (PWA_SRS = 0.9876; OA_SRS = 0.9943) and modest error rates in both error metrics (CE_HC = 0.0748; CE = 0.0950; OE_HC = 0.0342; OE = 0.0673). These metrics indicate a well-balanced classifier that minimized both false positives and false negatives. Pixel-level examples illustrate how Model_1i, Model_4i, and Model_8i identify nearly identical spatial patterns of water and non-water within the expert-elicited pixels (Fig. 7b). This similarity in model performance is further evidenced by the stratified random-sample results (Fig. 7c), which show virtually no red pixels, indicating that all three models produced very low commission errors. Figure 8 further displays Model_1i's performance superiority in identifying water compared to the top-performing spectral index, WRI, and the most commonly used spectral index, NDWI.

Model_4i, close behind Model_1i, produced similarly robust results (Kappa_HC = 0.8958; OA_HC = 0.9577; PWA_SRS = 0.9921; OA_SRS = 0.9963) (Fig. 6c; Fig. S2), with slightly higher omission errors (OE_HC = 0.0237; OE = 0.0485) and commission errors (CE_HC = 0.0977; CE = 0.1165). Model_8i and Model_10n also achieved high agreement (Model_8i:

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

OA_HC = 0.9592; Kappa_HC = 0.8953; PWA_SRS = 0.9900; Model_10n: OA_HC = 0.9585; Kappa_HC = 0.8942; PWA_SRS = 0.9885). The rest of the models displayed in Figure 6a, Model_4n, Model_9n, Model_1n, Model_8n, Model_9i, and Model_14i, performed strongly as well, each producing OA_HC values above 0.95, Kappa_HC values above 0.85, CE_HC below 0.12, OE_HC below 0.10 (Fig. S2), along with high stratified-sample accuracy.

Outside this top tier of high performing models, many excelled in only one error metric (Fig. 6b). Model_7i and Model_5i minimized omission errors (Model_7i OE_HC = 0.0200; Model_5i OE_HC = 0.0237) (Fig. 6c; Fig. S2) and achieved near-perfect stratified-sample results (Model_7i PWA_SRS = 0.9965; OA_SRS = 0.9986; Model_5i PWA_SRS = 0.9970; OA_SRS = 0.9983). However, this sensitivity came at the cost of higher commission errors (Model_7i CE_HC = 0.1905; Model_5i CE_HC = 0.1803), indicating a tendency to overclassify water and potentially overestimate IML boundaries. In contrast, Model_9n and Model_14n produced low commission errors (Model_9n CE_HC = 0.0253; Model_14n CE_HC = 0.0367) but higher omission errors (Model_9n OE_HC = 0.0932; Model_14n OE_HC = 0.1418), reflecting a more conservative bias that reduced false positives at the expense of missing water pixels. These error profiles highlight that the "ideal" model depends on the user's operational priority; whether they aim to minimize false negatives (complete water detection) or minimize false positives (boundary precision).

Taken together, the rank-sum integration confirms that a relatively small subset of clustering-to-random-forest configurations (Model_1i, Model_8i, Model_10n, Model_4i, and Model_9n/9i) consistently delivered the best balance of accuracy, agreement, and error trade-offs for Greenlandic IML mapping. At the same time, models with greater sensitivity (e.g., Model_7i, Model_5i) or more conservative behavior (e.g., Model_9n, Model_14n) may be valuable when specific operational constraints dictate different error tolerances.

3.3 Expert Intercomparison Dataset

Across the five surface-type classes, the 13-expert intercomparison dataset revealed substantial variability in manual interpretation using medium-resolution imagery (Fig. 9). Overall agreement was 72% (217/300 pixels) (Table 4), ranging from 88% at North Midternæs and 86% at Russell to 60% at Tininnilik (Fig. 9a) and Isvand (Fig. 9b). When restricting the dataset to high-confidence labels, agreement dropped sharply to 54% (162/300 pixels) (Table 4), demonstrating that even pixels experts felt certain about were not consistently interpreted across the group. Binary water vs. non-water classification increased agreement to 84% (251/300 pixels), yet the high-confidence subset again fell to 54% (163/300 pixels) (Table 4), indicating that ambiguity persists even when the decision is reduced to two classes. Figure 9 shows that disagreement clusters along IML margins, shadowed terrain, and debris-covered ice, which are the same zones where spectral indices and machine-learning models have the greatest difficulty.

These results underscore that ambiguity is fundamentally landscape-driven, not a failure of individual interpreters or methods. The same terrain features that challenge experts also produce elevated errors in automated approaches, suggesting that disagreement reflects inherent visual limitations of 30-meter imagery rather than inconsistent human judgment. This reinforces the need for automated classification methods that apply consistent decision rules across ambiguous regions and produce scalable, expert-aligned interpretations of Greenland's IMLs.

4 Intercomparison of spectral index and ML model performances and limitations

The clustering-to-random-forest models, particularly the 10 highest-ranked models displayed in Figure 6a, demonstrate strong potential for operational IML mapping across Greenland. These models consistently achieved high agreement with expert interpretations, produced low omission and commission errors, and showed robust stratified-sample performance, indicating reliable water detection under the challenging spectral conditions characteristic of proglacial environments. Their weighted rank-sum values (Fig. 6b) far exceed those of the highest-performing spectral indices (WRI and ILI; Fig. 4a), emphasizing the advantages of machine-learning approaches for IML delineation.

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

To highlight the general effectiveness of the clustering-to-random-forest models relative to the spectral index methods, the stratified random-sample results provide a clear comparison: 18 of the 28 models achieved both PWA_SRS and OA_SRS values above 0.9000, and nine exceeded 0.9900 in both metrics (Fig. S2), demonstrating that most model configurations reliably identify true water pixels. Meanwhile, only two spectral indices, WRI and MBWI, achieved PWA_SRS and OA_SRS values above 0.9000, with only WRI exceeding 0.9900. A similar pattern appears in the agreement metrics, where 19 clustering-to-random-forest models produced Kappa_HC values higher than the best-performing spectral index, WRI, indicating that even several middle-tier and lower-tier models achieved stronger agreement with the expert-validation dataset than the top spectral index (Fig. S2).

A key methodological insight from this comparison is the reproducible trade-off exhibited by spectral indices. WRI maximizes completeness and minimizes false negatives, making it well suited for applications requiring comprehensive water detection, such as drainage identification or volumetric change estimation, particularly when IMLs are spectrally distinct from surrounding ice. In contrast, ILI prioritizes boundary precision and minimizes false positives, making it preferable for accurate boundary delineation. These complementary behaviors suggest it could be useful in a potential workflow in which WRI identifies candidate water pixels and ILI refines boundaries (if computational resources and validation allow). Such an approach captures many of the benefits of machine-learning classification while retaining the simplicity and low computational cost of index-based methods. Notably, NDWI, despite its widespread use in IML and GLOF detection studies (Shugar et al., 2020; Zhang et al., 2024), performed poorly in Greenland's low solar angles, shadowed terrain, and debris-covered ice, regularly underestimating IML extent and producing high omission errors. These results highlight the need for caution when applying NDWI in high-latitude proglacial environments and reinforce the value of alternative indices or machine-learning approaches.

The strongest operational application of the clustering-to-random-forest models is the production of consistent, Greenland-wide IML water masks across the Landsat archive. Because the top-performing configuration relies solely on the seven Landsat surface-reflectance bands, it may be applicable retroactively to Landsat 5 and 7 imagery, following standard preprocessing steps such as radiometric normalization, band harmonization, and sensor-specific corrections but overall transferability still needs to be evaluated. This makes Model_1i and its close variants well suited for long-term monitoring of IML evolution, drainage frequency, and seasonal hydrologic behavior. The low omission errors and high agreement ensure that rapid drainage events are captured reliably. When applied across multi-decadal Landsat records, these models enable the construction of a historical catalog of IML drainages and, hence, their dynamics, providing a foundation for quantifying long-term trends in IML stability, drainage recurrence, and hydrological connectivity. When combined with other remote-sensing datasets, the clustering-to-random-forest outputs can also support GLOF identification, enabling comparisons of past and present drainage events to better understand how the retreat of the GrIS is impacting these events.

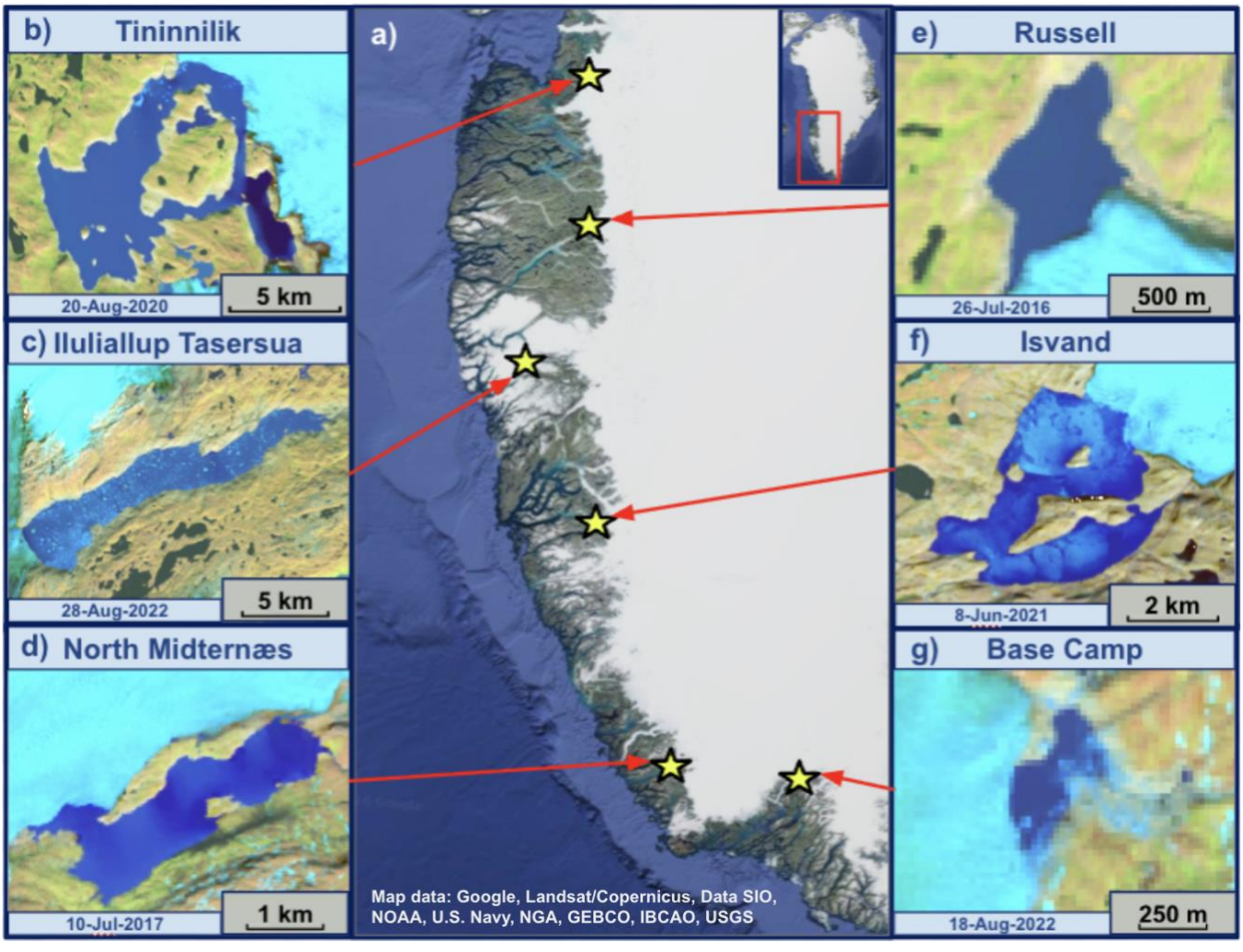
Different clustering-to-random-forest configurations exhibit distinct error structures, allowing users to select models with specific operational priorities. Models such as Model_7i and Model_5i minimize omission errors, so are well suited for applications where complete water detection is critical. Conversely, models like Model_9n and Model_14n minimize commission errors, so are better suited for precise boundary delineation. The top-ranked models strike a balance between these extremes, making them broadly applicable across most IML monitoring tasks.

The expert intercomparison dataset provides additional context for interpreting model performance. Manual delineation at the 30 m resolution of Landsat is inherently ambiguous, as shown by frequent expert disagreement, which is particularly noticeable along IML and glacier ice margins, in areas with both lake ice and glacier ice, and across debris-covered ice (Fig. 9). This disagreement is spatially structured rather than random, limiting the reliability of manual interpretation for region-wide or time-series analyses. A well-trained classifier can learn the stable spectral patterns that drive expert consensus and apply those rules consistently where experts diverge, producing repeatable masks that align with human judgment where it is most confident and reducing subjective variability where expert agreement is weakest. The comparison in Figure 8 illustrates this outcome: Model_1i aligns most closely with the true IML boundary, WRI tends to overestimate the IML extent, and NDWI systematically underestimates it.

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

5 Conclusions

This study demonstrates that clustering-to-random-forest approaches, and in particular the model we call 'Model_1i', which only uses Landsat 8's seven reflectance bands within the training data, provide the most consistent, accurate, and operationally scalable method for mapping IMLs. Model_1i balanced low omission and commission errors, outperformed spectral indices including the widely used NDWI, and matched or exceeded inter-expert agreement, making it the preferred method for generating a Greenland-wide Landsat-era IML dataset. When computational or data constraints necessitate index-based methods, practitioners should select WRI for comprehensive detection or ILI for conservative boundary mapping. Future work should prioritize testing geographic and seasonal transferability, incorporating temporal and ancillary data, and evaluating ensemble strategies to reduce ambiguity at the margins further and improve operational GLOF detection. It is also stressed that a model trained on Landsat 8 bands can be applied to earlier Landsat missions provided appropriate preprocessing: band-to-band harmonization, radiometric normalization, and any necessary spectral resampling or sensor-specific corrections are applied so that the input features match the training data. With those standard harmonization steps, the simplicity of Model_1i's band set makes it amenable to retrospective, Landsat-era mapping across multiple sensors.



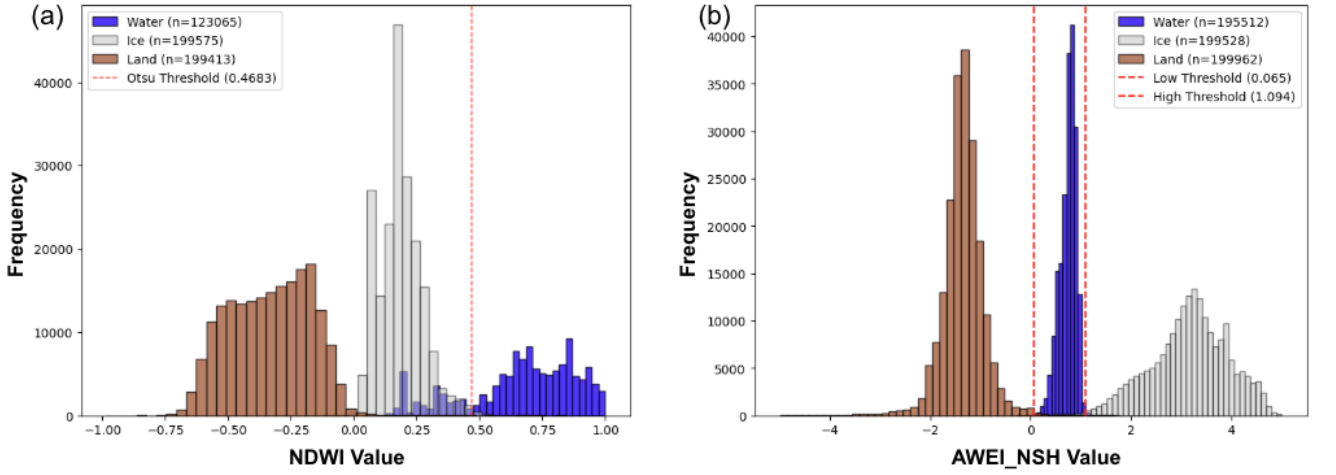
https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Figure 1. Study area and ice-marginal lake (IML) test sites in southwest Greenland. (a) Overview of the study area showing the locations of the six test IMLs (b–g). Map data: Google, Landsat/Copernicus, Data SIO, NOAA, U.S. Navy, NGA, GEBCO, IBCAO, U.S. Geological Survey. Imagery dates 14 December 2015 – 1 January 2021; some imagery has unknown acquisition dates. Image captured from Google Earth on 16 May 2026 (Google Earth, 2026).

Spectral Index	Abbreviation	Formula	Original Source
Normalized Difference Water Index	NDWI	$\frac{SR_B3 - SR_B5}{SR_B3 + SR_B5}$	Gao, 1996; Mcfeeters, 1996
Modified Normalized Difference Water Index	MNDWI	$\frac{SR_B3 - SR_B6}{SR_B3 + SR_B6}$	Xu, 2006
Augmented Normalized Difference Water Index	ANDWI	$\frac{SR_B2 + SR_B3 + SR_B4 - SR_B5 - SR_B6 - SR_B7}{SR_B2 + SR_B3 + SR_B4 + SR_B5 + SR_B6 + SR_B7}$	Rad et al., 2021
Multi-Band Water Index	MBWI	$(2 \times SR_B3) - SR_B4 - SR_B5 - SR_B6 - SR_B7$	Wang et al., 2018
Water Ratio Index	WRI	$\frac{SR_B3 - SR_B4}{SR_B5 + SR_B6}$	Shen and Li, 2010
Automated Water Extraction Index No Shadow	AWEI_NSH	$4 \times (SR_B3 - SR_B6) - (0.25 \times SR_B5 + 2.75 \times SR_B7)$	Feyisa, 2014
Automated Water Extraction Index Shadow	AWEI_SH	$SR_B2 + (2.5 \times SR_B3) - (1.5 \times (SR_B5 + SR_B6)) - (0.25 \times SR_B7)$	Feyisa, 2014
Adjusted Normalized Difference Snow Index	ANDSI	$\frac{(SR_B5 \times (SR_B3 + SR_B6)) - (SR_B7 \times (SR_B3 - SR_B6))}{(SR_B5 \times (SR_B3 + SR_B6)) + (SR_B7 \times (SR_B3 - SR_B6))}$	Mohammadi et al., 2023
Snow Water Index	SWI	$\frac{SR_B3 \times (SR_B5 - SR_B6))}{(SR_B3 + SR_B5) \times (SR_B5 + SR_B6))}$	Barbieux et al., 2018
Icy Lakes Index	ILI	$\frac{SR_B4 - SR_B7}{SR_B5 + SR_B6}$	Dixit et al., 2019

Table 1: Spectral indices tested in this study: Names, abbreviations, Landsat 8 formulas, and original citations are given. All formulas use Landsat 8 surface reflectance bands (SR_B#), listed with their corresponding color and wavelength range: SR_B2 (blue; 0.45 - 0.51 µm), SR_B3 (green; 0.53 - 0.59 µm), SR_B4 (red; 0.64 - 0.67 µm), SR_B5 (near-infrared; 0.85 - 0.88 µm), SR_B6 (shortwave infrared 1; 1.57 - 1.65 µm), SR_B7 (shortwave infrared 2; 2.11 - 2.29 µm).

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



(c)	Otsu Threshold Values
NDWI > 0.4683	0.065 < AWEI_NSH < 1.094
MNDWI > 0.9364	0.145 < AWEI_SH < 0.879
ANDWI > 0.8361	ANDSI < 0.882 OR 1.016 < ANDSI < 1.288
0.1072 < MBWI < 0.3716	ILI < 0.073 OR ILI > 3.0567
WRI < -0.0393 OR WRI > 3.2418	SWI < -0.251 OR 0.051 < SWI < 0.467 OR SWI > 0.740

Figure 2. Examples of histograms of sampled index values with Otsu thresholds (dashed red lines): (a) NDWI distribution showing the threshold that separates water from non-water, (b) AWEI_NSH distribution showing the threshold that distinguishes water from ice and land, and (c) Otsu threshold values for all spectral indices. See Table 1 for abbreviations of spectral indices.

Lake Ice Included in Training Data		Training Band Combinations
Yes	No	
Model_1i	Model_1n	SR_B1, SR_B2, SR_B3, SR_B4, SR_B5, SR_B6, SR_B7
Model_2i	Model_2n	SR_B1, SR_B2, SR_B3, SR_B4, SR_B5, SR_B6, SR_B7, NDWI, MNDWI, ANDWI, MBWI, WRI, AWEI_NSH, AWEI_SH, ANDSI, SWI, ILI
Model_3i	Model_3n	SR_B1, SR_B2, SR_B3, SR_B4, SR_B5, SR_B6, SR_B7, ANDWI, MBWI, WRI, AWEI_NSH, AWEI_SH, ANDSI, SWI, ILI
Model_4i	Model_4n	SR_B1, SR_B2, SR_B3, SR_B4, SR_B5, SR_B6, SR_B7, ANDWI, MBWI, WRI, AWEI_NSH, AWEI_SH, ILI
Model_5i	Model_5n	SR_B1, SR_B2, SR_B3, SR_B4, SR_B5, SR_B6, SR_B7, ANDWI, WRI, AWEI_NSH, AWEI_SH, ILI
Model_6i	Model_6n	SR_B1, SR_B3, SR_B4, SR_B6, SR_B7, ANDWI, AWEI_SH
Model_7i	Model_7n	SR_B1, SR_B3, SR_B4, SR_B7, ANDWI, AWEI_SH
Model_8i	Model_8n	SR_B1, SR_B2, SR_B3, SR_B4, SR_B5, SR_B6, SR_B7, WRI, AWEI_NSH, AWEI_SH, ILI
Model_9i	Model_9n	SR_B2, SR_B3, SR_B6, MBWI, AWEI_NSH
Model_10i	Model_10n	SR_B1, SR_B2, SR_B3, SR_B4, SR_B6, SR_B7, MBWI, WRI, AWEI_NSH, AWEI_SH, SWI
Model_11i	Model_11n	SR_B6
Model_12i	Model_12n	SR_B3, SR_B6
Model_13i	Model_13n	SR_B6, SR_B7
Model_14i	Model_14n	SR_B3, SR_B6, SR_B7

Table 2: Clustering-to-random-forest models tested in this study: Displaying training band combinations and whether lake ice was included within the training data. See Table 1 for abbreviations of spectral indices.

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

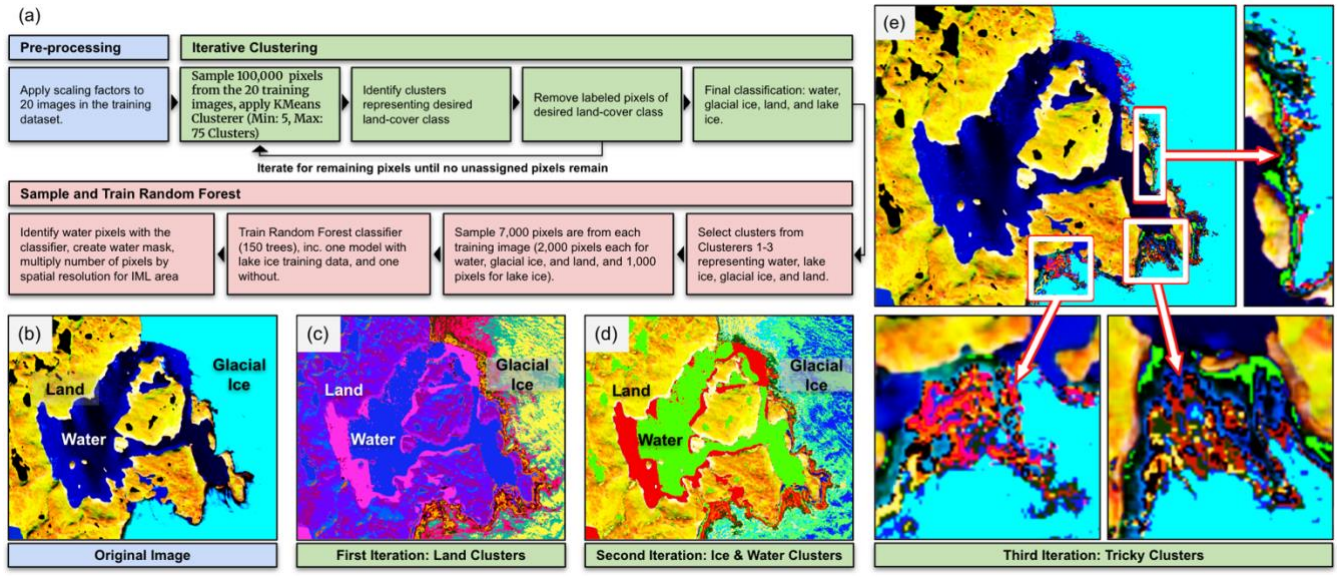


Figure 3: Clustering-to-random-forest workflow: (a) Supervised clustering and random forest training workflow (modified from Dell et al., 2022) for the clustering-to-random-forest methods using Lake Tininnilik as an example. (b) Landsat 8 image of Lake Tininnilik before clustering. (c) First clustering iteration applied to the full image; from this output, pixels from land clusters are identified, labeled, and removed from future iterations. (d) Second clustering iteration, which is restricted to the remaining pixels (non-land); pixels from ice and water clusters are identified, labeled, and removed from future iterations. (e) Final clustering iteration, applied to the remaining unclassified pixels, isolates pixels from ambiguous "tricky" clusters (dirty ice and mixed water), roughly outlined in red, to refine boundary precision.

Metric	Abbreviation	Importance (value)	Weight
high confidence Kappa coefficient	Kappa_HC	highest (5)	19.23 %
high confidence omission error	OE_HC	highest (5)	19.23 %
high confidence commission error	CE_HC	high (4)	15.38 %
stratified random sample pixel weighted accuracy	PWA_SRS	medium-high (3)	11.54 %
Kappa coefficient	Kappa	medium-high (3)	11.54 %
high confidence overall accuracy	OA_HC	medium (2)	7.69 %
overall accuracy	OA	low (1)	3.85 %
stratified random sample overall accuracy	OA_SRS	low (1)	3.85 %
omission error	OE	low (1)	3.85 %

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

commission error	CE	low (1)	3.85 %

Table 3: Weighted rank-sum metrics employed in this study: The metric, abbreviations, level of importance, and weighted rank associated with each metric used in determining the overall rankings for the methods evaluated in this study.

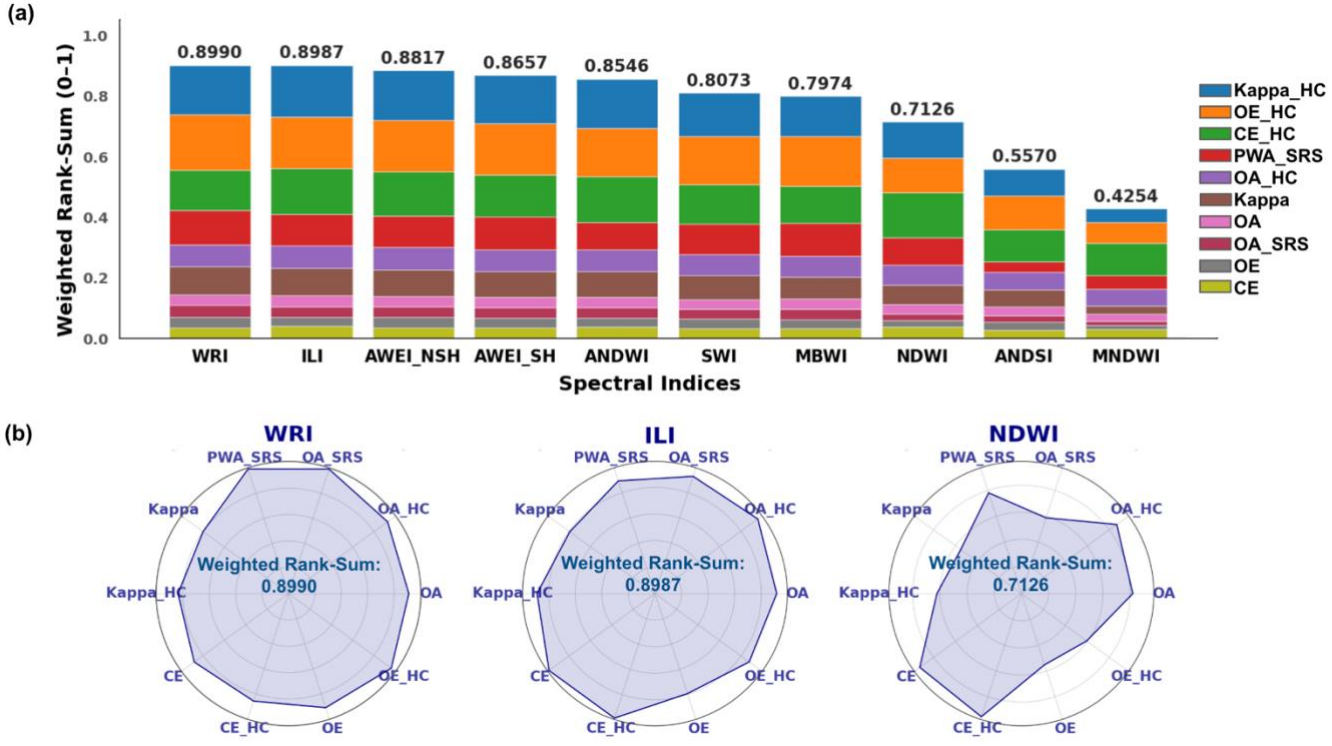


Figure 4: Spectral index weighted rank-sum results: (a) Bar chart showing weighted rank-sum values for all indices, illustrating the relative influence of each metric on the final composite score. (b) Spider plots showing the metric-level results for the two top-performing indices, WRI and ILI, alongside the commonly used NDWI. Each axis represents an individual metric and is linearly scaled from the plot center to the outer perimeter. For the metrics Kappa_HC, PWA_SRS, OA_HC, Kappa, OA, OA_SRS, the center represents 0 and the perimeter represents 1, whereas the error-based metrics CE_HC, OE_HC, CE, and OE are inverted, with the center representing 1 and the outer perimeter representing 0. For all metrics, values closer to the perimeter indicate better performance, with the composite score stated in the center of each spider plot.

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

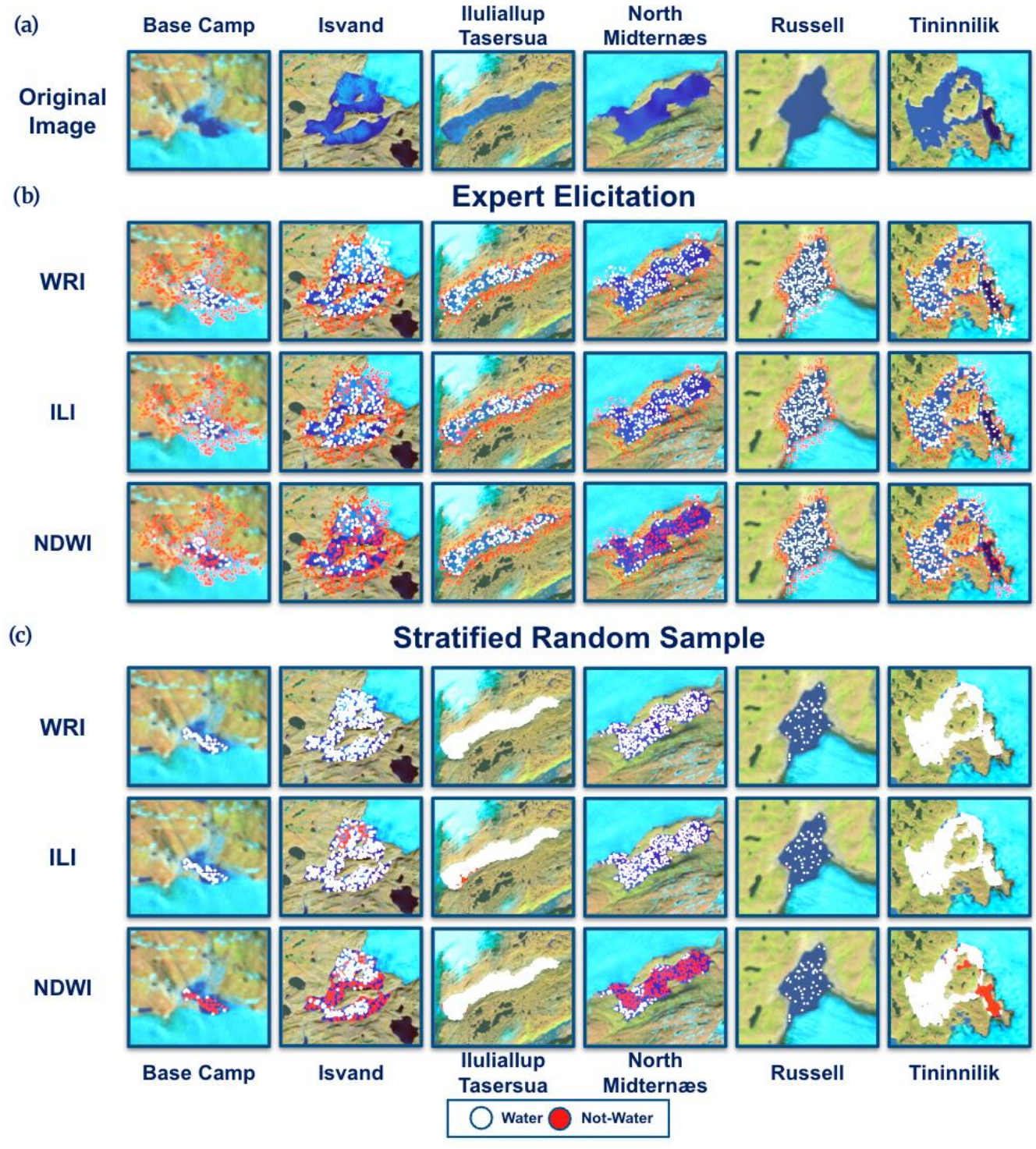
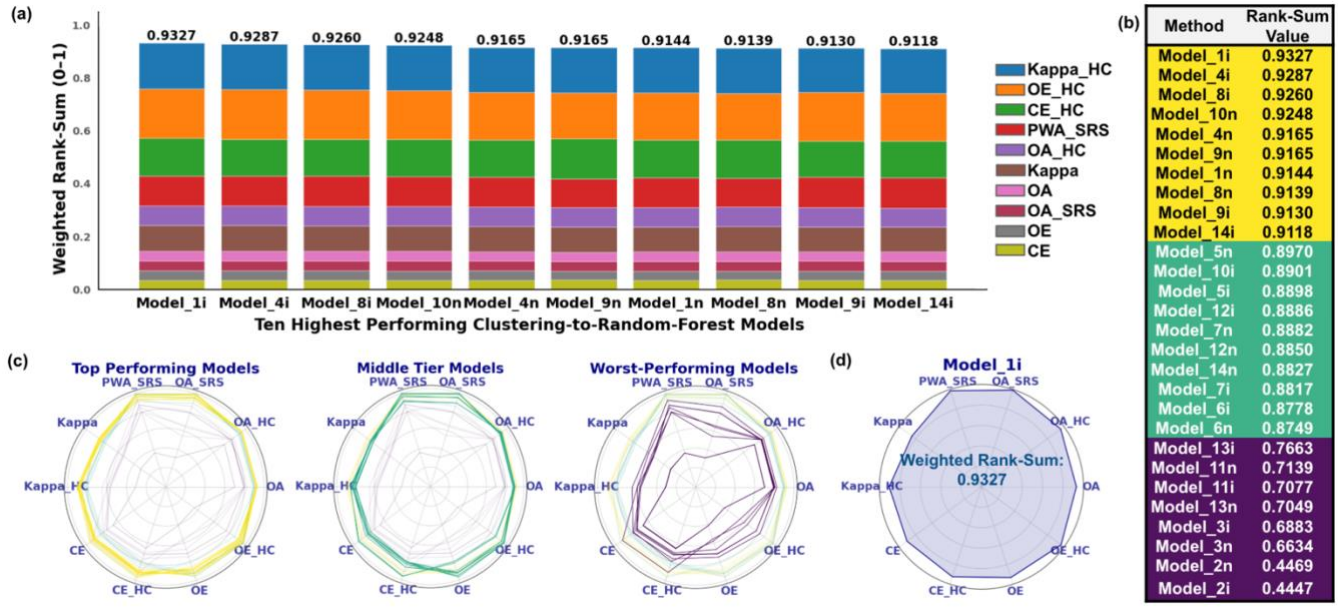


Figure 5: Pixel-level assessment of spectral index performance for each IML: (a) Original false-color Landsat 8 image for each test IML. (b) Pixels used in the expert-elicitation dataset, where white pixels indicate locations the WRI, ILI, or NDWI mask classified as water and red pixels indicate locations classified as non-water. Within the IML boundaries, red pixels represent omission errors

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

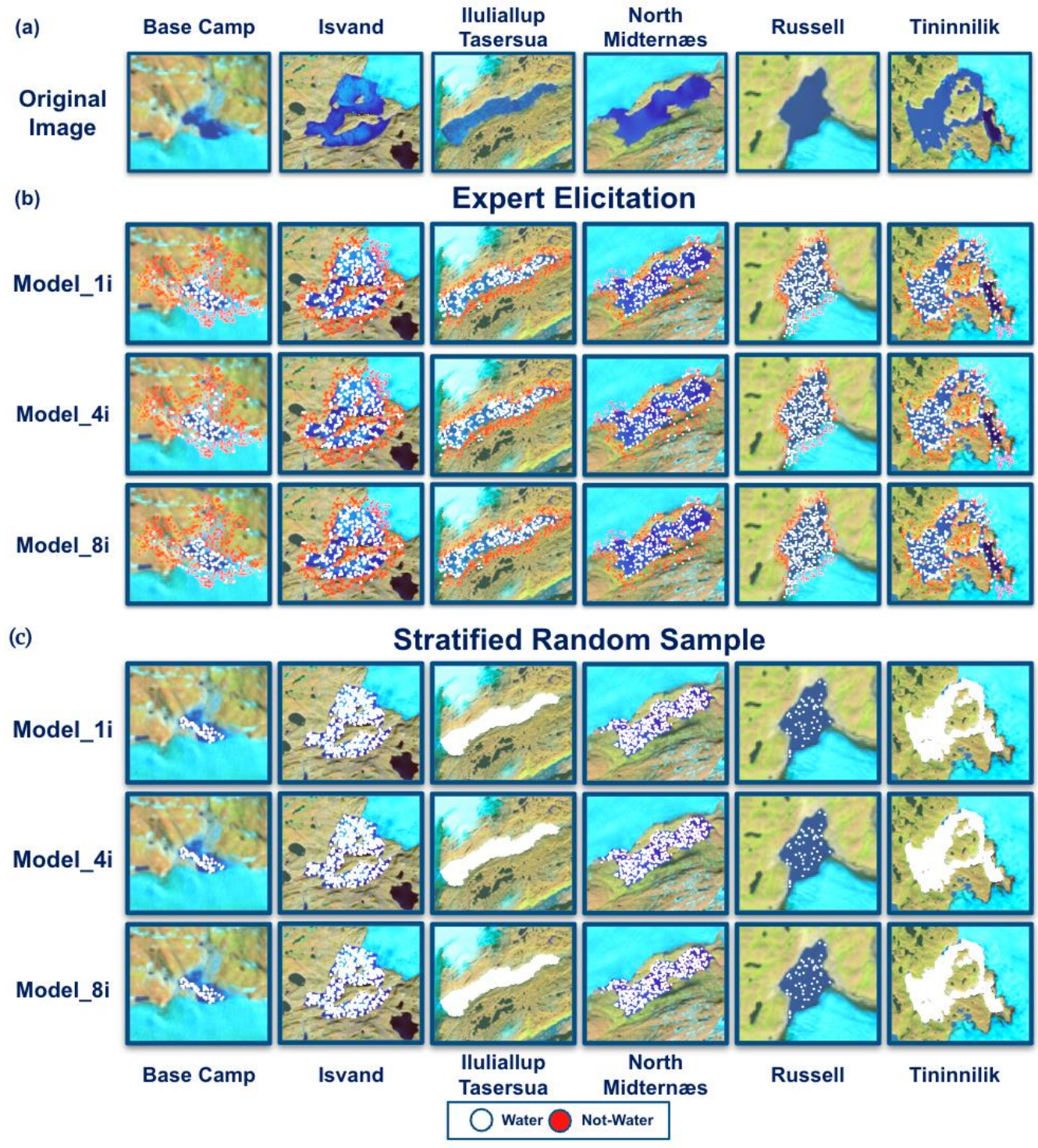
(water misclassified as non-water), while white pixels outside the lake boundary represent commission errors (non-water misclassified as water). (c) Pixels used in the stratified random-sample dataset, where all sampled pixels are known water pixels. White pixels indicate correct classifications by the spectral-index mask, and red pixels indicate omission errors where the mask incorrectly classified a water pixel as non-water.



Method	Rank-Sum Value
Model_1i	0.9327
Model_4i	0.9287
Model_8i	0.9260
Model_10n	0.9248
Model_4n	0.9165
Model_9n	0.9165
Model_1n	0.9144
Model_8n	0.9139
Model_9i	0.9130
Model_14i	0.9118
Model_5n	0.8970
Model_10i	0.8901
Model_5i	0.8898
Model_12i	0.8886
Model_7n	0.8882
Model_12n	0.8850
Model_14n	0.8827
Model_7i	0.8817
Model_6i	0.8778
Model_6n	0.8749
Model_13i	0.7663
Model_11n	0.7139
Model_11i	0.7077
Model_13n	0.7049
Model_3i	0.6883
Model_3n	0.6634
Model_2n	0.4469
Model_2i	0.4447

Figure 6: Clustering-to-random-forest weighted rank-sum results: (a) Bar chart summarizing weighted rank-sum values for the top-performing clustering-to-random-forest models, illustrating the relative influence of each metric on the final composite score. (b) Table displaying the weighted rank-sum value of each method; color is associated with the spider plots in (c), the 10 best-performing clustering-to-random-forest methods in yellow, the 10 middle-tier methods in teal green, and the eight worst-performing methods in purple. (c) Combined spider plots; the similar shapes indicate that the top-performing clustering-to-random-forest methods performed similarly. Each axis represents an individual metric and is linearly scaled from the plot center to the outer perimeter. For the metrics Kappa_HC, PWA_SRS, OA_HC, Kappa, OA, OA_SRS, the center represents 0 and the perimeter represents 1, whereas the error-based metrics CE_HC, OE_HC, CE, and OE are inverted, with the center representing 1 and the outer perimeter representing 0. For all metrics, values closer to the perimeter indicate better performance, with the composite score stated in the center of each spider plot. (d) Spider plot for the best-performing clustering-to-random-forest model, Model_1i. Error-based metrics are inverted so that values closer to the plot perimeter indicate better performance.

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



460 **Figure 7: Pixel-level assessment of clustering-to-random-forest model performance for each IML: (a) Original false-color Landsat 8 image for each test IML. (b) Pixels used in the expert-elicitation dataset, where white pixels indicate locations where Model_1i, Model_4i, or Model_8i classified pixels as water and red pixels indicate locations classified as non-water. Within the**

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

IML boundaries, red pixels represent omission errors (water misclassified as non-water), while white pixels outside the lake boundary represent commission errors (non-water misclassified as water). (c) Pixels used in the stratified random-sample dataset, where all sampled pixels are known water pixels. White pixels indicate correct classifications by the spectral-index mask, and red pixels indicate omission errors where the mask incorrectly classified a water pixel as non-water.

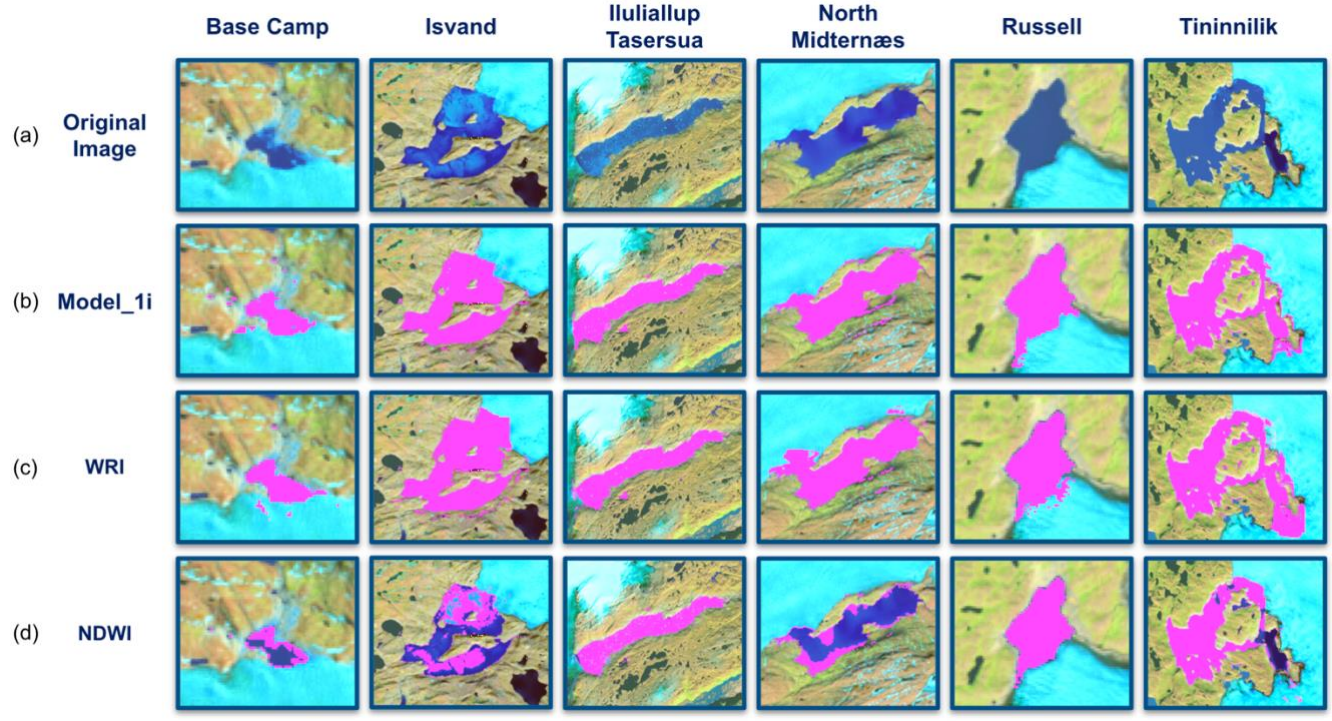


Figure 8: Application of water classification masks from three methods to original images for each of the six study lakes: (a) The original Landsat 8 image; and water masks produced by (b) the top-performing ML method, Model_1i; (c) the best performing spectral index, WRI; and d) the commonly used spectral index NDWI, with magenta representing areas of water.

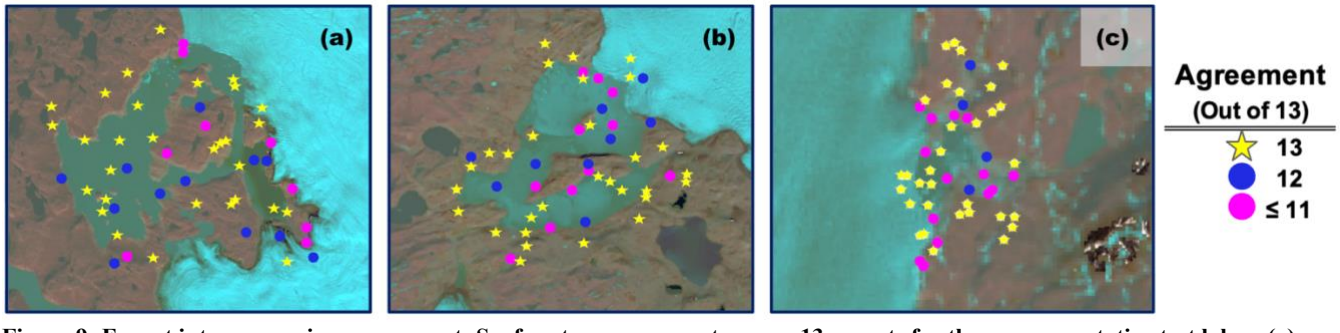


Figure 9: Expert intercomparison agreement: Surface type agreement among 13 experts for three representative test lakes: (a) Tininnilik, (b) Isvand, and (c) Base Camp. Gold Stars denote full expert agreement (13/13), blue circles indicate agreement among 12/13 experts, and magenta circles indicate agreement between 11/13 experts or fewer.

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Agreement Amongst Experts (50 pixels per lake)								
	Base Camp	Iluliallup Tasersua	Isvand	North Midternæs	Russell	Tininnilik	Total	Mean Agreement (%)
Surface type agreement	32	38	30	44	43	30	217	72
Surface type agreement and high confidence	25	29	14	37	33	24	162	54
Binary surface type agreement	42	41	41	47	44	36	251	84
Binary surface type agreement and high confidence	25	29	15	37	33	24	163	54

Table 4: Expert intercomparison agreement: Summary of expert agreement across all six test lakes, showing mean surface type agreement and binary surface type agreement, along with corresponding agreement levels when restricted to high-confidence interpretations.

Data availability

Landsat 8 OLI surface reflectance imagery was accessed through Google Earth Engine using the USGS Collection 2 Level-2 dataset. These collections are available directly within Google Earth Engine under the following identifier: LANDSAT/LC08/C02/T1_L2. All image processing and compositing were performed within the Google Earth Engine environment.

Supplement link

The supplement is available online.

Author contributions

EC conceptualized the research, carried out the main body of the research, and drafted the manuscript. MS, AFB, and MAD advised on all aspects of the project and assisted with editing of the manuscript. All other co-authors participated in the expert elicitation of the validation data and review of the manuscript.

Competing interests

The authors declare no conflicts of interest.

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



Disclaimer

Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements

Computation and data management were supported in part by CryoCloud, a cloud-based cyberinfrastructure platform for cryosphere research (NASA grant 80NSSC25K7940 and 80NSSC22K1877).

Financial support

EC was supported by the National Science Foundation Graduate Research Fellowship Program (NSF GRFP) under grant DGE-2444107. AFB was supported by NSF grant #2213702 (PI Banwell) awarded to the University of Colorado Boulder. P.H. acknowledges support from the ESA (European Space Agency) Living Planet Fellowship (4000136382/21/I-DT-lr). K.K.K. and J.K. acknowledge support from the Independent Research Fund in Denmark (grant ID 10.46540/3103-00234B). JM was supported by the UK Research and Innovation (UKRI) Canada-Inuit Nunangat-United Kingdom (CINUK) Arctic Research Programme IMAGINE Project (grant NE/X003868/1). DS was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, 2020-04207 and the Killam Foundation. This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-25-C-B010 in the amount of $144,500 to MS. AK was supported by IPCC fellowship 2026 and the Cuomo Foundation.

References

AMAP: AMAP Arctic Climate Change Update 2021: Key Trends and Impacts, Arctic Monitoring and Assessment Programme, Tromsø, Norway, viii+148 pp., available at: https://www.amap.no/documents/doc/amap-arctic-climate-change-update-2021-key-trends-and-impacts/3594, last access: 9 March 2026, 2021.

Arthur, D. and Sergei, V.: k-means++: The advantages of careful seeding, Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms, 7, 1027–1035, 2007.

Barbieux, K., Charitsi, A., and Merminod, B.: Icy lakes extraction and water-ice classification using Landsat 8 OLI multispectral data, International Journal of Remote Sensing, 39, 3646–3678, https://doi.org/10.1080/01431161.2018.1447165, 2018.

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Box, J. E., Hubbard, A., Bahr, D. B., Colgan, W. T., Fettweis, X., Mankoff, K. D., Wehrlé, A., Noël, B., Van Den Broeke, M. R., Wouters, B., Bjørk, A. A., and Fausto, R. S.: Greenland ice sheet climate disequilibrium and committed sea-level rise, Nat. Clim. Chang., 12, 808–813, https://doi.org/10.1038/s41558-022-01441-2, 2022.

Carr, E. P.: Greenland Glacial Lake Outburst Floods: Identifying Drainage Events and Assessing Trends in Frequency, Time of Year, and Magnitude in the Landsat Satellite Era, M.A. thesis, University of Colorado Boulder, Boulder, CO, USA, 2024.

Carrivick, J. L. and Quincey, D. J.: Progressive increase in number and volume of ice-marginal lakes on the western margin of the Greenland Ice Sheet, Global and Planetary Change, 116, 156–163, https://doi.org/10.1016/j.gloplacha.2014.02.009, 2014.

Carrivick, J. L. and Tweed, F. S.: A review of glacier outburst floods in Iceland and Greenland with a megafloods perspective, Earth-Science Reviews, 196, 102876, https://doi.org/10.1016/j.earscirev.2019.102876, 2019.

Carrivick, J. L., Turner, A. G. D., Russell, A. J., Ingeman-Nielsen, T., and Yde, J. C.: Outburst flood evolution at Russell Glacier, western Greenland: effects of a bedrock channel cascade with intermediary lakes, Quaternary Science Reviews, 67, 39–58, https://doi.org/10.1016/j.quascirev.2013.01.023, 2013.

Carrivick, J. L., How, P., Lea, J. M., Sutherland, J. L., Grimes, M., Tweed, F. S., Cornford, S., Quincey, D. J., and Mallalieu, J.: Ice-Marginal Proglacial Lakes Across Greenland: Present Status and a Possible Future, Geophysical Research Letters, 49, e2022GL099276, https://doi.org/10.1029/2022GL099276, 2022.

Chaaban, F., El Khattabi, J., and Darwishe, H.: Accuracy Assessment of ESA WorldCover 2020 and ESRI 2020 Land Cover Maps for a Region in Syria, J geovis spat anal, 6, 31, https://doi.org/10.1007/s41651-022-00126-w, 2022.

Chen, J., Wang, Y., Wang, J., Zhang, Y., Xu, Y., Yang, O., Zhang, R., Wang, J., Wang, Z., Lu, F., and Hu, Z.: The Performance of Landsat-8 and Landsat-9 Data for Water Body Extraction Based on Various Water Indices: A Comparative Analysis, Remote Sensing, 16, 1984, https://doi.org/10.3390/rs16111984, 2024.

Dell, R. L., Banwell, A. F., Willis, I. C., Arnold, N. S., Halberstadt, A. R. W., Chudley, T. R., and Pritchard, H. D.: Supervised classification of slush and ponded water on Antarctic ice shelves using Landsat 8 imagery, J. Glaciol., 68, 401–414, https://doi.org/10.1017/jog.2021.114, 2022.

Dell, R. L., Willis, I. C., Arnold, N. S., Banwell, A. F., and De Roda Husman, S.: Substantial contribution of slush to meltwater area across Antarctic ice shelves, Nat. Geosci., 17, 624–630, https://doi.org/10.1038/s41561-024-01466-6, 2024.

Dixit, A., Goswami, A., and Jain, S.: Development and Evaluation of a New “Snow Water Index (SWI)” for Accurate Snow Cover Delineation, Remote Sensing, 11, 2774, https://doi.org/10.3390/rs11232774, 2019.

Dømgaard, M., Kjeldsen, K. K., Huiban, F., Carrivick, J. L., Khan, S. A., and Bjørk, A. A.: Recent changes in drainage route and outburst magnitude of Russell Glacier ice-dammed lake, West Greenland, https://doi.org/10.5194/egusphere-2022-566, 3 August 2022., 2022.

Dømgaard, M., Kjeldsen, K., How, P., and Bjørk, A.: Altimetry-based ice-marginal lake water level changes in Greenland, Commun Earth Environ, 5, 365, https://doi.org/10.1038/s43247-024-01522-4, 2024.

Earth Resources Observation and Science (EROS) Center: Landsat 8-9 Operational Land Imager / Thermal Infrared Sensor Level-2, Collection 2, U.S. Geological Survey [data set], https://doi.org/10.5066/P9OGBGM6, 2020.

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

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.

Gao, B.: NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space, Remote Sensing of Environment, 58, 257–266, https://doi.org/10.1016/S0034-4257(96)00067-3, 1996.

Google: Google Earth, https://earth.google.com, last access: 16 May 2026.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R.: Google Earth Engine: Planetary-scale geospatial analysis for everyone, Remote Sensing of Environment, 202, 18–27, https://doi.org/10.1016/j.rse.2017.06.031, 2017.

Halberstadt, A. R. W., Gleason, C. J., Moussavi, M. S., Pope, A., Trusel, L. D., and DeConto, R. M.: Antarctic Supraglacial Lake Identification Using Landsat-8 Image Classification, Remote Sensing, 12, 1327, https://doi.org/10.3390/rs12081327, 2020.

Hanna, E., Box, J. E., Fettweis, X., Van Den Broeke, M. R., Wake, L., Luu, L., Maddison, J., Mote, T., Mottram, R., Noël, B., Preece, J., Topál, D., Trusel, L., and Van Der Laan, L. N.: Past, present and future Greenland Ice Sheet surface melt, 1500–2200 CE, Nat Rev Earth Environ, https://doi.org/10.1038/s43017-026-00800-3, 2026.

Harpur, C., Carrivick, J. L., Sutherland, J. L., and Mallalieu, J.: The emerging importance of ice-marginal lakes across Greenland, Geography, 110, 6–15, https://doi.org/10.1080/00167487.2024.2437946, 2025.

Hengst, A. M., Armstrong, W., Rick, B., and McGrath, D.: Short-Term Variability in Alaska Ice-Marginal Lake Area: Implications for Long-Term Studies, Remote Sensing, 13, 3955, https://doi.org/10.3390/rs13193955, 2021.

How, P., Messerli, A., Mätzler, E., Santoro, M., Wiesmann, A., Caduff, R., Langley, K., Bojesen, M. H., Paul, F., Kääb, A., and Carrivick, J. L.: Greenland-wide inventory of ice marginal lakes using a multi-method approach, Sci Rep, 11, 4481, https://doi.org/10.1038/s41598-021-83509-1, 2021.

How, P., Petersen, D., Kjeldsen, K. K., Raundrup, K., Karlsson, N. B., Messerli, A., Rutishauser, A., Carrivick, J. L., Lea, J. M., Fausto, R. S., Ahlstrøm, A. P., and Andersen, S. B.: The Greenland Ice-Marginal Lake Inventory Series from 2016 to 2023, Earth Syst. Sci. Data, 17, 6331–6351, https://doi.org/10.5194/essd-17-6331-2025, 2025.

Kjeldsen, K. K., Mortensen, J., Bendtsen, J., Petersen, D., Lennert, K., and Rysgaard, S.: Ice-dammed lake drainage cools and raises surface salinities in a tidewater outlet glacier fjord, west Greenland: GODTHÅBSFJORD, J. Geophys. Res. Earth Surf., 119, 1310–1321, https://doi.org/10.1002/2013JF003034, 2014.

Kjeldsen, K. K., Khan, S. A., Bjørk, A. A., Nielsen, K., and Mouginot, J.: Ice-dammed lake drainage in west Greenland: Drainage pattern and implications on ice flow and bedrock motion, Geophysical Research Letters, 44, 7320–7327, https://doi.org/10.1002/2017GL074081, 2017.

Lützow, N. and Veh, G.: Glacier Lake Outburst Flood Database, https://doi.org/10.5281/ZENODO.7330344, 2023.

Lützow, N., Veh, G., and Korup, O.: A global database of historic glacier lake outburst floods, Earth Syst. Sci. Data, 15, 2983–3000, https://doi.org/10.5194/essd-15-2983-2023, 2023.

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Maciel, D. A., Pahlevan, N., Barbosa, C. C. F., De Novo, E. M. L. D. M., Paulino, R. S., Martins, V. S., Vermote, E., and Crawford, C. J.: Validity of the Landsat surface reflectance archive for aquatic science: Implications for cloud-based analysis, Limnol Oceanogr Letters, 8, 850–858, https://doi.org/10.1002/lol2.10344, 2023.

Mallalieu, J., Carrivick, J. L., Quincey, D. J., and Raby, C. L.: Ice-marginal lakes associated with enhanced recession of the Greenland Ice Sheet, Global and Planetary Change, 202, 103503, https://doi.org/10.1016/j.gloplacha.2021.103503, 2021.

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.

Mohammadi, B., Pilesjö, P., and Duan, Z.: The superiority of the Adjusted Normalized Difference Snow Index (ANDSI) for mapping glaciers using Sentinel-2 multispectral satellite imagery, GIScience & Remote Sensing, 60, 2257978, https://doi.org/10.1080/15481603.2023.2257978, 2023.

Oqaasileriffik (Language Secretariat of Greenland): Nunat Aqqi – Greenlandic place names, https://nunataqqi.gl/, last access: 2 September 2026.

Otsu, N.: A Threshold Selection Method from Gray-Level Histograms, IEEE Trans. Syst., Man, Cybern., 9, 62–66, https://doi.org/10.1109/TSMC.1979.4310076, 1979.

Park, J.-Y., Schloesser, F., Timmermann, A., Choudhury, D., Lee, J.-Y., and Nellikkattil, A. B.: Future sea-level projections with a coupled atmosphere-ocean-ice-sheet model, Nat Commun, 14, 636, https://doi.org/10.1038/s41467-023-36051-9, 2023.

Rad, A. M., Kreitler, J., and Sadegh, M.: Augmented Normalized Difference Water Index for improved surface water monitoring, Environmental Modelling & Software, 140, 105030, https://doi.org/10.1016/j.envsoft.2021.105030, 2021.

Rosser, J. F., Leibovici, D. G., and Jackson, M. J.: Rapid flood inundation mapping using social media, remote sensing and topographic data, Nat Hazards, 87, 103–120, https://doi.org/10.1007/s11069-017-2755-0, 2017.

Schiøtt, S., Tejsner, P., and Rysgaard, S.: Inuit and Local Knowledge on the Marine Ecosystem in Ilulissat Icefjord, Greenland, Hum Ecol, 50, 167–181, https://doi.org/10.1007/s10745-021-00277-2, 2022.

Shen, L. and Li, C.: Water body extraction from Landsat ETM+ imagery using adaboost algorithm, in: 2010 18th International Conference on Geoinformatics, 1–4, https://doi.org/10.1109/GEOINFORMATICS.2010.5567762, 2010.

Shetty, S., Gupta, P. K., Belgiu, M., and Srivastav, S. K.: Assessing the Effect of Training Sampling Design on the Performance of Machine Learning Classifiers for Land Cover Mapping Using Multi-Temporal Remote Sensing Data and Google Earth Engine, Remote Sensing, 13, 1433, https://doi.org/10.3390/rs13081433, 2021.

Shugar, D. H., Burr, A., Haritashya, U. K., Kargel, J. S., Watson, C. S., Kennedy, M. C., Bevington, A. R., Betts, R. A., Harrison, S., and Strattman, K.: Rapid worldwide growth of glacial lakes since 1990, Nat. Clim. Chang., 10, 939–945, https://doi.org/10.1038/s41558-020-0855-4, 2020.

Song, C., Fan, C., Ma, J., Zhan, P., and Deng, X.: A spatially constrained remote sensing-based inventory of glacial lakes worldwide, Sci Data, 12, 464, https://doi.org/10.1038/s41597-025-04809-z, 2025.

https://doi.org/10.5194/egusphere-2026-5265
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



Stehman, S. V.: Sampling designs for accuracy assessment of land cover, International Journal of Remote Sensing, 30, 5243–5272, https://doi.org/10.1080/01431160903131000, 2009.

Stokes, C. R., Bamber, J. L., Dutton, A., and DeConto, R. M.: Warming of +1.5 °C is too high for polar ice sheets, Commun Earth Environ, 6, 351, https://doi.org/10.1038/s43247-025-02299-w, 2025.

Wang, X., Xie, S., Zhang, X., Chen, C., Guo, H., Du, J., and Duan, Z.: A robust Multi-Band Water Index (MBWI) for automated extraction of surface water from Landsat 8 OLI imagery, International Journal of Applied Earth Observation and Geoinformation, 68, 73–91, https://doi.org/10.1016/j.jag.2018.01.018, 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.

Yu, Y., Wang, C., Fu, Q., Kou, R., Huang, F., Yang, B., Yang, T., and Gao, M.: Techniques and Challenges of Image Segmentation: A Review, Electronics, 12, 1199, https://doi.org/10.3390/electronics12051199, 2023.

Zhang, G., Carrivick, J. L., Emmer, A., Shugar, D. H., Veh, G., Wang, X., Labedz, C., Mergili, M., Mölg, N., Huss, M., Allen, S., Sugiyama, S., and Lützow, N.: Characteristics and changes of glacial lakes and outburst floods, Nat Rev Earth Environ, 5, 447–462, https://doi.org/10.1038/s43017-024-00554-w, 2024.

Zhao, C., Wei, H., Feyisa, G. L., De Castro Tayer, T., Ma, G., Wu, H., and Pan, Y.: Evaluating spectral indices for water extraction: Limitations and contextual usage recommendations, International Journal of Applied Earth Observation and Geoinformation, 139, 104510, https://doi.org/10.1016/j.jag.2025.104510, 2025.