

DOI: 10.37943/23KUWI4163

**Daniker Chepashev**

Master's degree, PhD Candidate in Space Technic and Technologies, Head of Laboratory, Laboratory of Space Monitoring of Emergencies  
d.chepashev@ionos.kz, orcid.org/0000-0002-8417-3990  
LLP Ionosphere Institute, Kazakhstan

**Yelizaveta Maximkina**

Master's degree, Junior Researcher, Laboratory of Space Monitoring of Emergencies  
maximkina.y@ionos.kz, orcid.org/0009-0007-6925-1202  
LLP Ionosphere Institute, Kazakhstan

**Gulsezim Zhussupova**

Master's degree, Junior Researcher, Laboratory of Space Monitoring of Emergencies  
zhussupova@ionos.kz, orcid.org/0009-0009-5330-9697  
LLP Ionosphere Institute, Kazakhstan

**Ruslan Zhilkibayev**

Bachelor's degree, Engineer, Laboratory of Space Monitoring of Emergencies  
zhilkibayev@ionos.kz, orcid.org/0009-0006-2057-9197  
LLP Ionosphere Institute, Kazakhstan

**Dias Merekeyev**

Bachelor's degree, Engineer, Laboratory of Space Monitoring of Emergencies  
merekeyev.d@ionos.kz, orcid.org/0009-0000-1815-3046  
LLP Ionosphere Institute, Kazakhstan

## RECOGNITION OF THE WATER SURFACE ACCORDING TO ICEYE DATA USING MACHINE LEARNING

**Abstract:** The growing frequency of floods and the resulting socio-economic losses highlight the need for accurate and automated tools for detecting and monitoring water surfaces. This study presents a methodology for automatic water surface recognition based on high-resolution ICEYE synthetic aperture radar (SAR) data. The algorithm is implemented in the Google Earth Engine environment and uses the Random Forest machine-learning model trained on manually labeled “water” and “land” classes derived directly from the radar imagery. Preprocessing, performed in ESA SNAP, included radiometric calibration, Range-Doppler terrain correction, and speckle filtering to ensure accurate backscatter representation. The trained model was applied to ICEYE VV-polarized images acquired over Uralsk, Kazakhstan, on April 20–21, 2024, during a major regional flood.

To validate the results, the Random Forest-derived masks were compared with those obtained using traditional methods such as Otsu and fixed-threshold classification, as well as optical masks generated from Sentinel-2 NDWI and MNDWI indices. Quantitative evaluation showed an overall accuracy of 76.8 % and a kappa coefficient of 0.535, while the area under the ROC curve (AUC = 0.91) indicated strong discriminatory capability. The Random Forest model demonstrated greater spatial precision and reduced false-positive mapping compared to threshold-based methods, confirming its suitability for operational flood monitoring.

The proposed approach highlights the potential of ICEYE data for near-real-time water surface mapping, especially under cloud-covered conditions where optical sensors are ineffective. Moreover, the developed workflow ensures reproducibility and can be integrated into automated flood-response systems for rapid situation assessment. In the future, incorporating additional polarimetric and texture features is expected to further enhance model performance and extend its applicability to diverse hydrological environments.

**Keywords:** ICEYE, synthetic aperture radar, machine learning, Random Forest, decision trees, water recognition

### Introduction

The rapid increase in the frequency of hydrometeorological disasters, including floods, necessitates the development of accurate, timely, and automated methods for monitoring water surfaces [1]. Satellites equipped with Synthetic Aperture Radar (SAR) are among the most promising data sources for such tasks due to their high spatial resolution and independence from cloud cover and illumination conditions. In particular, the use of data from the private X-band sensor ICEYE significantly enhances monitoring capabilities through frequent revisit times and robustness against atmospheric disturbances.

A key challenge in processing radar imagery is accurate detection of water surfaces, which is complicated by speckle noise, radar shadows, and ambiguous reflections from urban areas or marshy soils. Traditional methods based on thresholding radar intensity values can achieve high accuracy in homogeneous areas but often lack flexibility and adaptability compared to machine learning algorithms. Within this context, the Random Forest algorithm has proven effective for binary classification with a limited number of features, demonstrating resilience to noise and overfitting.

The objective of this study is to develop and test an automated water surface recognition algorithm based on ICEYE VV-polarization data, implemented in the Google Earth Engine environment using the Random Forest algorithm. The study region encompasses the city of Uralsk in the West Kazakhstan region, which systematically experiences floods during spring thaw periods. The most devastating recent flood occurred in spring 2024. The satellite images used in this study were acquired within the framework of the Program-Targeted Financing project for the purpose of monitoring flooding in the vicinity of Uralsk. The images were purchased due to the lack of radar sensor observations over this region. This circumstance somewhat limits the dataset to only two scenes; however, each image covers an area of approximately 2,900 square kilometers, which compensates for the restricted number of acquisitions. The objective of the study is to develop and test an algorithm for water surface detection using radar data, which in the future will allow the approach to be extrapolated to larger datasets of scenes as well as to other types of sensors.

The model was trained in manually labeled "water" and "land" classes and validated on two consecutive images from April 20 and April 21, 2024. To evaluate the effectiveness of the method, water masks were also obtained using more traditional methods, such as the Otsu method and the threshold method. Additionally, classification results were compared with optical water masks generated using NDWI and MNDWI indices derived from Sentinel-2 imagery captured on the same dates, as well as assessed quantitatively using Out-of-Bag (OOB) accuracy metrics.

### Methods and Materials

A significant tool for flood forecasting and monitoring is water surface detection using Synthetic Aperture Radar (SAR) data. Optical sensors often prove less useful for flood monitoring due to persistent cloud coverage accompanying precipitation events, which substantially con-

tribute to excessive runoff. In contrast, radar satellites demonstrate robustness against cloud cover and effectively differentiate water surfaces due to their distinctive reflectivity.

Several key approaches can be identified among contemporary methods for water detection on radar imagery [2]. The first relies on pixel-based thresholding of  $\sigma^0$  (backscatter coefficient), as smooth water surfaces typically exhibit low backscatter, allowing rapid identification of water bodies. This method underpins WaMaPro and other simple algorithms. The second direction employs object-oriented image analysis, incorporating multi-level segmentation, hydrologically consistent refinement based on Digital Elevation Models (DEMs), and the generation of accurate water masks, even in heterogeneous scenes. An example of this approach is implemented in the RaMaFlood algorithm. The third approach integrates automatic classification with probabilistic frameworks, fuzzy post-processing, and the incorporation of multiple auxiliary datasets, such as slopes, land-cover types, and others. In all the methods described, overall accuracy typically exceeds 90% under simple scene conditions with high contrast and smooth water surfaces. However, factors complicating water-land classification, such as wind-induced waves or radar shadows from mountain ranges, necessitate incorporating additional datasets and increasing system complexity to mitigate unfavorable imaging conditions and artifacts.

Remote sensing literature on flood delineation from SAR has evolved from threshold-based segmentation toward fully automated processing pipelines and learning-based schemes that cope better with scene heterogeneity and acquisition variability. Early operational frameworks (e.g., the DLR “Water Suite”) combine simple backscatter thresholds, object-based post-processing, and hydrologic constraints, enabling near-real-time (NRT) products but still relying on scene-specific tuning in complex settings. Comparative analyses across these operational approaches show that while thresholding can excel in homogeneous, smooth-water scenes, its performance degrades with wind-roughening, emergent or floating vegetation, and urban double-bounce, leading to commission and omission errors that are not trivially removed without ancillary data [3].

A widely used Otsu threshold maximizes between-class variance under an implicit bimodality assumption of the image histogram, which often breaks down in mixed land covers and heterogeneous incidence angles typical of wide-swath SAR acquisitions. As a result, Otsu can over-map “water-like” low-backscatter surfaces (e.g., damp soils, shadows) unless constrained by morphology or topography[4]. To improve robustness and timeliness, fully automated Sentinel-1 chains fuse radiometric normalization, terrain correction, and context-aware thresholding/change-detection to deliver flood extents within tens of minutes after data publication. Evidence from an operational NRT system shows high reliability across diverse European floods, yet residual weaknesses persist in urban areas (layover/shadow), forests (volume scattering), and mountainous terrain without thorough radiometric terrain correction [5]. Time-series and anomaly-based detection further mitigate single-scene biases by comparing events against local baselines; these approaches report markedly improved stability across incidence angles and seasons, but they require dense historical archives and careful handling of multi-annual land-cover dynamics [6].

Learning-based methods span classical ensembles to modern deep networks. Random Forests (RF) have proven attractive for low-feature, small-sample regimes because they are resilient to speckle-induced outliers and capture non-linear class boundaries without distributional assumptions; however, they can be conservative (high precision, lower recall) when trained on limited feature sets (e.g., intensity only), and their transferability across sensors, polarizations, and incidence angles is not guaranteed without re-calibration. Comparative studies that benchmark ML/DL models on Sentinel-1 indicate that deeper architectures (e.g., U-Net variants and nested U-Nets) can achieve higher recall and better shape fidelity – particularly in complex urban scenes – provided domain shifts are addressed via curated training and aug-

mentation. Nonetheless, these gains come with higher data/compute demand and potentially reduced interpretability relative to RF [2], [7]. In parallel, coherence-based change metrics complement intensity thresholds by revealing inundation beneath vegetation and in built-up areas, but they require SLC data and are sensitive to temporal decorrelation from non-flood processes [8].

Optical indices (NDWI, MNDWI, AWEI) remain essential comparators in flood studies. NDWI (green – NIR) is sensitive to open water but overestimates in urban shadows and dark surfaces; MNDWI (green – SWIR) better suppresses built-up/vegetation confusion; AWEI targets shadow-related false positives. These indices, however, are fundamentally constrained by cloud cover and illumination, making them complementary rather than substitutive for SAR in time-critical response. Multi-sensor assessments over Europe conclude that operational flood detection benefits from the synergy of Sentinel-1 and Sentinel-2 but remains limited by revisit geometry and event duration; systematic detection is not guaranteed for short-lived floods or under persistent cloud [9], [10], [11].

Across all SAR methods, rigorous pre-processing is pivotal. Radiometric terrain correction (flattening  $\gamma^0$ ) reduces topography-induced backscatter biases; omission of this step can induce false “drying/wetting” patterns along slopes and in layover/shadow zones, which thresholding and ML alike may mistake for inundation. Even with terrain correction, residual incidence-angle trends, speckle, and textural ambiguity (e.g., roads, airstrips) persist, motivating multi-feature designs (texture, slope, distance-to-hydrography) and post-filters (morphology/CRF) [12].

In high-resolution constellations (e.g., X-band), these issues are amplified by finer-scale heterogeneity; consequently, model portability and calibration strategies are an active research area for operational adoption [13].

This study is centered on developing a machine learning model for water surface detection using two consecutive ICEYE sensor images acquired in VV (vertical-vertical) polarization on April 20 and 21, 2024. The area of interest covers the city of Uralsk and its surrounding territories (Figure 1), which, like many other regions of Kazakhstan, experienced severe and widespread flooding in spring 2024. The hydrologic regime of the study area is dominated by the Zhaiyk (Ural) River, which courses directly through Uralsk before draining to the Caspian Sea. During the April 2024 freshet the Zhaiyk (Ural) River stage in Uralsk peaked at  $\approx 8.64$  m (864 cm), overtopping the 8.50 m critical flood mark and triggering emergency reinforcement of levees together with the evacuation of tens of thousands of residents. Hydrometeorological briefings further projected that a flood-wave discharge of about  $1\,700\text{ m}^3\text{ s}^{-1}$  would propagate downstream, inundating extensive riparian zones and substantially enlarging the regional flood footprint [14]. According to official data, the flood resulted from a combination of precipitation events and preceding snowmelt. Notably, snowfall volumes during the winter of 2023–2024 significantly exceeded average values in northern regions of the country. Consequently, the land surface became considerably saturated, which, combined with recurring frosts, impeded natural drainage and contributed to elevated river and reservoir levels [14].



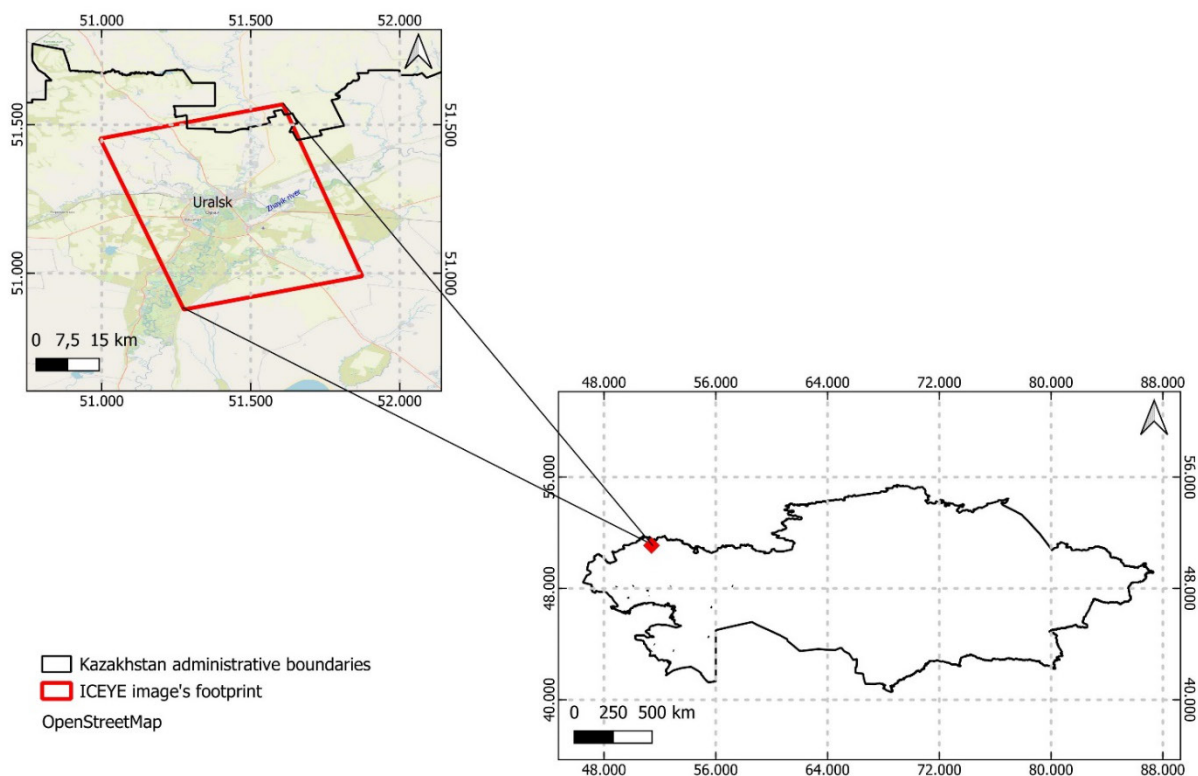


Figure 1. Region of interest

The initial stage of data handling is pre-processing, implemented within the ESA SNAP environment. The pre-processing workflow commences with radiometric calibration, converting digital numbers into  $\sigma_0$  values (backscatter coefficients). This step removes the dependency of scene brightness on viewing angle [15]. Following calibration, Range-Doppler terrain correction is conducted using precise ICEYE orbital data and the SRTMv3 digital elevation model integrated directly into ESA SNAP. This procedure projects the imagery onto the ground plane and mitigates overlay effects, foreshortening (occurring when features at different elevations, such as mountain bases and peaks, simultaneously reflect radar waves, resulting in distorted object shapes), and radar shadows (arising when a mountain's summit obstructs the radar signal from reaching the opposite slope). Hence, terrain correction significantly enhances the accuracy of subsequent classification tasks.

The final pre-processing step involves speckle filtering. It is important to note that speckle itself is not noise in the conventional sense; rather, it is a coherent interference pattern arising from the summation of radar waves reflected and scattered by numerous similar targets within a scene. Filtering thus involves local statistical averaging of scene brightness. In this study, a Boxcar filter with a  $7 \times 7$  window size was employed. This choice is justified by the fact that averaging over a sufficiently large window includes a greater number of scatterers, increasing the equivalent number of looks. Consequently, this approach preserves critical details, such as waterbody shorelines, and enhances the contrast between water surfaces and adjacent land areas [8].

Figure 2 illustrates the flowchart of the pre-processing steps using the imagery from April 20, with corresponding intermediate outputs.

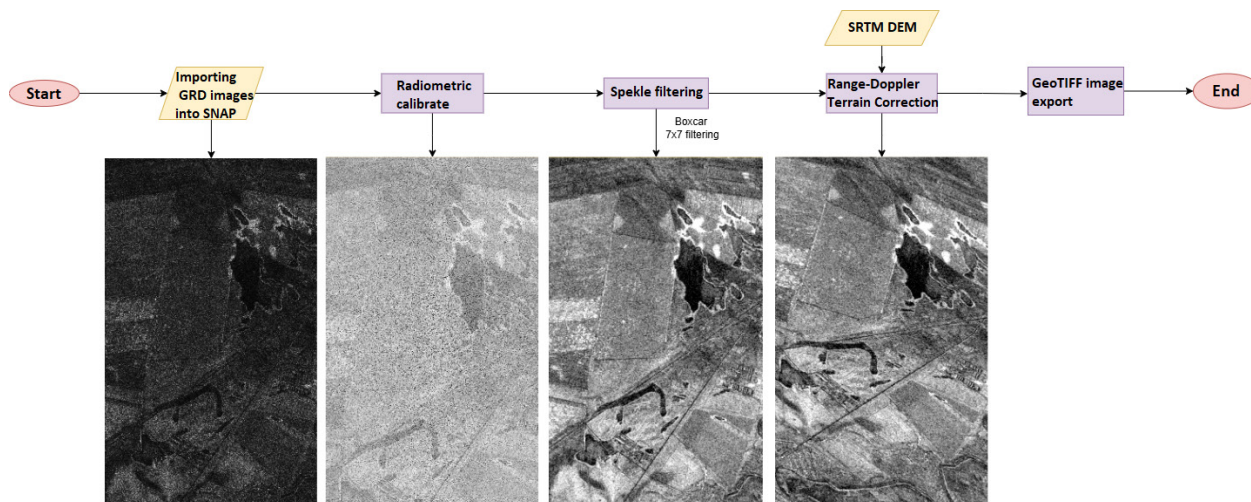


Figure 2. A block diagram of the preprocessing algorithm in the ESA SNAP software

At the conclusion of the pre-processing step, the resulting image is exported as a GeoTIFF file for subsequent integration and analysis within the Google Earth Engine (GEE) environment.

The water surface detection algorithm consists of several key stages. The first stage involves defining training classes and generating samples. The training dataset was manually created in GEE by delineating two distinct geometry sets – 'land' and 'water'. These polygons outlined regions empirically determined as water or land, based on visual inspection of Google Hybrid base imagery, the SAR image itself, and the NDWI map for the area, thus preventing false positives. The resulting dataset was subsequently used to train the Random Forest algorithm. Polygons are converted into a collection, and each pixel is assigned a binary class label (1 for water, 0 for land). To reduce computational load, radar intensity values are sampled at 10-meter intervals. From the resulting mask-intensity dataset, 8,000 random samples per class were selected – a quantity sufficient for accurate classification.

Figure 3 illustrates a flowchart depicting the subsequent classification and water surface detection process. The Random Forest algorithm was chosen for model development due to its demonstrated effectiveness with a limited number of features (in this case, just one – the radar backscatter intensity). Its advantage lies in using different decision thresholds at each node and random subsets of the data, thus producing diverse decision trees. Additionally, Random Forest makes no assumptions about data distribution, making even a single feature informative enough for building an accurate model. Another advantage is that each tree is trained on a random subset of the original data, significantly mitigating noise influence. Any noisy artifact present in a single tree will have minimal impact, as it is typically offset by the majority vote from other trees – provided the input data has been pre-processed properly to ensure a limited number of noisy artifacts.

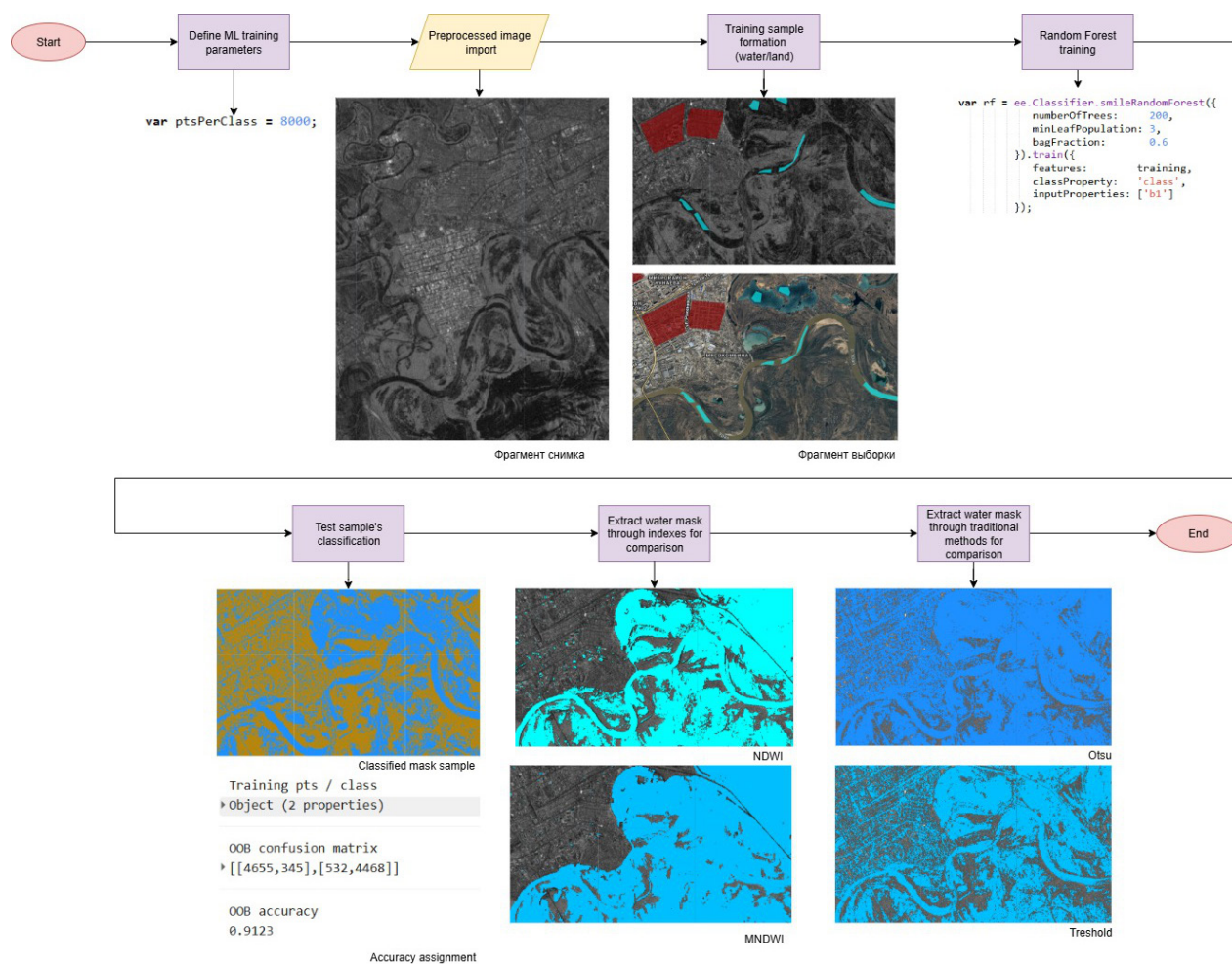


Figure 3. A block diagram of the water surface recognition algorithm operation on a preprocessed radar image

Comparative analysis showed that the classification accuracy (Overall Accuracy, OA) ranged from 75.9–76.8%, and the agreement coefficient  $k$  – from 0.52 to 0.54. The best results were achieved with the parameters RF\_100t\_0.5, RF\_300t\_0.6 and RF\_500t\_0.5, where OA was 76.7–76.8%, and  $k$  – 0.534–0.536. These values show a moderate but steady improvement compared to other configurations.

An increase in the number of trees from 100 to 500 did not result in a noticeable increase in accuracy: the maximum values were observed already with 100 trees (bagFraction = 0.5). With the further growth of the ensemble, the quality remained at a similar level, which indicates the saturation of the model. Varying the bagFraction parameter showed that the optimal value is 0.5–0.6, whereas at 0.7 there was a deterioration in quality (OA decreased to ~76.0%,  $k$  to ~0.52). This is because too high a sampling coefficient reduces the diversity of trees in the ensemble, which leads to overfitting and increased errors.

In general, the results indicate that the Random Forest model demonstrates resistance to changes in the number of trees, provided that the bagFraction parameter is selected correctly. RF\_300t\_0.6 (300 trees, minLeafPopulation = 3, bagFraction = 0.6) was adopted as the optimal set of parameters for further work, which provided the maximum value of  $k$  (0.535) and an overall accuracy of 76.8%, with an acceptable computational load.

So the training is performed using a Random Forest classifier comprising 300 decision trees, each built with a bagFraction of 0.6 (the fraction of data randomly selected for each tree) and



a minimum leaf population of 3. Testing different leaf population values did not affect model's accuracy, so the value 3 was chosen as the optimal. Each tree within the forest makes decisions through sequential binary splitting of the feature space. Every ensemble member is trained on a randomly selected subset of the initial dataset, employing random subsets of features. Since the single predictor is the radiometric intensity from band b1 of the radar image, the method relies on differences in the statistical distributions of signal amplitudes over water and land, resulting from distinct backscatter mechanisms. The training dataset is balanced, containing an equal number of samples for each class. Thus, the classifier learns to identify characteristic backscatter distributions: water surfaces, being smooth, tend to reflect radar waves away completely, yielding minimal returned signals; land areas, conversely, exhibit heterogeneous reflections dependent upon the specific type of surface. Examples illustrate the reflectivity of various surfaces are shown in Figure 4 [16].

C-band Microwave Energy Scattering on Ground/Water Surfaces and Linear Relationships

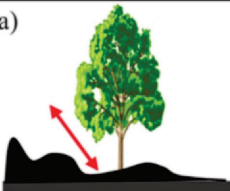
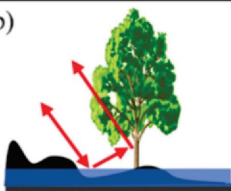
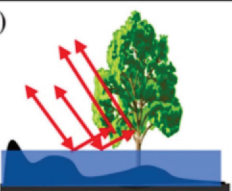
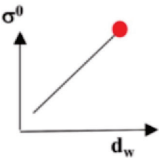
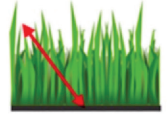
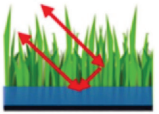
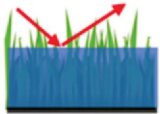
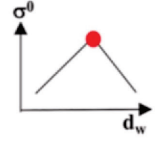
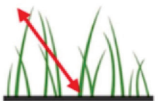
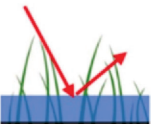
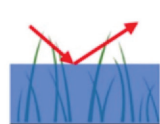
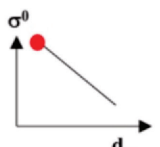
Vegetation	Unflooded	Shallow Water	Deep Water	$\sigma^0 - d_w$ Linear Relationship
Sparse Woody	(a)  Surface Scattering	(b)  Double-bounce and multiple-path scattering	(c)  Enhanced double-bounce and multiple-path scattering	(d) Lang and Kasischke, 2008 C-HH, C-VV 
Medium Dense Herbaceous	(e)  Surface Scattering	(f)  Double-bounce and multiple-path scattering	(g)  Specular Reflection	(h) Yuan et al., 2015 L-HH (Supposed relationship for C-band data) 
Sparse Herbaceous	(i)  Surface Scattering	(j)  Specular Reflection	(k)  Specular Reflection	(l) Kasischke et al., 2003, 2009 C-VV 

Figure 4. Backscatterer types on SAR images [17]

The developed model is applied directly to the original image without reducing its spatial resolution, thus preserving the quality of the output mask. In this scenario, the feature space consists of radar signal intensity values from band b1. Each decision tree selects an optimal threshold  $\theta$ , at which the division of the dataset into water and land classes maximizes information gain. The Gini impurity index (1) is employed as the criterion for evaluating information gain.

$$Gini(D) = 1 - \sum_{i=1}^C p_i^2, \quad (1)$$

where  $p_i$  – is the proportion of objects belonging to class  $i$  within the current node of the tree, and  $C$  is the total number of classes (water and land). During the classification stage, the intensity value  $x$  of an unknown pixel is input into each tree  $T_j$ . Each tree then returns to a class prediction  $c_j \in \{0, 1\}$ . Ultimately, the final prediction (2) is determined by majority voting across all trees:

$$\hat{c} = mode(T_1(x), T_2(x), \dots, T_N(x)), \quad (2)$$

where  $N$  is the total number of trees, which in this case is 200.

As a result, the output is a binary mask in which pixels corresponding to water surfaces have a value of 1, whereas pixels representing land have a value of 0. The second image in the analyzed series undergoes the same processing workflow, thereby yielding two masks corresponding to two consecutive days – April 20 and April 21, 2024.

## Results

Otsu's method was used as an unsupervised baseline to delineate open water directly from the SAR backscatter ( $b_1$ , dB). Let a histogram of the image within the analysis region be defined by bin centers  $x_i$  and counts  $n_i$  ( $i=1, \dots, L$ ) with  $N = \sum_i n_i$  and global mean  $\mu_T = \sum_i \frac{x_i n_i}{N}$ . For a candidate threshold at bin  $k$ , the cumulative class probabilities and means are:

$$\omega_0(k) = \frac{1}{N} \sum_{i=1}^k n_i, \quad (3)$$

$$\omega_1(k) = 1 - \omega_0(k), \quad (4)$$

$$\mu_0(k) = \frac{1}{N \omega_0(k)} \sum_{i=1}^k x_i n_i, \quad (5)$$

$$\mu_1(k) = \frac{\mu_T - \omega_0(k) \mu_0(k)}{\omega_1(k)}. \quad (6)$$

Otsu selects the threshold that maximizes the between-class variance:

$$\sigma_b^2 = \omega_0(k) \omega_1(k) (\mu_0(k) - \mu_1(k))^2, \quad (7)$$

equivalently minimizing the within-class variance. Applied to the histogram in Figure 5, the optimal threshold was  $t^* = -16.625621235853288$  dB (approximately  $-16.63$  dB). Pixels are then classified as water by a simple indicator function

$$W(x) = 1(b_1(x) < t^*), \quad (8)$$

which is theoretically consistent with specular returns from open water producing low backscatter in SAR. The resulting binary mask was restricted to the study geometry and converted to area by summing pixel areas, is the native pixel area.



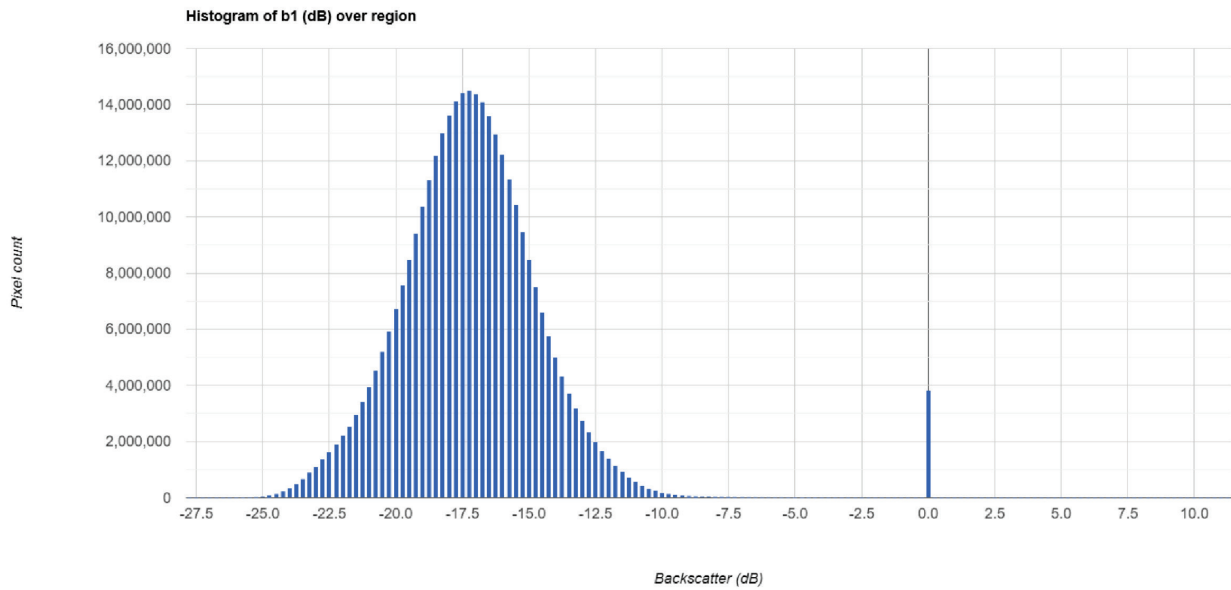


Figure 5. Histogram of the preprocessed ICEYE image

As a second baseline, water was delineated from the GRD backscatter (band b1, dB) using a single, empirically chosen threshold. Guided by the regional backscatter histogram and visual checks against reference water, the threshold was set to  $t = -18$  dB. Pixels were labeled water via a binary indicator:

$$W(x) = 1(b_1(x) < t) \quad (9)$$

which accords with the physical expectation that open water in C-band yields low, specular returns. The mask was computed over the analysis region and then clipped to the reporting geometry. Water area was obtained by summing pixel areas within the mask. This fixed-threshold estimate provides a training-free comparator to the Random Forest approaches, while remaining explicitly site- and acquisition-dependent.

The results of all three methods are displayed on Figure 6. There are large areas of false positive results for the Otsu method. Urban areas and roadways are classified as water, according to the method, which is not true. The threshold method worked much better, however, too many objects that are not water are classified as it.

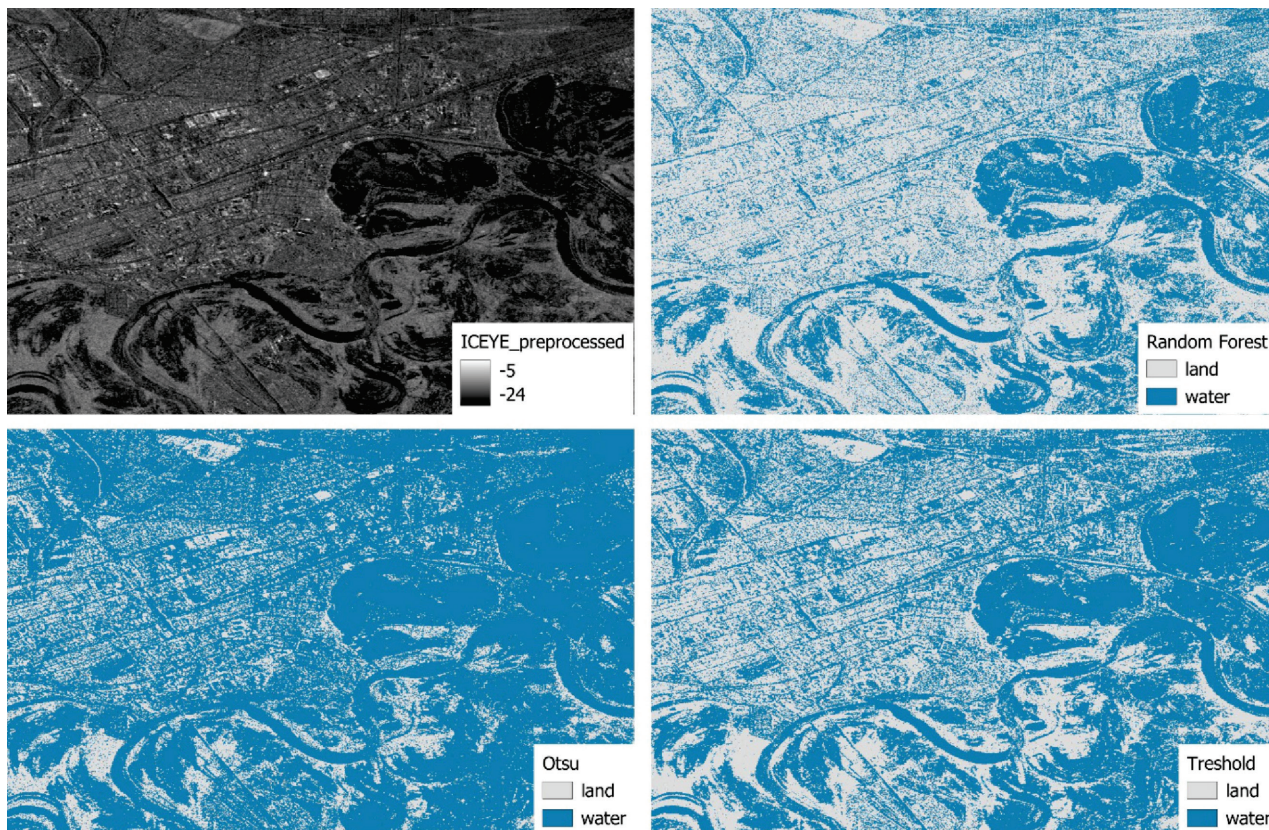


Figure 6. Water surface mask according to the ICEYE image (RF, Otsu, threshold methods)

Spectral indices such as the NDWI (Normalized Difference Water Index) (3), MNDWI (Modified NDWI) (10), among others, have been widely used in optical remote sensing tasks for water surface detection. Each of these indices possesses distinct advantages and limitations. The NDWI (3)[11] relies on differences in surface reflectance between the green and near-infrared (NIR) spectral ranges:

$$NDWI = \frac{GREEN - NIR}{GREEN + NIR} \quad (10)$$

The index demonstrates high sensitivity to open water bodies and can be effectively utilized for monitoring clear water surfaces. However, its application in urban areas, or in the presence of shadows, dark vegetation, or low-albedo objects, is often associated with a high likelihood of false positives [5].

The Modified NDWI (MNDWI) (11)[17] replaces the near-infrared (NIR) channel with the short-wave infrared (SWIR) channel, thereby enabling more effective suppression of reflectance from vegetation and artificial structures:

$$MNDWI = \frac{GREEN - SWIR}{GREEN + SWIR} \quad (11)$$

The use of MNDWI provides some advantages over the limitations inherent in NDWI; however, in scenarios involving weak water signatures—such as ice, geothermal waters, turbid waters, and similar conditions – the spectral response can become distorted, leading to increased omissions and false positives, analogous to those encountered with NDWI. The results of differential indexes outcome are displayed on Figure 7.



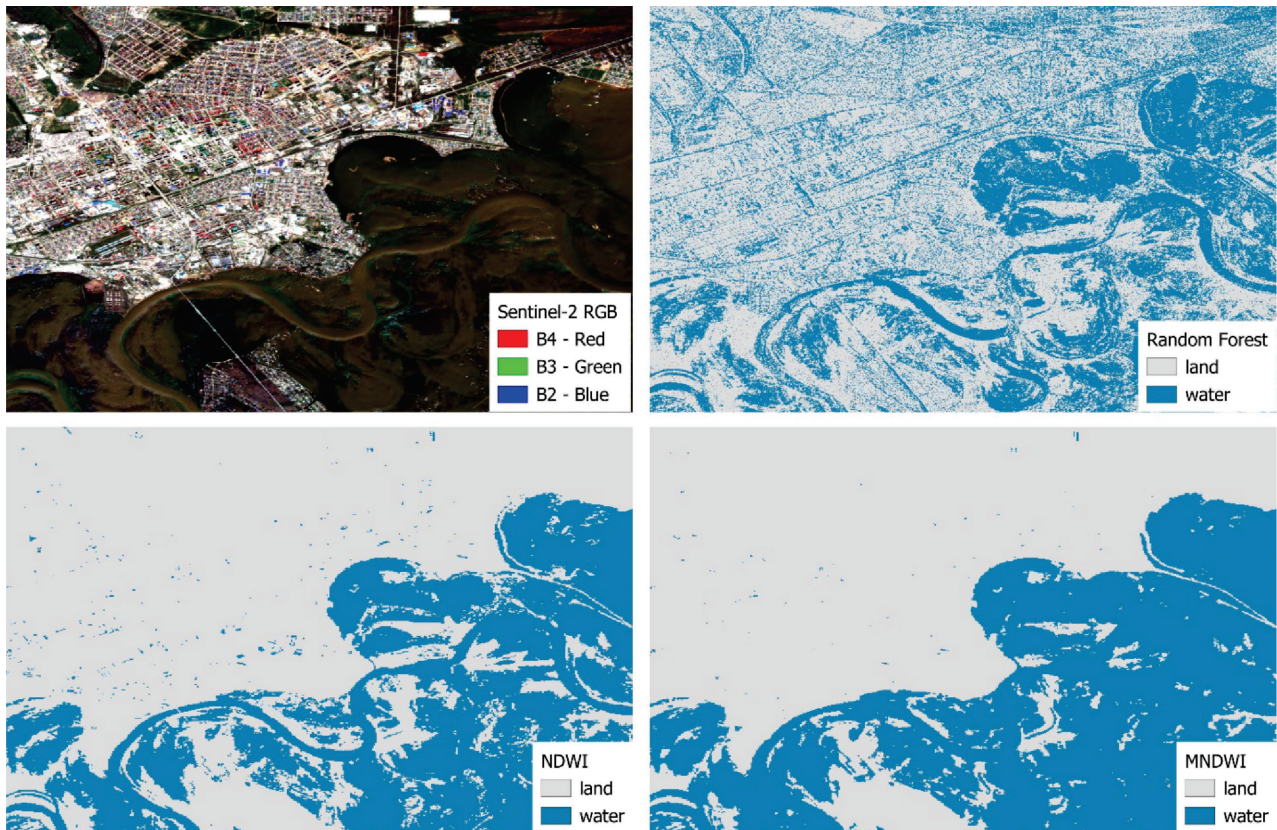


Figure 7. Water surface masks calculated from the Sentinel-2 image (NDWI, MNDWI)

A detailed examination also reveals significantly greater spatial detail in the RF-derived mask compared to the optical masks. This is primarily attributed to the considerably higher spatial resolution of the ICEYE X-band imagery (approximately 3 meters), in contrast to the Sentinel-2 imagery, which offers a resolution of 10 to 20 meters depending on the spectral band used [18].

As part of the comparison between the resulting classification and the reference optical water masks, the water surface area was calculated for each optical index-based mask as well as for the radar-derived mask.

As illustrated in the image, there are minor discrepancies in the estimated water surface area across the different methods. These differences can be attributed, first, to the spatial resolution disparities between the optical and radar imagery, as previously noted, and second, to the inherent characteristics of optical and radar sensing. Optical masks are susceptible to the influence of clouds and surface reflections, which can either inflate or deflate index values. In contrast, radar satellites are not affected by cloud cover but may misclassify objects with similar backscatter intensities – such as roadways, as mentioned earlier – as water surfaces.

Figure 8 presents the outcome of this analysis, showing the area values as output in the GEE console.

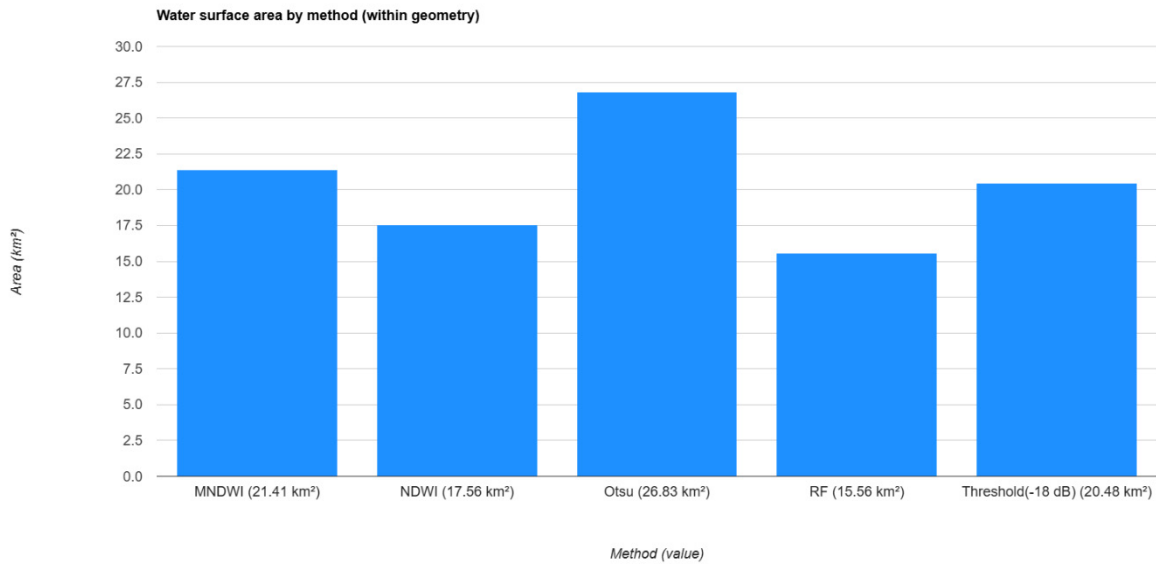


Figure 8. Water surface areas according to different extraction methods

Figures 9 and 10 present a comparison between the resulting masks and the original imagery. The algorithm demonstrates reasonably accurate performance: areas covered by water were identified with a satisfactory level of accuracy and reliability.

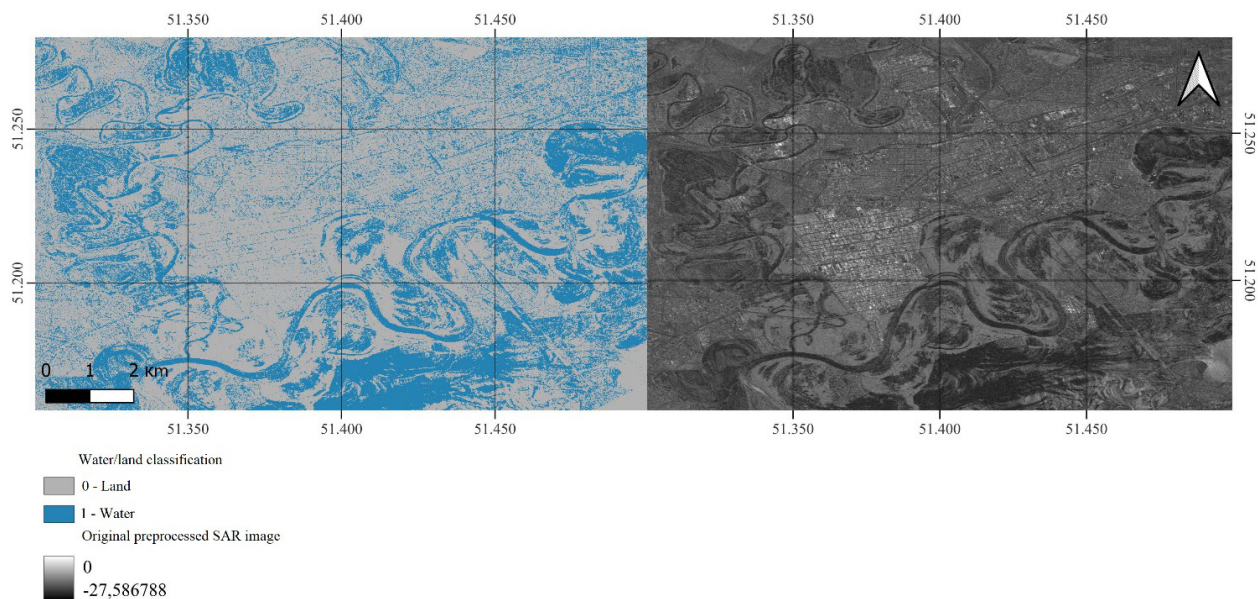


Figure 9. Water recognition on ICEYE image from April 20<sup>th</sup>



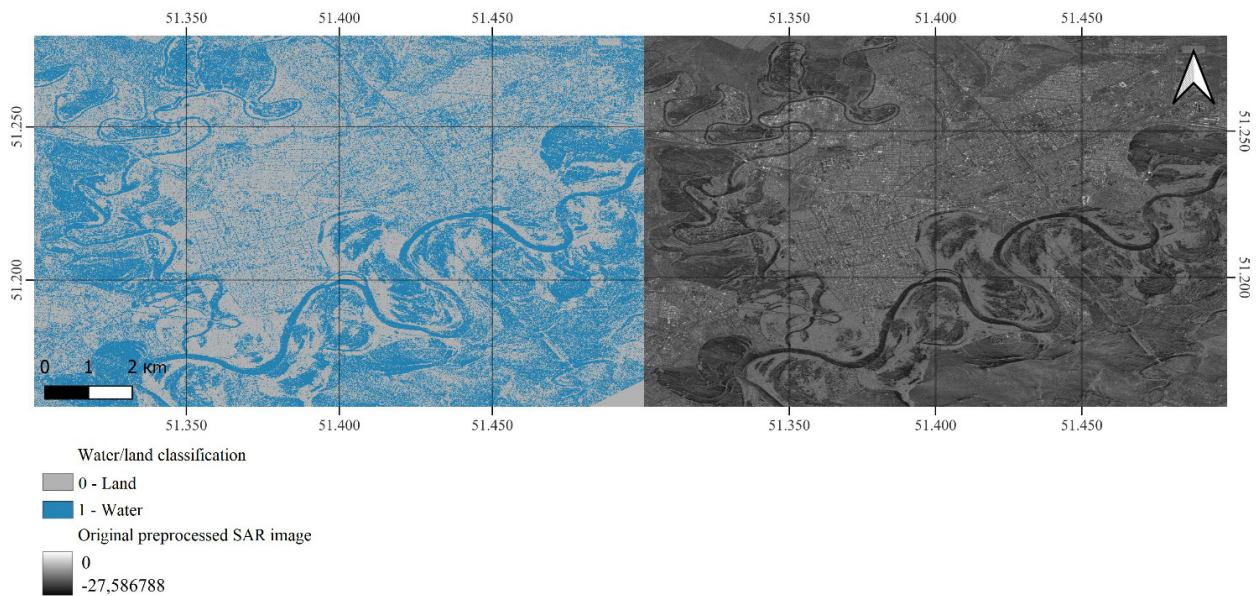


Figure 10. Water recognition on ICEYE image from April 21st

Figure 11 provides a more detailed view of the water surface detection results. The displayed images clearly show that major water bodies are accurately identified due to the high contrast in backscatter signals compared to non-water areas. However, the resulting mask contains some artifacts, such as segments of road networks. This type of misclassification occurs when the radar signal intensity values for hard surfaces (e.g., asphalt roads) are similar to those of specular reflections from calm water surfaces. One potential solution to this issue is to enhance the existing machine learning model by incorporating additional features such as texture parameters and polarization characteristics. This approach could reduce the incidence of false positives, particularly over road surfaces, by enabling the classifier to better distinguish between water and non-water targets with similar backscatter intensities.

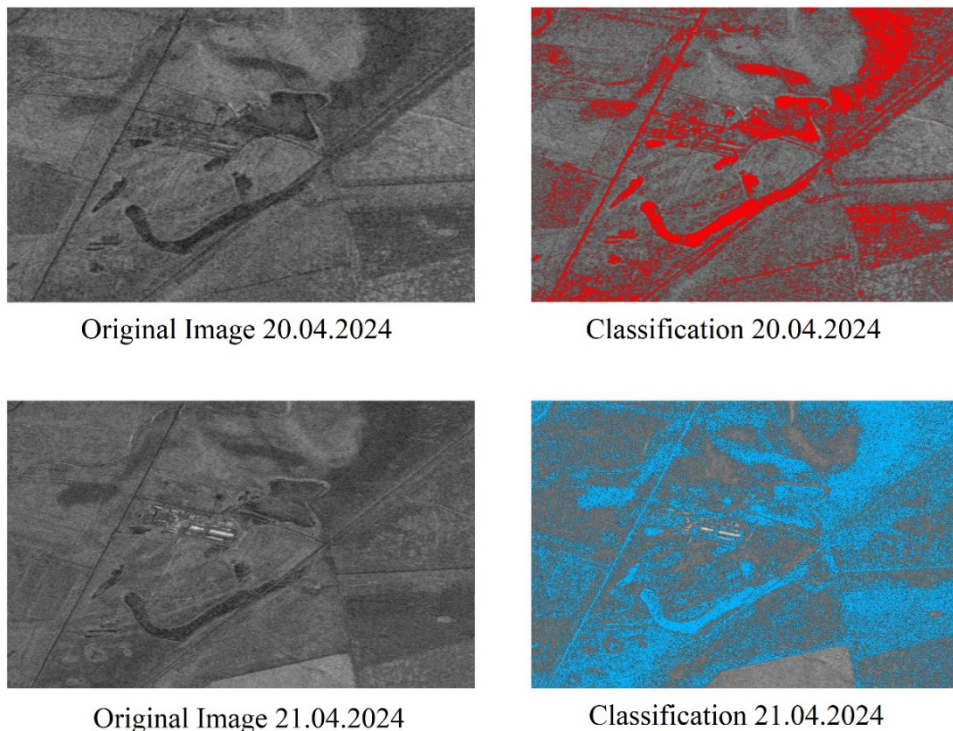


Figure 11. Closer look on water recognition



Another type of observed artifact is speckling within agricultural fields—individual pixels classified as water surrounded by pixels classified as land. This effect also arises due to the similar backscatter intensity of fields and water surfaces. During periods of active vegetation growth or when fields are highly saturated, including cases where they are flooded due to overflow, the distinction between surface types may become minimal. To address this issue, a post-processing procedure should be introduced following RF-based classification. A median filter applied over a moving window can be effective, smoothing isolated pixels while preserving the boundaries of larger objects. Alternatively, the existing model could be combined with operations specifically designed to eliminate false positives. This would result in a final output that is more reliable and directly usable for further applications.

### Discussion

During classifier training, an error assessment matrix was generated using out-of-bag (OOB) validation, which enables the calculation of standard performance metrics on independent data. The primary metric used was overall accuracy (12):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad (12)$$

where TP is the number of True Positive classified pixels, TN is the number of True Negative classified pixels, FP is the number of False Positive classified pixels, and FN is the number of False Negative classified pixels. Figure 12 shows the result of calculating the classification accuracy from the GEE console, according to which TP=6667 (water pixels correctly classified as water), TN=7609 (land pixels correctly classified as land), FP=391 (false positive recognition of land as water), FN=1333 (false negative recognition of water as land). Then the accuracy (13) is calculated as follows:

$$Accuracy = \frac{1601+1579}{1601+1579+521+442} = 0,7676, \quad (13)$$

which is considered sufficiently high value for machine learning tasks involving a limited number of classification features (Table 1).

Table 1 – Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	False Positives (FP = 521)	True Positives (TP = 1601)
Predicted Negative (0)	True Negatives (TN = 1579)	False Negatives (FN = 442)

It can also be concluded that Type I errors – 442 land pixels misclassified as water—constitute a relatively small proportion, indicating that the model rarely confuses land with water. However, water is more frequently misclassified as land. This issue is likely due not to deficiencies in the model itself, but rather to artifacts in the input imagery. The presence of wind and surface ripples on water bodies can directly affect their backscattering properties, leading to false negatives in classification.

To eliminate the dependence on the decision threshold, a Receiver Operating Characteristic (ROC) curve was constructed [19]. This curve illustrates how the two key error metrics of a binary classifier –  $TPR(\theta)$ ,  $FPR(\theta)$  (14, 15) change as the decision threshold  $\theta$  is gradually shifted across the range from 0 to 1.

$$TPR(\theta) = \frac{TP(\theta)}{TP(\theta)+FN(\theta)} \quad (14)$$

$$FPR(\theta) = \frac{FP(\theta)}{FP(\theta)+TN(\theta)}. \quad (15)$$

In this case, an algorithm was implemented to vary the probability of a pixel belonging to the "water" class from 0 to 1 in increments of 0.05. For each probability value, the sensitivity and false positive rate were calculated and plotted on a graph. For a given threshold value  $\theta$ , the sensitivity–false positive rate pair is computed as follows (16):

$$(FPR(\theta), TPR(\theta)) = \left( \frac{\sum_{i=1}^N 1(p_i \geq \theta, y_i = 1)}{\sum_{i=1}^N 1(y_i = 1)}, \frac{\sum_{i=1}^N 1(p_i \geq \theta, y_i = 0)}{\sum_{i=1}^N 1(y_i = 0)} \right) \quad (16)$$

According to Figure 12, the curve representing the relationship between sensitivity and the false positive rate lies well above the diagonal of random guessing, particularly in the region of low FPR values. Notably, when the false positive rate is around five percent, the sensitivity reaches approximately seventy-five percent.

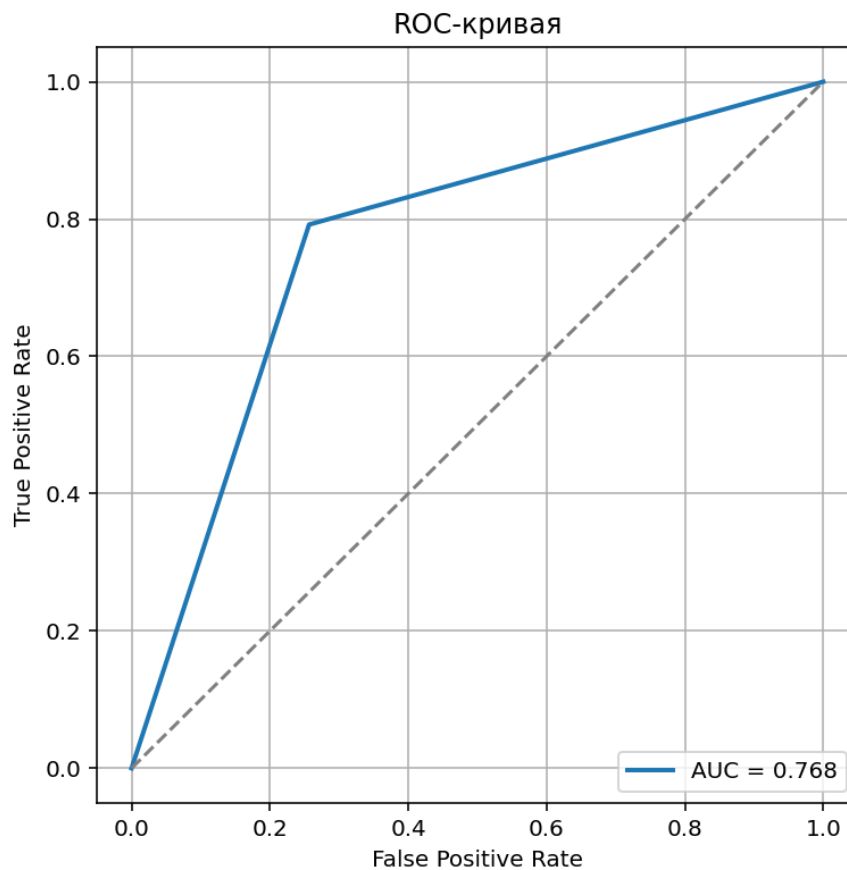


Figure 12. Effectiveness of the binary classifier model plot

The integral characteristic of the curve – namely, the area under the corresponding curvilinear trapezoid (17) – serves as a measure of the model's performance and is calculated as follows:

$$AUC = \int_0^1 TPR(x) dx \approx \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) \frac{TPR_{i+1} + TPR_i}{2} \quad (17)$$

Substituting the values obtained from formulas (9) and (10) yields a result of 0.91. This value indicates that there is a 91% probability that a randomly selected water pixel receives a

higher probabilistic score than a randomly selected land pixel, thereby demonstrating a high level of model performance. The upward trend of the curve observed in the region of high sensitivities, as shown in the figure, reflects a typical trade-off – further reduction in missed water pixels can only be achieved at the expense of an increased rate of false classification of land pixels. Thus, the final threshold selection should be aligned with specific operational requirements for the model.

The kappa coefficient ( $\kappa$ ) is a measure of agreement proposed by Cohen [20] and is widely used to evaluate classification quality while accounting for the probability of random agreement. Unlike Overall Accuracy, kappa shows to what extent the model's results exceed the level of random classification. The value of  $\kappa$  ranges from  $-1$  (worse than random) to  $1$  (perfect agreement), with  $0$  corresponding to pure chance.

Formally, the metric is defined as:

$$k = \frac{P_o - P_E}{1 - P_E}, \quad (18)$$

where  $P_o$  is the observed accuracy:

$$P_o = \frac{TP + TN}{N}, \quad (19)$$

and  $P_E$  is the expected accuracy under random assignment:

$$P_E = \frac{(row_1 \cdot col_1) + (row_2 \cdot col_2)}{N^2}, \quad (20)$$

where  $row_1$  and  $col_1$  are the row and column sums of the confusion matrix.

In this case  $P_o = 0.7676$ ,  $P_E = 0.5007$ , which yields to  $k \approx 0.535$ . This indicates a moderate level of agreement, exceeding random guessing by more than 26%.

Further reduction of omission errors for water detection would inevitably lead to a substantial increase in false inundation of land areas. Consequently, the choice of the decision threshold should be guided by the intended application: if a strict limit on false positives is required, it is advisable to fix the FPR; whereas in situations where the risk of missing water is more critical, the criterion of maximizing the difference between sensitivity and false positive rate (Youden's index) [19] offers a more balanced solution. Overall, the set of derived metrics confirms the algorithm's high reliability and its suitability for rapid flood mapping, even when using a limited set of radar-based features.

A spatial agreement analysis was performed to compare three SAR-based water-delineation approaches using backscatter (b1, dB): a supervised Random Forest (RF), an unsupervised Otsu threshold (data-driven;  $t^*$  dB), and a fixed manual threshold ( $-18$  dB). For each method pair, agreement maps within the study geometry partitioned pixels into four categories (both land, A-only water, B-only water, both water); per-method water extent was computed as "only + both," and pairwise consistency was summarized by the Jaccard index (intersection/union).

The results demonstrated a consistent ordering of mapped water extent (Figure 13): Otsu  $\gg$  Threshold ( $-18$  dB)  $>$  RF, yielding 26.83, 20.47, and 15.55 km<sup>2</sup>, respectively. Consistency was highest for Otsu vs Threshold (Jaccard  $\approx 0.76$ , intersection 20.48 km<sup>2</sup>), moderate for RF vs Threshold ( $\approx 0.71$ , intersection 15.00 km<sup>2</sup>), and lowest for RF vs Otsu ( $\approx 0.57$ , intersection 15.40 km<sup>2</sup>). The Otsu–Threshold comparison contained no Threshold-only water, indicating that the  $-18$  dB mask is a subset of Otsu, as expected from the less stringent Otsu threshold. Disagreements were dominated by Otsu-only areas relative to RF (11.43 km<sup>2</sup>) and relative to Threshold (6.35 km<sup>2</sup>), and by Threshold-only areas relative to RF (5.47 km<sup>2</sup>), suggesting that thresholding admits additional low-backscatter, ambiguous zones (e.g., roughened water, undated vegetation, wet soils, SAR shadows).

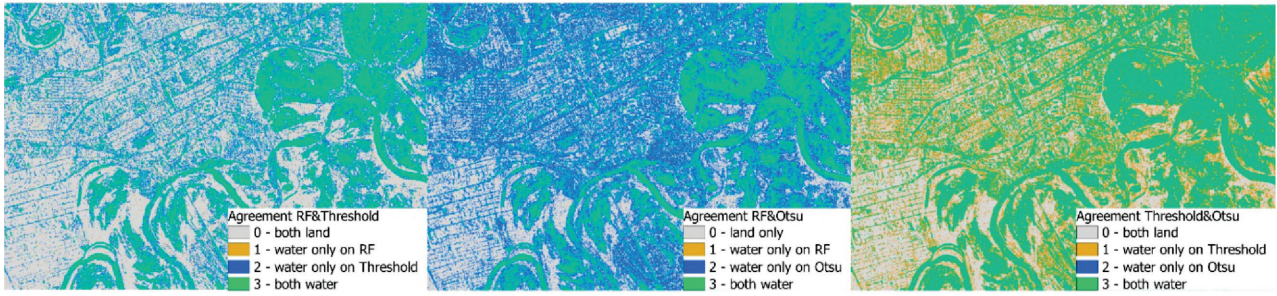


Figure 13. Agreement maps for SAR water extraction methods

These findings indicate that RF behaves conservatively, producing the smallest water extent and likely higher spatial purity by excluding borderline pixels that thresholding methods include. RF is therefore suitable when precision (low commission error) is prioritized over completeness, whereas Otsu (and, to a lesser degree, the  $-18$  dB threshold) emphasize recall by expanding the water mask. Definitive statements on effectiveness should be corroborated with independent reference data; within the agreement framework reported here, RF exhibits a selective delineation that trades some recall for reduced false positives.

The methodology for evaluating the binary water classification model from SAR data was carried out in two stages. At the first stage, the discriminatory capacity of the model was assessed independently of any fixed threshold using the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC). For this purpose, true class labels

$$y_{true} \in \{0, 1\} \quad (21)$$

and continuous model outputs

$$y_{score} \in [0, 1] \quad (22)$$

representing the probability of belonging to the positive class (water) were employed. The ROC curve was constructed by varying the threshold  $\tau \in [0, 1]$ .

The predicted class  $y_{pred}$  is assigned the value of 1 (water) if the probability of belonging to the water class, estimated by the model  $y_{score}$ , is greater than or equal to the selected threshold  $\tau$ . Otherwise, the predicted class is assigned the value of 0 (land).

For each  $\tau$ , the True Positive Rate was computed as (14) and the False Positive Rate as (15). The area under the ROC curve was then integrated in a standard manner. On the evaluated dataset, the resulting AUC was approximately 0.768, which indicates a robust discriminatory ability of the model and a clear deviation from random guessing (AUC = 0.5). This demonstrates that the classifier effectively orders examples by “water-likeness” even prior to thresholding.

At the second stage, threshold optimization was carried out to obtain a final binary map and threshold-dependent metrics. The optimization criterion was the maximum of the harmonic mean of precision and recall, namely the F1 score:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \quad (23)$$

where

$$Precision = \frac{TP}{TP + FP}, \quad (24)$$

and

$$Recall = \frac{TP}{TP+FN}. \quad (25)$$

The threshold was searched across a dense grid on the interval  $[0, 1]$   $[0,1]$ . Alternative criteria such as minimizing the Euclidean distance to the ideal ROC point (FPR=0, TPR=1) and maximizing Youden's index

$$J = TPR - FPR, \quad (26)$$

were also examined. The maximum F1 value on the evaluated sample was reached at a very low threshold, around  $\tau \approx 0.005$ . Such a localization of the optimum near zero reflects the conservative nature of the probability estimates produced by the model: the distribution of  $y_{score}$  is skewed toward small values, and even a minimal cutoff secures a good separation of the positive class. The stability of this solution is confirmed by the plateau of metrics observed in the range  $\tau \in [0.005, 0.05]$ , where further increases in threshold have virtually no effect on the balance of errors. This insensitivity of metrics to the precise cutoff demonstrates robustness in the neighborhood of the optimum.

The empirical performance of the model at the optimal threshold reveals the following profile. On the validation set containing 4143 observations, with a positive class prevalence of approximately 0.493, the confusion matrix

$$\begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix} = \begin{pmatrix} 1561 & 539 \\ 425 & 1618 \end{pmatrix} \quad (27)$$

was obtained. From this, the following values were calculated: precision  $\approx 0.750$ , recall  $\approx 0.792$ , F1  $\approx 0.770$ , and Intersection over Union (IoU)

$$IoU = \frac{TP}{TP+FP+FN} \approx 0.6327. \quad (28)$$

Overall accuracy reached  $\approx 0.767$ , while sensitivity (TPR) was 0.792 and specificity

$$TNR = \frac{Tn}{TN+FP} \approx 0.743 \quad (29)$$

Balanced accuracy, defined as  $(TPR + TNR)/2$ , was consistent with the observed overall accuracy, reflecting the near balance of class prevalence. These values jointly indicate that the classifier achieves a coherent trade-off between omission and commission errors at the chosen threshold.

From a methodological perspective, choosing the threshold by maximizing F1 is appropriate when false positives and false negatives have comparable cost, and the task requires balancing under-detection of flooded areas with false alarms over land. If domain-specific priorities differ, the thresholding strategy can be adapted accordingly: lowering the threshold increases recall at the expense of precision, while raising it favors precision at the cost of recall. The observation that the F1 optimum occurs at an extremely low cutoff further suggests imperfect calibration of probability outputs from the random forest. In practice, probability calibration (e.g., isotonic regression or Platt scaling) could be applied before thresholding, or alternatively, optimization could be based on the precision-recall curve in cases of class imbalance. Nevertheless, even in the current configuration, the model exhibits consistent performance: the AUC of 0.77 corresponds to a stable balance of precision and recall at a low threshold, with IoU exceeding 0.62, thus demonstrating the model's adequacy for practical water body delineation in the study area.



## Conclusion

The results of this study demonstrated the high effectiveness of the Random Forest algorithm for the automatic classification of water surfaces using ICEYE synthetic aperture radar (SAR) data. The proposed methodology, based on training the model with a manually labeled sample and utilizing a single feature—radiometric backscatter intensity—achieved a classification accuracy of 79.8%, as indicated by the out-of-bag (OOB) accuracy metric. The resulting water masks showed both visual and quantitative agreement with reference masks generated from optical indicators (NDWI, MNDWI), with the added advantage of significantly higher spatial detail due to the finer resolution of SAR imagery.

The analysis identified characteristic classification errors, including false positives over hard surfaces, which can be attributed to surface roughness and wind-induced ripples affecting the smoothness of water bodies. The proposed approach can be extended to other regions and applied in operational flood monitoring tasks, particularly under cloud-covered conditions when optical methods are ineffective. It may serve as a foundation for the development of regional or national flood information systems based on ICEYE and similar SAR platforms.

## Acknowledgment

This research is funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24992865 "Development of a multi-functional system of ground-space monitoring and early warning of natural and technogenic emergencies") for the period 2023–2025.

## References

- [1] Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., ... & Zhou, B. (2021). *Climate change 2021: The physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, 2(1), 2391.
- [2] Martinis, S., Kuenzer, C., Wendleder, A., Huth, J., Twele, A., Roth, A., & Dech, S. (2015). Comparing four operational SAR-based water and flood detection approaches. *International Journal of Remote Sensing*, 36(13), 3519–3543.
- [3] Chini, M., Pelich, R., Pulvirenti, L., Pierdicca, N., Hostache, R., & Matgen, P. (2019). Sentinel-1 InSAR coherence to detect floodwater in urban areas: Houston and Hurricane Harvey as a test case. *Remote Sensing*, 11(2), 107.
- [4] DeVries, B., Huang, C., Armston, J., Huang, W., Jones, J. W., & Lang, M. W. (2020). Rapid and robust monitoring of flood events using Sentinel-1 and Landsat data on Google Earth Engine. *Remote Sensing of Environment*, 240, 111664.
- [5] Feyisa, G. L., Meilby, H., Fensholt, R., & Proud, S. R. (2014). Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. *Remote Sensing of Environment*, 140, 23–35.
- [6] Ghosh, B., Garg, S., Motagh, M., & Martinis, S. (2024). Automatic flood detection from Sentinel-1 data using a nested U-Net model and a NASA benchmark dataset. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 92, 1–18.
- [7] McFeeters, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7), 1425–1432.
- [8] Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66.
- [9] Toma, A., Şandric, I., & Mihai, B.-A. (2024). Flooded area detection and mapping from Sentinel-1 imagery: Complementary approaches and comparative performance evaluation. *European Journal of Remote Sensing*, 57(1), Article 2414004.
- [10] Twele, A., Cao, W., Plank, S., & Martinis, S. (2016). Sentinel-1-based flood mapping: A fully automated processing chain. *International Journal of Remote Sensing*, 37(13), 2990–3004.

- [11] Xu, H. (2006). Modification of normalized difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14), 3025–3033.
- [12] Malakhov, D. V., Dolbnya, O. V., & Kurbanova, R. A. (2025). Flooding of 2024 in Turgay-Irgyz interflow: The impact on the biodiversity assessed by satellite data. *Geografija i vodnye resursy*, (1), 22–30.
- [13] Small, D. (2011). Flattening gamma: Radiometric terrain correction for SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 49(8), 3081–3093.
- [14] Cardona-Mesa, A. A., Vásquez-Salazar, R. D., Travieso-González, C. M., & Gómez, L. (2025). Comparative analysis of despeckling filters based on generative artificial intelligence trained with actual synthetic aperture radar imagery. *Remote Sensing*, 17(5), 828.
- [15] Zhang, B., Wdowinski, S., Gann, D., Hong, S. H., & Sah, J. (2022). Spatiotemporal variations of wetland backscatter: The role of water depth and vegetation characteristics in Sentinel-1 dual-polarization SAR observations. *Remote Sensing of Environment*, 270, 112864.
- [16] Welbeck, R. N. A. (2021). Assessing the effects of sea level rise on urban floods: A 1D2D satellite-based flood inundation modelling approach in Accra coastal zone (Master's thesis, University of Twente).
- [17] Ignatenko, V., Dogan, O., Radius, A., Nottingham, M., Muff, D., Lamentowski, L., ... & Vilja, P. (2024, April). ICEYE Microsatellite SAR Constellation: SAR data quality improvements and new Dwell imaging mode. In *EUSAR 2024; 15th European Conference on Synthetic Aperture Radar* (pp. 1118–1192). VDE.
- [18] Janssens, A. C. J., & Martens, F. K. (2020). Reflection on modern methods: Revisiting the area under the ROC curve. *International Journal of Epidemiology*, 49(4), 1397–1403.
- [19] Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1), 6086.
- [20] Hanegraaf, P., Wondimu, A., Mosselman, J. J., De Jong, R., Abogunrin, S., Queiros, L., ... & Van Der Schans, J. (2024). Inter-reviewer reliability of human literature reviewing and implications for the introduction of machine-assisted systematic reviews: a mixed-methods review. *BMJ Open*, 14(3), e076912.