A 10 m resolution global lake database based on Sentinel-2 MSI data and deep learning

Beihui Hu1, and Lian Feng1

1 School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, China

Corresponding author: L. Feng, ([fengl@sustech.edu.cn)](mailto:email@address.edu))

Key Points:

* We provide a global lake database with ~12 million lakes > 0.005 km2
* Small lakes with size <1 km2 dominate the global lake count (account for 98%)
* The widely distributed ponds in Southeast Asia have been accurately and comprehensively mapped.

Abstract

It is critical to understanding the abundance and distribution of global lakes, but existing global lake databases lack comprehensive coverage of small lakes, thus hardly depict global lake distribution accurately. Here, we developed a 10 m resolution global lake database using Sentinel-2 MSI data from March 28, 2017, to April 10, 2022 and a deep learning method. Our GLAKESplus database covers ~12 million lakes with areas larger than 0.005 km2, amounting to a total area of ~3.4×106 km2, providing detailed information on their average boundaries and spatial distribution. Our results indicate that small lakes (< 1 km2) account for 98.3% of the total number of lakes worldwide, further highlighting the importance of small lakes in the global ecosystem. Compared with existing databases, GLAKESplus offers higher spatial resolution and more comprehensive coverage of small lakes, providing a more accurate and reliable data foundation for related research.

1 Introduction

Lakes and reservoirs, hereafter simply “lakes”, play an important role in global hydrological, biogeochemical and carbon cycles (Lehner & Döll, 2004; Verpoorter et al., 2014). Their functions are closely related to their geometric characteristics (Messager et al., 2016). Due to their sensitivity to climate changes and human activities, global lakes are constantly changing (Pi et al., 2022; Williamson et al., 2009). Understanding the spatial distribution and variability of global lakes is crucial to the study of related earth system process and water resource regulation.

Mapping global lakes remains challenging because of their vast distribution, diverse morphology and sheer quantity. GLWD (Lehner & Döll, 2004) and HydroLAKES (Messager et al., 2016) are two global lake databases compiled from multiple data sources with varying resolutions and mapping times, making them difficult to accurately depict global lake distributions. Remote sensing, with its wide coverage, timeliness and rich information, has been widely used in large-scale lake observation. Coarse-resolution satellite sensors such as AVHRR (spatial resolution of 1001 m) and MODIS Terra/Aqua (spatial resolution of 0.25 km to 1 km) are only suitable for studying large lakes.

Since Landsat data became freely available in 2008, medium-resolution Landsat satellites have been widely used in global lake mapping. For example, Verpoorter et al. (2014) processed Landsat7 ETM+ images from around 2000 to create the GLOWABO, mapping global lakes larger than 0.002 km2. Similarly, Sheng et al. (2016) used Landsat8 OLI data from around 2015 to create the Circa-2015 database, providing representative water areas for global lakes larger than 0.004 km2. However, due to the diverse forms of surface water bodies, GLOWABO and Circa-2015 may contain misclassification and omission errors and their global mapping accuracy has not been assessed.

Pekel et al. (2016) used an expert system to classify water in each Landsat image, producing the Global Surface Water Occurrence (GSWO) product, which shows the probability of water presence. Building on GSWO, Pi et al. (2022) used deep learning to capture global lake changes from 1984 to 2019, revealing the importance of small lakes. The resulting GLAKES database covers ~3.4 million lakes larger than 0.03 km2, providing maximum lake boundaries and time-series weighted lake areas over the study period.

Despite the minimum mapping unit of 0.002 km2 in existing Landsat-based global lake databases, small lakes remain poorly understood. Although their contribution to global lake area is relatively small, small lakes dominate the global lake count. Studies have shown that small lakes are significant source of inland water carbon flux and exhibit more dramatic changes in their size (Holgerson & Raymond, 2016; Pi et al., 2022). Therefore, there is a need to generate a global lake database with improved small-lake coverage. The increasing availability of high-resolution satellites makes this goal achievable. Compared to costly commercial satellites (e.g., SPOT, QuickBird), Sentinel-2 provides free data with a spatial resolution up to 10 meters, offering significant advantages for lake observation. While previous studies have explored regional-scale applications, such as Yang et al. ( 2020), who used Sentinel-2 satellite data to estimate monthly surface water extent in France, achieving global 10 m resolution lake mapping remain a major challenge due to the intensive computational and storage demands.

Recent advances in cloud computing platforms have improved the efficiency and accessibility of processing massive remote sensing data. The Google Earth Engine (GEE) platform provides Sentinel-2 L2A products since March 28, 2017, enabling global satellite image processing with reduced local storage requirement. In addition, deep learning techniques with its high accuracy, speed, and automation, offers significant advantages for large-scale remote sensing observations. For example, Brandt et al. (2020) used sub-meter resolution satellite imagery and deep learning to count over 1.8 billion trees with crown sizes larger than 3 m2 in the West African Sahara and Sahel region. GLAKES were generated by deep learning as well. Deep learning models can learn both spectral and geometric features of ground features, providing a significant advantage traditional lake identification methods.

Motivated by previous research, this study aims to develop a 10 m resolution global lake database using Sentinel-2 remote sensing data and a deep learning method. The specific research objectives include: 1) training a semantic segmentation model suitable for Sentinel-2 global lake mapping; 2) mapping global lakes to create a global lake database with improved small-lake coverage and more precise boundaries; and 3) analyzing the global lake distribution and comparing it with existing databases.

2 Materials and methods

The production process of GLAKESplus was as follows: 1) Image preprocessing, where the pixel-wise average of NDWI (McFEETERS, 1996), red, green, blue, and near-infrared bands of Sentinel-2 data from the study period were calculated and downloaded in slices to local storage. 2) Sample preparation, where worldwide samples were generated to form a training dataset (divided into training, validation and test sets. 3) Model application, where a U-Net model was trained to extract lake features from sentinel-2 images and predict a raw global lake classification map. 4) Post-processing, where several post-processing steps were applied to reduce commission and omission errors in the raw global lake classification map, ultimately producing the GLAKESplus database. The flowchart is illustrated inFigure 1.

Figure 1. Flowchart for developing the GLAKESplus database.

2.1 Sentinel-2 data and preprocessing

Sentinel-2 is an earth observation mission under the Copernicus program of the European Space Agency (ESA), consisting of Sentinel-2A (launched in 2015) and Sentinel-2B (launched in 2017), with a revisit period of 2 to 5 days. The B2 (blue), B3 (green), B4 (red), and B8 (near-infrared) bands of sentinel-2 have a spatial resolution of 10 m, enabling the delineation of finer lake boundaries. The near-infrared bands (B11, B12) have a spatial resolution of 20 m, while the remaining bands have a spatial resolution of 60 m. We selected three visible bands (B2, B3, B4), one short-wave infrared band (B11, resampled to 10 m) and the Normalized Difference Water Index (NDWI) (McFEETERS, 1996) for lake mapping. NDWI is a commonly used water enhancement index, calculated for Sentinel-2 as:

NDWI effectively suppresses vegetation signals and enhances water features. However, it has limitations in distinguishing water bodies from impervious surface. Incorporating SWIR band information helps mitigate this issue, improving the accuracy of water body classification.

Using the GEE platform, we obtained Sentinel-2 L2A images with label “percentages of cloud pixel” <60% from March 28, 2017 to April 10, 2022 and performed several preprocessing operations.

First, the Sentinel-2 cloud probability product (S2Cloudless) was used to remove cloud pixels. Based on empirical thresholds, all pixels with a cloud probability greater than 50% were masked, and cloud shadow pixels were filtered using NIR band dark pixels and cloud projection intersections. Additionally, snow and ice pixels were masked using the Scene Classification Layer (SCL) of Sentinel-2 L2A products. Subsequently, we calculated the NDWI index for each image and averaged all images pixel-wise to obtain a mean composite image that minimizes seasonal variations and removes transient disturbances (e.g., cloud residuals, algal blooms, sediment plumes). Finally, nearly 10 TB of Sentinel-2 composite images were downloaded in tiles to the local computer for subsequent lake extraction.

2.2 Sample preparation

We selected representative sample regions globally and manually labeled lake boundaries for model training. For each sample region, we initially generated lake labels using a threshold-based segmentation method on the NDWI band, with thresholds determined manually. Extensive manual refinement was then performed to remove commission and omission errors ensuring high-quality final labels. In Sentinel-2 images, lakes typically have higher NDWI values than the background, lower reflectance in the RGB and SWIR bands, and a round, flat shape, making them easy to distinguish from the background. Regions that were easy to distinguish were labeled as (1) Normal Regions (NR). Additionally, we observe several regions requiring careful identification: (2) Alongside Rivers Regions (AR), Surface water bodies exhibit diverse morphologies, requiring careful differentiation between rivers and lakes, especially oxbow lakes, which share similar shapes with river channels; (3) Built-up Regions (BR), where buildings and their shadows may be misclassified as lakes in the NDWI band and need to be removed using other bands; and (4) Ice Lake Regions (IL), where the glaciers have similar features to lakes in the NDWI and SWIR bands are similar to lakes, requiring careful delineation of lake boundaries; (5) Salt Lakes Regions (SL), salt lakes have lower NDWI thresholds and exhibit high reflectance in other bands; Finally, A total of 799 labeled sample regions were created and stratified into training (60%), validation (20%) and test (20%) sets following the stratified random sampling method. The spatial and size distribution of the sample regions is shown in Figure 2. To be notice, the sample type of each region are defined by their dominant feature.

Figure 2. Spatial distribution and statistical characteristics of sample regions. (a) Global spatial distribution of sample regions, categorized into five region types: NR, AR, BR, IL and SL. Different colors represent the dataset splits: training (sky blue), validation (orange), and test (green). (b) Histogram of sample region sizes, divided into 27 logarithmically spaced bins, with different dataset splits stacked for visualization. (c) Summary statistics of sample regions, including the count of regions and their total area for each region type.

2.3 Model application

deep learning refers to the automatic learning and extraction of complex features from input data through multi-layer neural networks, has become a key driver and process of artificial intelligence over the past decades (Brandt et al., 2020).Among the powerful models in deep learning, U-Net (Ronneberger et al., 2015) and its variants have achieved state-of-the-art performance in semantic segmentation task. U-Net is a fully convolutional neural network, the left side of the network functions as an encoder, extracting hierarchical features, while the right side acts as a decoder, reconstructing spatial information through up-sampling. During the encoding and decoding process, U-Net fuses deep and shallow features via skip connections, thereby improving the accuracy of semantic segmentation. By employing the overlap-tile strategy, U-Net can seamlessly segment arbitrarily large images (Ronneberger et al., 2015). Building on previous research (Brandt et al., 2020), Pi et al. (2022) applied the U-Net model to global lake mapping, achieving promising results. Therefore, we made subtle adjustments to their code and adapted the U-Net model for global lake mapping at a 10 m resolution using Sentinel-2 imagery.

Due to GPU memory constraints, the model input size was set to 576×576. Since the sample regions were too large for the model input, a random sampling method was employed to generate patches of the same size as the model input from the training and test sets. The probability of each sample region being randomly selected was proportional to its size to avoid undersampling large regions and oversampling small regions.

The training set and validation set were used to model training. We adopted the same loss function and optimization algorithms as Pi et al. (2022). During training, the gradient of loss function were calculated to optimize the model's parameters, making predictions as close as possible to the true labels. We keep the model with the smallest loss error on Validation set were used to save the best model. In our study, we keep the model with the smallest loss error on Validation set. Training was terminated prematurely when the validation loss did not decrease for 50 consecutive epochs to avoid overfitting of the training data. The specific hyperparameter settings for training are shown in Supplementary Table 1.

After training, the final model were used to predicte lakes from global grid images. A sliding window was used to crop large grid images into small patches. The prediction of each patch was stitched together to form raw global lake classification map. To improve accuracy, the prediction of edge pixels within a 100-pixel margin of each patch were discarded due to the insufficient contextual information at the edges.

2.4 Post processing

Several post-processing operations were applied to the raw global lake classification map. Due to the relatively small input size, the deep learning model struggled to distinguish local features similar to large lakes, such as oceans and river with large widths. Thus, we removed oceanic and river residuals by other databases. The coastline data sets of Openstreetmap (OSM) (Goodchild, 2007) were used to remove ocean residuals. Its land polygons can be downloaded from <https://osmdata.openstreetmap.de/data/land-polygons.html> . All polygons that were not within land polygons were considered as ocean residuals and been removed. Lagoons connected to the ocean were not included in our consideration because their different characteristics from inland lakes. To remove river residuals, we adopted and modified the method proposed by Pi et al.(2022). Firstly, we use OSM river data and the Global River Widths from Landsat (GRWL) database to remove river pixels from the raw global lake classification map. Subsequently, we retained reservoir polygons that intersected with other reservoir datasets,including the GeoDAR dataset (Wang et al., 2022) and OSM reservoir data. The river and reservoir data from OSM were extracted from OSM's global dataset (<https://planet.openstreetmap.org/>). Next, we improved the accuracy of river masking using the GLAKES dataset and the Area Ratio (AR) of each polygon. The AR is calculated using the following formula:

Where Areabefore masking and Areaafter masking represent the area of the polygon before and after river masking, respectively. Polygons with AR closer to 1 are more likely to be a lake connected to a river. Conversely, polygons with AR closer to 0 are more likely to be a river residuals. Specifically, polygons intersecting with GLAKES with AR > 0.8 were retained as river-connected lakes(Fig. 3a). Polygons not intersecting GLAKES with AR < 0.8 were entirely removed, as they are essentially river residuals (Fig. 3d). In other cases, polygons after initial mask were retained (Fig. 3b, c). Finally, extensive manual inspection were performed to minimize errors.

Figure 3. Post-processing of river mask and the corresponding results.(a) Target polygons intersect with GLAKES with an area ratio ≥0.8; (b) Targets polygon not intersect with GLAKES with an area ratio ≥0.8; (c) Targets polygon intersect with GLAKES with an area ratio <0.8; (d) Targets polygon not intersect with GLAKES with an area ratio <0.8.

Based on the accuracy assessments results (Figure 5), we set the minimum lake area threshold as 0.005 km2 and removed all polygons smaller than this threshold; After the above operation, the commission errors of raw global lake classification map were largely eliminated, we further performed a lake-completion operation to reduce the commission error. Due to the limited input size, parts of large lakes with features similar to rivers were missed by the model. Additionally, in the Sentinel-2 mean basemap, lakes in arid regions were often misclassified as land due to their high surface reflectance. GLAKES has been proven accurate for mapping large global lakes, providing comprehensive maximum lake boundaries over multiple years, making it suitable for large lake completion. PLD (Wang et al., 2023) is a global lake dataset integrating multiple data sources, with a minimum lake coverage area of 0.01 km². Its primary data source, Circa-2015, has shown good performance of highly dynamic lakes in Oceania (Sheng et al., 2016). Therefore, the PLD dataset was used to supplement lakes in arid regions. We defined arid regions as areas with an Arid Index < 0.2, obtained form the Arid Index database (Zomer et al., 2022).

Before merging, we processed the GLAKES dataset to remove misclassification errors caused by incomplete floodplain definitions. SHIFT (Zheng et al., 2024) is a global geomorphic floodplain dataset based on DEM-mapping with a 90 m resolution, offering more comprehensive coverage than existing floodplain data. We applied a 30% occurrence threshold to mask GLAKES within the SHIFT dataset and outside arid regions, effectively removed the floodplain residuals (Fig. 4a, b) and agriculture zones (Fig. 4c, d).

Figure 4. Comparison of GLAKES before and after applying the flood mask. (a) and (b) demonstrate that the flood masking operation effectively removed misclassified floodplains, while (c) and (d) illustrate its effectiveness in eliminating misclassified paddy fields. For (a-d), the left figures show the water occurrence (ranging from 0 to 100) and GLAKES before mask (orange line), the right figures show the GLAKES polygon after mask (blue).

Finally, natural lakes larger than 1 km² and all reservoirs from the floodplain-masked GLAKES dataset, along with all PLD lakes located in arid regions, were merged into the raw global lake classification map. After extensive manual inspection and modification, the final global lake classification map was compiled GLAKESplus dataset.

3 Results

3.1 Accuracy assessments

An independent test set consisting of 158 sample regions was used to evaluate the model classification accuracy. The assessment metrics included Recall, Precision, F1 score and IoU. Recall represents the proportion of correctly predicted lakes among lake labels, while precision represents the proportion of correctly predicted lakes among predicted lakes; F1 score is the harmonic average of Recall and Precision and IoU represents the proportion of correctly predicted lakes to the union between lake labels and prediction. First, we evaluated model performance at the patch level. A total of 2424 patches (each 576×576 pixels) were extracted sequentially from the test set for type-wise accuracy assessment (Table 1). Overall, our model showed good performance with IoU of 87.5% and other metrics exceeding 92.7%. Performance varied slightly across different region types. In AR regions, precision and IoU were relatively lower (89.77% and 84.28%, respectively; Table 1), primarily due to the river residues of large rivers (Supplementary Figure 1c). In SL regions, our model exhibited lower recall and IoU (72.71% and 69.80%, respectively; Table 1), primarily due to the omission of salt lakes (Supplementary Figure 1f). Our model performed well in other regions (NR, BR, and IL).

Table 1 Accuracy assessments of our developed deep-learning algorithm at patch level with different region type.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Region type | Patch count | recall (%) | precision (%) | F1 (%) | IoU (%) |
| NR | 798 | 93.00 | 96.26 | 94.29 | 89.71 |
| AR | 698 | 93.08 | 89.77 | 91.08 | 84.28 |
| BR | 308 | 95.78 | 90.83 | 93.16 | 87.29 |
| IL | 522 | 96.79 | 94.76 | 95.72 | 91.87 |
| SL | 98 | 72.71 | 95.60 | 81.46 | 69.80 |
| total | 2424 | 93.37 | 93.35 | 92.70 | 87.50 |

After merging the prediction results from the test set, we conducted an overall accuracy evaluation of the river masking operation (Table 2) and performed a lake-scale accuracy assessment (Figure 5). The results indicated that the river mask operation improved Precision by 1.77% while slightly reducing Recall by 0.1% (Table 2), the true predicted area for the removed river residuals approached zero (Fig. 5a).

Table 2 The improvement of river mask operate in test regions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | recall (%) | precision (%) | F1 (%) | IoU (%) |
| Before river mask | 93.50 | 92.95 | 93.22 | 87.31 |
| After river mask | 93.40 | 94.72 | 94.05 | 88.78 |

At the lake scale, the true predicted area showed strong correlations with both corresponding labeled area and the predicted area, with R2 values of 0.98 and slopes greater than 0.92 (Figure 5a). However, the correlation between the correctly predicted area and the labeled area was slightly weaker for smaller lakes (Fig. 5a). For lakes larger than 0.005 km2, both the mean recall of the labeled lakes and the mean precision of the predicted lakes exceed 86%, and improve as lake size increases. Notably, in 1–20 km2 size group, the mean recall is 93.1%, slightly lower than the 95.5% observed in the [0.1–1 km²] size group.

Figure 5. Validation in lake-entity of the deep-learning algorithm. (a) The Area of each label and each predicted polygon and its corresponding true predicted area (b) The mean recall of labels and mean precision of predicted polygons in different lake size groups.

3.2 Global lake abundance and distribution

GLAKESplus includes ~12 million lakes larger than 0.005 km2, with a total lake area of ~3.4 × 106 km2. Consistent with previous studies, large lakes (>100 km2) dominate global lake area (accounting for 56.1%; Figure 6a), while small lakes (<1 km2) overwhelmingly dominate the global lake count (accounting for 98.3%; Figure 6b), with lakes smaller than 0.01 km2 (not covered by existing publicly available datasets) contributing 32.6% of the total lake count. Figure 6 illustrates the spatial distribution of lake count and total area in GLAKESplus within 1° × 1° grid cells, as well as their longitudinal and latitudinal profiles. Globally, 49% of global lakes or 30% of total lake area are located north of 57°N, mainly distributed in the Canadian Shield, Scandinavia, and Western Siberian Plain; South of 57°N, lake count gradually decreases, while lake area remains stable until 36°N, after which it begins to decline; Small peaks in lake count appear around 30°N and 22°N, mainly contributed by by dense lake distributions in the Mississippi River Basin (North America) and the lower Yangtze River floodplain (Asia). In tropical regions, both lake count and total area are relatively low, but several high-density regions exist, such as coastal southern China, Cambodia, and eastern India. Additionally, the presence of large lakes in the Amazon Basin (South America) and the East African Rift (Africa) results a small peak in lake area between 1°S and 5°S. In the longitudinal profiles, 54% of global lakes (or 48% of total lake area) are distributed in the Western Hemisphere (-20°W to 160°E). In the Eastern Hemisphere, lake area exhibits two prominent peaks at 29–35°E (East African Rift lakes,) and 49–53°E (Caspian Sea), while lake count shows multiple peaks between 66–120°E (Figure 6d).

Figure 6. Spatial distribution of GLAKESplus. Lakes in GLAKESplus were mapped, showing (a) lake area density (total lake area/grid area) and (b) lake count per 1°×1° grid cell. The latitudinal and longitudinal lake profile, summarizing global lake count and total lake area by 1°, were shown in (c) and (d), while (a) and (b) present the statistics for small (<1 km2), medium (1–100 km2), and large (>100 km2) lakes, respectively.

3.2 Comparison with other global lake databases

We compared GLAKESplus with GLAKES and PLD to evaluate its strengths and limitations in representing global lake distributions.The comparison involved differences in total lake count and lake area across different size groups (Figure 7) as well as pixel-scale boundary differences. It should be noted that the three datasets define lake boundaries differently. For lakes larger than 0.01 km2, all three datasets showed an increase in lake count and a decrease in total lake area as size decreased. All datasets contained 16 lakes larger than 10,000 km2, but the total lake area in PLD was slightly smaller than that in GLAKESplus and GLAKES. In other size groups, GLAKESplus consistently included more lakes and a larger total area than the other datasets, with the differences increasing as lake size decreased. GLAKES did not include lakes smaller than 0.3 km2, resulting in the smallest lake count and area in the 0.01–1 km2 size group, whereas it exceeded PLD in all other size groups; Compared to PLD, GLAKESplus mapped an extra 3.9 million lakes smaller than 0.01 km2, with a total lake area of 27,643 km2. At pixel-level scale (Figure 8) , GLAKESplus provided more detailed lake boundaries than GLAKES and PLD due to higher spatial resolution (30 m to 10 m). In pond-dense regions, such as the Yangtze River Basin in China (Figure 8a), GLAKESplus accurately and comprehensively delineated small lake boundaries, while GLAKES underperformed, and PLD did not account for these ponds at all. For salt lakes in arid regions, GLAKES and PLD showed partial omissions, whereas GLAKESplus integrated lakes from GLAKES and PLD, leading to more comprehensive lake boundaries (Figure 8b). In some floodplain regions (e.g., the lower Ob River floodplain; Figure 8c) and semi-arid regions (e.g., eastern Argentina; Figure 8d), the lake count and area of GLAKES and PLD are larger than GLAKESplus, GLAKESplus omitted some lakes with low NDWI value in sentinel-2 mean imagery.

Figure 7. Comparisons of lake count and area in different size groups among GLAKESplus, GLAKES and PLD.

Figure 8. Regional comparison among GLAKESplus, GLAKES and PLD. (a) Ponds in the Yangtze River Basin. (b) The Uyuni Salt Flat in Bolivia. (c) Lakes in the lower Ob River floodplain. (d) Seasonal lakes in eastern Argentina. For (a-d), the left figures show the NDWI basemap and lake extend of GLAKES and PLD, the right figures show the RGB basemap and lake extend of GLAKESplus.

We conducted a more detailed spatial comparison between GLAKESplus and PLD (Figure 8). For lakes larger than 0.1 km2, the total lake count and area distributions along latitude and longitude in GLAKESplus are generally similar to those in PLD, but for lakes smaller than 0.1 km², GLAKESplus exhibits significantly higher values than PLD (Fig. 9a, b, d, e). We further compared the total area and count of lakes with area between 0.01–0.1 km2, which are included in both datasets, within 1° × 1° grid cells (Fig. 8c). The results indicate that, except for a few floodplains and arid inland regions mentioned earlier, GLAKESplus contains more lakes and a greater total lake area than PLD across most regions. The differences are particularly pronounced in the Yangtze River Basin in China, eastern India, and Southeast Asia.

Figure 8 The differences in the spatial distribution of total lake area and count between GLAKESplus and PLD (excluding the Caspian Sea, which contains the Garabogazköl lagoon). Lake count of two databases (a) per latitudinal degree and (b) per longitude degree. Total lakes area of two databases (d) per latitudinal degree and (e) per longitude degree. For (a, b, d and e), lakes greater than 0.1 km2 and smaller than 0.1 km2 were plotted on opposite axes, with the axes were stretched for better visual presentation. GLAKESplus was represented by stacked bar charts figure in different colors, while PLD was represented by line charts in different colors and line widths. (c) The difference of count and total area of lakes between 0.01 and 0.1 km2 between two databases per 1°×1° grid cell.

4 Discussion and Conclusions

Based on Sentinel-2 satellite data collected between March 28, 2017 to April 10, 2022, and a deep learning method, we proposed the first 10 m resolution global lake database. GLAKESplus contains ~12 million lakes larger than 0.005 km2, covering a total area of 3.4 × 106 km2. Accuracy assessments indicate that our model achieves high accuracy, with recall, precision, and F1-score exceeding 92.75 and an IoU above 82.7 on the test set.

However, our model exhibits some limitations in certain regions. Specifically, due to the relatively small input image size, the model faces challenges in distinguishing large-scale water bodies. This results in the misclassification of some wide rivers as lakes (with a precision of 89.77% in AR regions) and the omission of certain large lakes with river-like characteristics (with an average recall of 93.1% for test labels >1 km2, slightly lower than that of smaller size groups). Increasing the input size would significantly raise GPU memory demands, and 576 × 576 pixels is the largest input size our GPU memory (48GB) can support during training. Additionally, our model omitted some salt lakes (with a recall of 72.71% in SL regions) due to their spectral characteristics being similar to those of land. Our post-processing operations effectively reduced the omission and commission errors of raw prediction. First, ocean and river residuals were effectively removed by other databases. The river masking operation improving the precision of test regions by 1.77%. Second, polygons smaller than 0.005 km2 with low mean recall and mean precision (both below 85%) were deleted. Finally, a lake completion procedure was implemented to restore boundaries for large lakes and arid-region lakes, ensuring more comprehensive lake delineation. Additionally, we conducted extensive manual inspections to ensure the accuracy of the mapping results.

Compared to existing global lake datasets (Verpoorter et al., 2014; Messager et al., 2016; Sheng et al., 2016; Pi et al., 2022; Wang et al., 2023), GLAKESplus has higher spatial resolution (10m) and more comprehensive coverage of small lakes. Lakes with size <1 km2 contribute 98% of the global lake count, further highlights the importance of small lakes. GLAKESplus surpasses GLAKES and PLD in both the lake count and total lake area across all size groups, while also providing more detailed spatial distribution information. A key highlight is that GLAKESplus comprehensively maps numerous ponds in the middle and lower Yangtze River Plain and the coastal regions of Southeast Asia, which are underestimated in PLD and GLAKES. However, GLAKESplus also has some limitations. In certain areas, its lake predictions are lower than those of GLAKES and PLD. For example, in some floodplains, GLAKESplus tends to predict only perennial water bodies, while in arid regions, it tends to overlook seasonal or dried-up lakes. Additionally, unlike GLAKES, GLAKESplus provides only a single temporal snapshot, lacking information on lake dynamics over time.

The global lake distribution information provided by GLAKESplus contributes to the accurate assessment of inland lake carbon fluxes, supports research on climate change and the biogeochemistry of inland waters, and is of great significance for global water resource conservation. Future research will focus on enhancing the model's ability to distinguish lakes at different scales and exploring the application of the 10m resolution global lake mapping method to different time periods for dynamic monitoring of global lakes.

Supplementary Figure 1. The performance of the deep learning algorithm in different regions. (a) NR; (b) Small river in LR; (c) Large river in LR; (d) BR; (e) IR and (f) SR. For (a-f), the left figures show the RGB basemap, the right figures show the corresponding labels and predictions. The river residuals were shown in grey color.

Supplementary Table 1 Keep hyperparameters tested and adjusted in the U-Net Model.

|  |  |
| --- | --- |
| Hyperparameters | Setting |
| Optimizer | Adadelta |
| Loss function | Dice loss |
| Batch size | 16 |
| steps\_per\_epoch | 691 |
| validation\_steps | 224 |
| Earlystopping patient | 50 |
| Epoch | 200 |
| Patch size | 576,576,6 |

**References**

Brandt, M., Tucker, C. J., Kariryaa, A., Rasmussen, K., Abel, C., Small, J., Chave, J., Rasmussen, L. V., Hiernaux, P., Diouf, A. A., Kergoat, L., Mertz, O., Igel, C., Gieseke, F., Schöning, J., Li, S., Melocik, K., Meyer, J., Sinno, S., … Fensholt, R. (2020). An unexpectedly large count of trees in the West African Sahara and Sahel. *Nature*, *587*(7832), 78–82. https://doi.org/10.1038/s41586-020-2824-5

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

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

Lehner, B., & Döll, P. (2004). Development and validation of a global database of lakes, reservoirs and wetlands. *Journal of Hydrology*, *296*(1), 1–22. https://doi.org/10.1016/j.jhydrol.2004.03.028

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. https://doi.org/10.1080/01431169608948714

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., & Schmitt, O. (2016). Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms13603

Pekel, J.-F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*, *540*(7633), 418–422. https://doi.org/10.1038/nature20584

Pi, X., Luo, Q., Feng, L., Xu, Y., Tang, J., Liang, X., Ma, E., Cheng, R., Fensholt, R., Brandt, M., Cai, X., Gibson, L., Liu, J., Zheng, C., Li, W., & Bryan, B. A. (2022). Mapping global lake dynamics reveals the emerging roles of small lakes. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-33239-3

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation* (arXiv:1505.04597). arXiv. https://doi.org/10.48550/arXiv.1505.04597

Sheng, Y., Song, C., Wang, J., Lyons, E. A., Knox, B. R., Cox, J. S., & Gao, F. (2016). Representative lake water extent mapping at continental scales using multi-temporal Landsat-8 imagery. *Remote Sensing of Environment*, *185*, 129–141. https://doi.org/10.1016/j.rse.2015.12.041

Verpoorter, C., Kutser, T., Seekell, D. A., & Tranvik, L. J. (2014). A global inventory of lakes based on high-resolution satellite imagery. *Geophysical Research Letters*, *41*(18), 6396–6402. https://doi.org/10.1002/2014GL060641

Wang, J., Pottier, C., Cazals, C., Battude, M., Sheng, Y., Song, C., Sikder, M. S., Yang, X., Ke, L., Gosset, M., Oliveira, R. R. A., Grippa, M., Girard, F., Allen, G. H., Biancamaria, S., Smith, L., Crétaux, J.-F., & Pavelsky, T. M. (2023). *The Surface Water and Ocean Topography Mission (SWOT) Prior Lake Database (PLD): Lake mask and operational auxiliaries* [Preprint]. Preprints. https://doi.org/10.22541/au.170258987.72387777/v1

Wang, J., Walter, B. A., Yao, F., Song, C., Ding, M., Maroof, A. S., Zhu, J., Fan, C., McAlister, J. M., Sikder, S., Sheng, Y., Allen, G. H., Crétaux, J.-F., & Wada, Y. (2022). GeoDAR: Georeferenced global dams and reservoirs dataset for bridging attributes and geolocations. *Earth System Science Data*, *14*(4), 1869–1899. https://doi.org/10.5194/essd-14-1869-2022

Williamson, C. E., Saros, J. E., Vincent, W. F., & Smol, J. P. (2009). Lakes and reservoirs as sentinels, integrators, and regulators of climate change. *Limnology and Oceanography*, *54*(6part2), 2273–2282. https://doi.org/10.4319/lo.2009.54.6\_part\_2.2273

Yang, X., Qin, Q., Yésou, H., Ledauphin, T., Koehl, M., Grussenmeyer, P., & Zhu, Z. (2020). Monthly estimation of the surface water extent in France at a 10-m resolution using Sentinel-2 data. *Remote Sensing of Environment*, *244*, 111803. https://doi.org/10.1016/j.rse.2020.111803

Zheng, K., Lin, P., & Yin, Z. (2024). SHIFT: A spatial-heterogeneity improvement in DEM-based mapping of global geomorphic floodplains. *Earth System Science Data*, *16*(8), 3873–3891. https://doi.org/10.5194/essd-16-3873-2024