https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

Glacial-Lake-Bench: A Global Multi-Sensor Benchmark Dataset for Evaluating Deep Learning Models for Glacial Lake Mapping

Saurabh Kaushik[1*], Beth Tellman[1], Ian Howat[2], Umesh Haritashya[3,4]

[1]University of Wisconsin, Madison-USA.
[2]Byrd Polar and Climate Research Center, The Ohio State University, Columbus, USA.
[3]Department of Earth and Environmental Geosciences, University of Dayton, Dayton, OH, USA
[4]Sustainability Program, University of Dayton, Dayton, OH, USA
Correspondence: skaushik8@wisc.edu

Abstract

Glacial lakes are among the most sensitive indicators of climate change, closely linked to natural hazards and are natural reservoirs of freshwater resources. Thus, automated mapping and monitoring of glacial lakes is imperative. However, most automated approaches either remain regional in scope or show limited performance under challenging conditions such as cloud cover, shadows, and spatially small or frozen lakes. The global scale analysis and comparative evaluation of deep learning models is primarily hindered by the lack of readily available datasets for training data. To address this gap, we present Glacial Lake-Bench (GLB), a multisource remote sensing dataset comprising Sentinel-2, Sentinel-1, and Copernicus DEM-derived terrain (11 channels in total). GLB consists of 19,115 image-label pairs (256x256x11) spanning all Randolph Glacier Inventory (RGI) regions except Antarctica, providing the first-ever global, multi-sensor dataset for glacial lake segmentation. In addition, we compiled Glacial Lake-Bench-Challenge (GLBC), a curated subset of 1,105 image-label pairs representing scenes with cloud cover, shadow, frozen lake surfaces, and small lakes to establish a community standard for evaluating model robustness under difficult conditions. Labels are derived from Zhang et al. (2024); we independently quantify label quality through stratified sampling of 50 image-label pairs from each RGI region, amounting to 900 chips in total. Our quality assessment reveals a mean Intersection over Union (mIoU) of 0.95, precision of 0.99, recall of 0.96, and per-region agreement that is consistent with the known difficulty of small, turbid, shadowed lakes in high-mountain terrain. To demonstrate that the dataset is usable, well-posed, and appropriately challenging, we provide reference baselines from

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

two convolutional networks (U-Net, DeepLabv3+) and two Geo-Foundation Models (GFMs) (DOFA, Prithvi-EO-2.0), evaluated with a recommended leave-one-region-out (LORO) protocol that minimizes spatial autocorrelation, alongside a random split and the GLBC subset. Baseline mIoU reaches 0.80–0.85 on GLB and drops to 0.74–0.79 on GLBC, confirming that the challenge subset isolates genuinely difficult conditions. The GLB dataset is available at https://zenodo.org/records/17917359 (Kaushik, 2026)

1. Introduction

Glacial lakes can be categorized depending on their origin and geomorphological setting. For example, ice-dammed lakes may form on the glacier surface as supraglacial lakes or be impounded by glacial ice between the glacier margin and adjacent valley walls. Another common type is moraine-dammed lakes, which develop between the retreating ice terminus and the end moraine following glacier recession. Additional lake types include cirque lakes and lakes within glaciated landscapes that are not directly connected to contemporary glaciers (King et al., 2019; Zhang et al., 2023). The response of such glacial lakes toward climate change in terms of areal expansion or development of new glacial lakes is directly linked with glacier mass wastage (King et al., 2019; Zhang et al., 2023), an acceleration of surface velocity due to meltwater reaching the bed (Pronk et al., 2021), and enhanced absorption of solar radiation that further accelerates glacier change (Zhang et al., 2024). The continued formation and expansion of glacial lakes also increase the risk of Glacial Lake Outburst Floods (GLOFs), which can cause large-scale damage to infrastructure and fatalities in downstream regions (Taylor et al., 2023; Zheng et al., 2021). Thus, glacial lakes represent one of the key indicators of climate change and are closely associated with glacial hazard (Taylor et al., 2023; Zheng et al., 2021). In addition to their role as a barometer of glacier response, glacial lakes are considered natural freshwater reservoirs, though they remain largely unutilized due to engineering challenges (Immerzeel et al., 2020).

Given the high relevance of glacial lakes, Zhang et al., (2024) mapped global glacial lakes (>0.002 km^2) for two time periods (1990 and 2020), demonstrating 54%, 11%, and 9% increases in global glacial lake number, area, and volume, respectively. However, interannual and seasonal variability remains understudied. Li et al. (2025) reported that 66% of global lakes exhibit strong seasonality, particularly in regions where 90% of the world's population resides. Similar findings by Pi et al. (2022) highlight the prominent role of smaller lakes in global lake-area variability. For

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

example, small lakes (<1 km^2) constitute only ~15% of total global lake area, yet dominate variability in half of all inland lake regions (Pi et al., 2022). However, minimum lake area thresholds used in both studies (Li et al., 2025; Pi et al., 2022) were limited to 0.03 km^2. Despite their importance, the interannual and seasonal variability of small glacial lakes remains poorly characterized.

In the last decade, automated mapping and monitoring methods have gained wide scientific attention. Deep Learning (DL) approaches are primarily dominated by Convolutional Neural Networks, with U-Net and DeepLab variants being the most widely used (Tom et al., 2025). Overall, DL based methods have been found to be most successful across various spatiotemporal scales (Dirscherl et al., 2021; Jiang et al., 2025; Kaushik et al., 2020, 2022; Ma et al., 2025; Tang et al., 2024; Wang et al., 2022). For example, Tang et al., (2024) generated the first temporal glacial lake inventory across High Mountain Asia (HMA) using Landsat data and DeepLabv3+ architecture. However, the method exhibited reduced performance in several regions (e.g., West Kun Lun and Western Himalaya), achieving only F1 scores of 0.70 and 0.79, respectively. Similarly, the method proposed by (Jiang et al., 2025) also showed severe limitations in spatial transferability, demonstrating only 0.56 mean Intersection over Union (mIoU) in West Greenland. The literature reveals that most automated approaches (Dirscherl et al., 2021; Hu et al., 2024; Kaushik et al., 2022; Ma et al., 2025; Wang et al., 2022) are either regionally constrained, sensitive to lake size, or exhibit reduced performance under challenging conditions such as cloud cover, shadow, smaller lakes, and frozen lakes. As a result, globally applicable glacial lake mapping algorithms remain lacking, particularly those demonstrating comprehensive assessments of spatial transferability on the global scale. This limitation has limited the measurement of spatiotemporal variations in glacial lakes at both annual and seasonal scales, especially for small lakes (<0.03 km^2).

Despite significant development in the domain, lack of readily available, multisource remote sensing dataset for training and evaluating DL models remains a key limiting factor in establishing globally applicable glacial lake mapping solutions and setting a community standard for improving glacial lake mapping in challenging conditions. The systematic evaluation of model architecture and the establishment of methodological advances are directly hindered by the lack of a community-standard benchmark dataset. To address this, we compiled the first ever global

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

multisource remote sensing dataset of glacial lakes, Glacial Lake-Bench (GLB) collected across each RGI region (except Antarctica). The dataset consists of 19,115 image-label pairs (256 x 256 x 11), consists of 11 channels: Blue, Green, Red, NIR, SWIR1, SWIR2, NDWI, Slope, Elevation, SAR-VV, and SAR-VH. In addition, we generated a subset test dataset, Glacial Lake-Bench-Challenge (GLBC), consisting of 1,105 image-label pairs representing challenging conditions (e.g., sparse or fully cloud-covered scenes, frozen lakes, shadow, and small lakes $\leq$0.05 km^2), reflecting a highly imbalanced class distribution. These curated global samples provide a valuable resource for testing the applicability of newly developed models, particularly in scenarios where existing algorithms face significant limitations.

To evaluate the dataset's applicability, we conducted a comprehensive comparative evaluation using state-of-the-art CNN models and Vision Transformer (ViT)-based Geo-Foundation Models (GFMs): Prithvi 2.0 (Szwarcman et al., 2024) and Dynamic One-For-All (DOFA) (Xiong et al., 2024), as representatives of pretrained encoders. This evaluation highlights the advantages of recently developed GFMs over traditional CNNs for global glacial lake semantic segmentation, particularly under challenging environmental conditions. We further performed leave-one-region-out cross-validation on the top models to quantify spatial transferability. Baseline results for both global and challenging conditions are provided. The GLB and GLBC datasets together serve as a benchmark resource to advance robust, globally scalable glacial lake mapping capabilities and improving site specific algorithms in most challenging conditions.

Our key contributions can be summarized as follows:

1) Dataset Release: We compile and release the first-ever readily available multisource remote sensing dataset, Glacial Lake Benchmark (GLB), designed for evaluating deep learning models under global-scale and challenging conditions for glacial lake mapping.
2) Model Evaluation: We conduct a systematic comparative evaluation of state-of-the-art deep learning models for global spatial transferability using a leave-one-region-out approach.
3) An efficient data-processing pipeline that seamlessly generates multi-sensor image–label pairs using only a label shapefile, an AOI shapefile, and user-specified date range and cloud-cover thresholds.

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

2. Glacial Lake-Bench Dataset

2.1. Data Acquisition and processing

To generate Glacial Lake-Bench, we developed a comprehensive pipeline to systematically acquire and process Sentinel-1, Sentinel-2, and Digital Elevation Model (DEM) data (Fig. 1). The pipeline requires the following inputs: a glacial lake boundary shapefile, a date range, an Area of Interest (AOI) shapefile, and a cloud-cover threshold. We used global glacial lake boundaries provided by Zhang et al., (2024), which were derived from 2020 Sentinel-2 images with minimal cloud cover during the ablation season. To ensure consistency, we followed the same framework by collecting 2020 Sentinel-2 images during the ablation season and filtering those with cloud cover of ≤15%.

The pipeline begins by acquiring Sentinel-2 data from the AWS client based on the specified date range, region, and cloud-cover threshold. To optimize computing resources, preliminary maps are generated to inform the user about tile extents and area coverage, with specific cloud cover overlaid on the glacial lake shapefile. This step provides precise information on the total glacial lake area within the specified region and the portion covered by satellite imagery under the given constraints. We additionally generate preview maps showing tile extents, cloud-cover distribution, and spatial overlap with lake polygons to facilitate data quality control. Once Sentinel-2 data are acquired, the SWIR bands are resampled to 10 m to match the spatial resolution of the other multispectral bands (Fig. 1). We selected the Blue, Green, Red, NIR, SWIR-1, and SWIR-2 bands, and computed the Normalized Difference Water Index (NDWI =(Green − NIR) / (Green + NIR)) to enhance water detection. These six bands are then stacked for further processing alongside DEM and Sentinel-1 SAR data. We acquire the Copernicus DEM (30 m) from the AWS S3 bucket and reproject and resample using bilinear interpolation to match the spatial resolution and projection of the Sentinel-2 imagery. Slope is then computed from the DEM, and slope and elevation are stacked as Bands 8 and 9, respectively (Fig. 1). Sentinel-1 Radiometrically Terrain-Corrected (RTC) images are acquired within a ±5-day temporal window, with preference given to the closest available acquisition via the Planetary Computer API. We extract both Vertical–Vertical (VV) and Vertical–Horizontal (VH) polarization images, which are stacked as Bands 10 and 11.

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

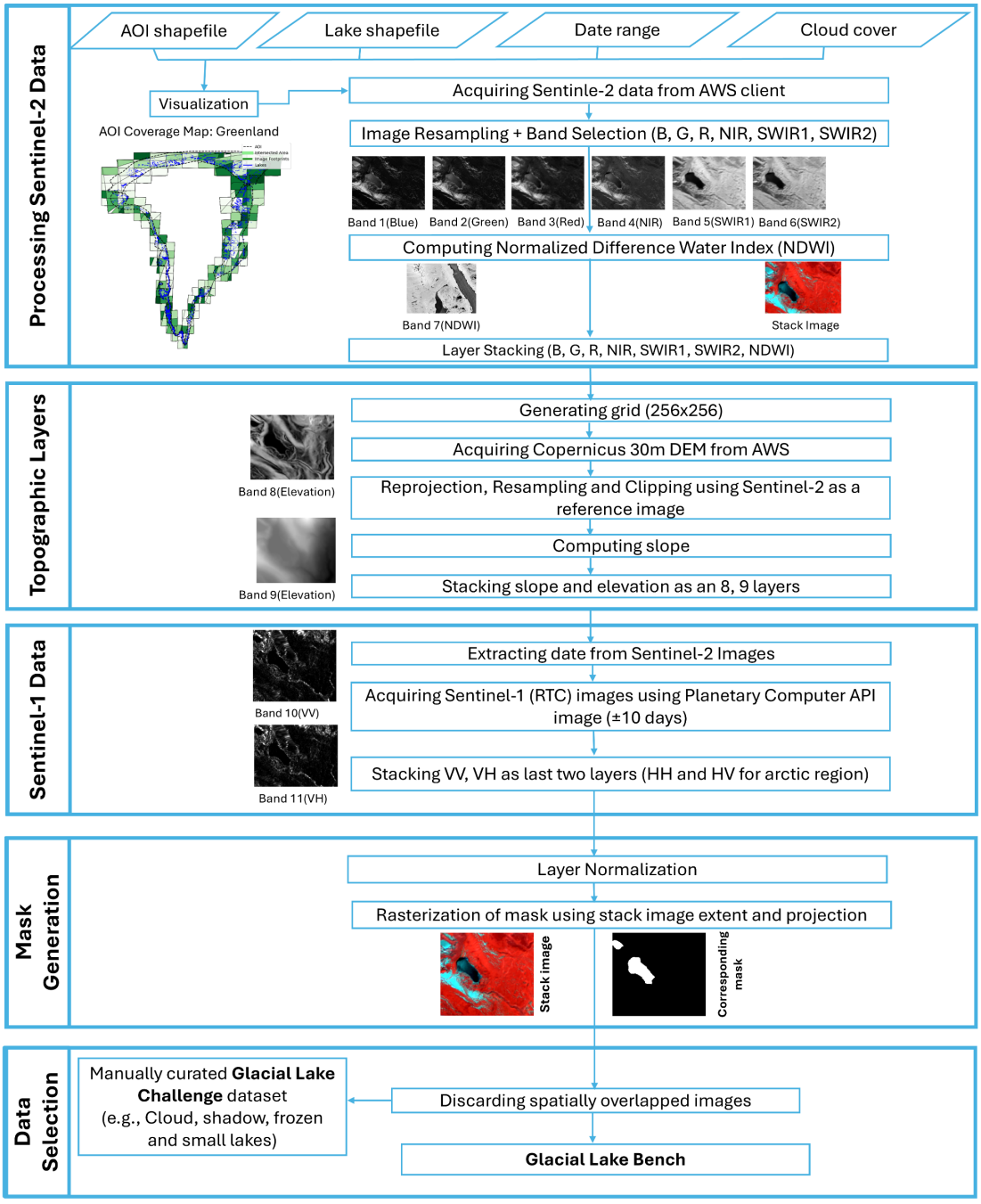
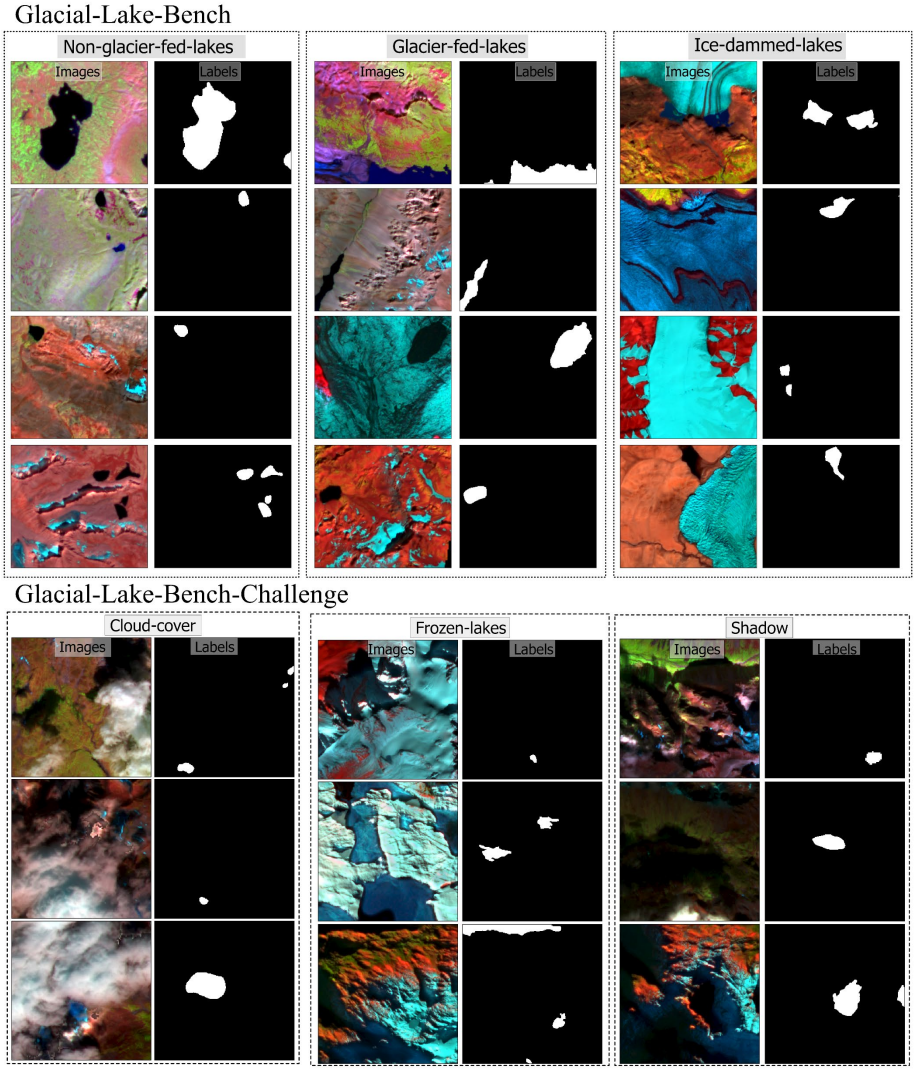


Fig. 1 Overall workflow for generating the Glacial Lake-Bench dataset. We used glacial lake boundaries provided by Zhang et al. (2024) and acquired data during the ablation season to

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

153 minimize the influence of seasonal snow. The Glacial Lake Challenge dataset is curated through
154 manual selection of images captured under challenging conditions, including heavy and partial
155 cloud cover, frozen lakes, shadows, and small lake features.



156

157 **Fig. 2** Examples of three major types of glacial lakes included in the Glacial Lake-Bench dataset
158 across the globe. The images display a Near-Infrared (NIR), Red, and Green band combination.
159 Lake boundaries are derived from (Zhang et al., 2024). The generated labels indicate lake pixels
160 with a value of 1 and background pixels with a value of 0.

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

For Arctic regions, where VV and VH data may be unavailable, Horizontal–Horizontal (HH) and Horizontal–Vertical (HV) polarization images are used instead. Once all bands are assembled, the stacked images are normalized using a min-max stretch to a range of 0–1 to ensure compatibility with deep learning models. We generate segmentation masks from glacial lake boundaries provided by Zhang et al. (2024) . Lake polygons are rasterized onto the 10 m grid of each image using consistent extent, projection, and spatial resolution. Pixels falling within lake boundaries are assigned a value of 1, while all other pixels are labeled as 0, representing background. In the final step, spatially overlapping tiles are discarded from the GLB dataset to ensure uniqueness and avoid redundancy. We additionally curated a manually selected subset of the GLB dataset through detailed visual analysis, focusing on the most challenging glacial lake mapping conditions. This subset includes scenes affected by cloud cover; frozen or partially frozen lakes; lakes influenced by terrain shadow; lakes exhibiting mixed spectral signatures or high turbidity; and small lakes (≤ 0.05 km^2). This targeted subset is designed to enable rapid and focused evaluation of newly developed or existing algorithms, particularly those aimed at improving glacial lake mapping performance under challenging conditions. Finally, examples of compiled GLB and GLBC data are shown in Figure 2.

2.2. Spatial Coverage

To support global-scale analysis and algorithm development, we curated the Glacial Lake Bench (GLB) dataset across all RGI regions (Table 1, Fig. 3), enabling end users to evaluate algorithms at both regional and global scales. The inclusion of data from diverse geographical and climatological regions introduces substantial variability, allowing models to learn the appearance of glacial lakes across a wide range of environmental settings. In terms of spatial coverage, GLB exhibits its strongest representation in the Alaska region, covering 89.42 % of the total lake area. Other regions including North Asia, Svalbard, Central Asia, and South Asia West also demonstrate strong coverage, with 64.15 %, 78.71 %, 65.20 %, and 78.81 %, respectively. However, we acknowledge that, likely due to the applied cloud-cover threshold (15 %), GLB under-samples certain regions such as South Asia East, Iceland, and Scandinavia, with coverage of 9.21 %, 11.67 %, and 19.08 %, respectively. Nevertheless, given the overall size and spatial extent of the dataset, GLB still provides a sufficient and representative sample to support the development of scalable, globally applicable approaches.

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

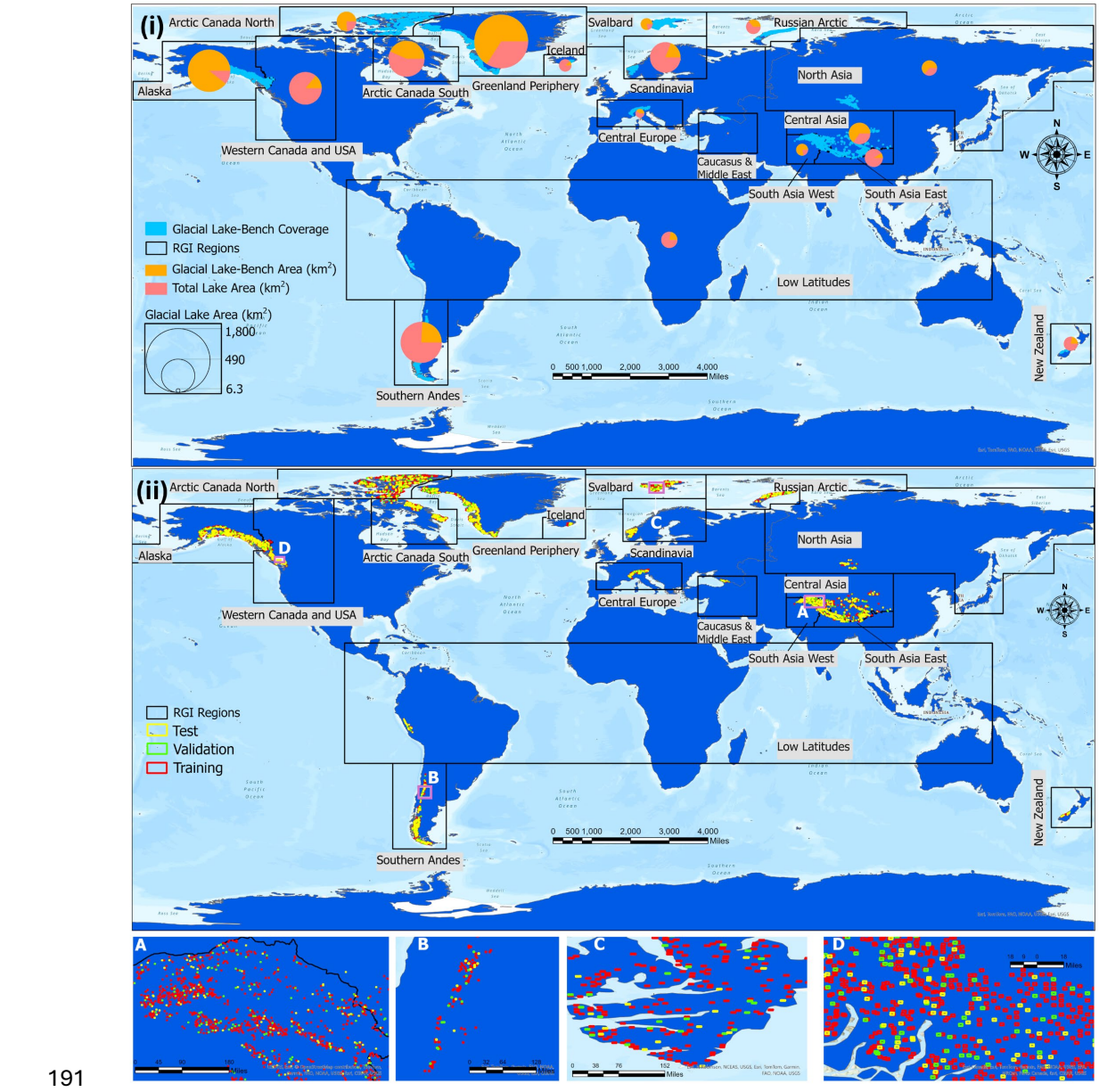


Fig. 3 Compilation of Glacial Lake-Bench collected globally across each RGI region. (i) spatial footprint of Glacial Lake-Bench coverage, along with the ratio between the total glacial lake area and the lake area represented in our compiled dataset. (ii) exhibiting spatial distribution of training, test, and validation sites.

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

196 **Table 1.** Summary of the compiled Glacial Lake-Bench dataset, including the number of images
197 per RGI region, the total lake area covered by the dataset, and a comparison with the total glacial
198 lake area reported by Zhang et al. (2024), from which the lake labels were derived.

RGI Region	No. Images	Total Lake Area (km²)	Area Covered by GLB Dataset (km²)	% Area Covered by GLB
Alaska	3323	3549.55	3174.64	89.42%
Iceland	65	306.50	35.78	11.67%
New Zealand	133	393.67	87.43	22.21%
North Asia	261	467.02	299.75	64.15%
Scandinavia	595	1840.02	351.08	19.08%
Svalbard	487	278.30	219.05	78.71%
Arctic Canada North	1558	767.37	583.21	75.99%
Arctic Canada South	1074	2567.49	1112.91	43.36%
Central Asia	1982	921.79	600.95	65.20%
Central Europe	521	156.84	63.92	40.77%
Greenland	2287	5725.42	3775.64	65.94%
Low Latitudes	412	524.83	116.06	22.11%
Middle East *(Caucasus)*	97	6.31	3.81	60.37%
Russian Arctic	344	420.33	152.12	36.19%
Southern Andes	2077	3329.34	802.32	24.09%
South Asia East	638	649.21	59.80	9.21%
South Asia West	1684	283.99	223.88	78.81%
Western Canada and USA	1612	2107.74	296.66	14.07%

199

200 ### *2.3. Temporal Coverage*

201 GLB is a single-timestamp dataset sampled during the 2020 ablation season, chosen to coincide
202 with the imagery and timing used to generate the Zhang et al. (2024) labels while minimizing the
203 impact of seasonal snow. This design supports robust spatial benchmarking and within-season
204 transfer, but it does not capture seasonal or interannual change. We state this scope explicitly so
205 that users do not over-interpret the dataset for multi-temporal tasks, and we identify multi-temporal
206 extension as the primary planned development.

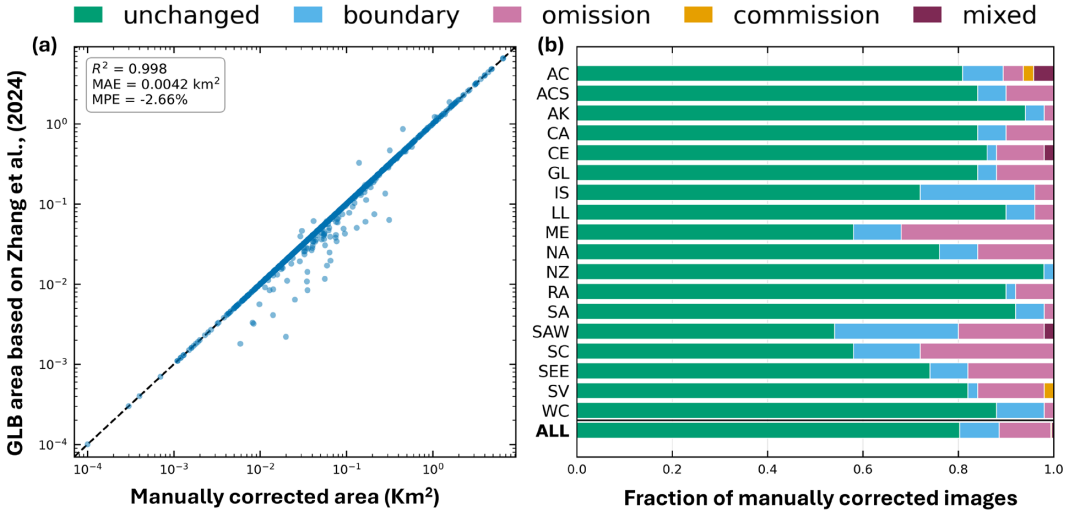
207

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

2.4. *Data quality and Uncertainty*

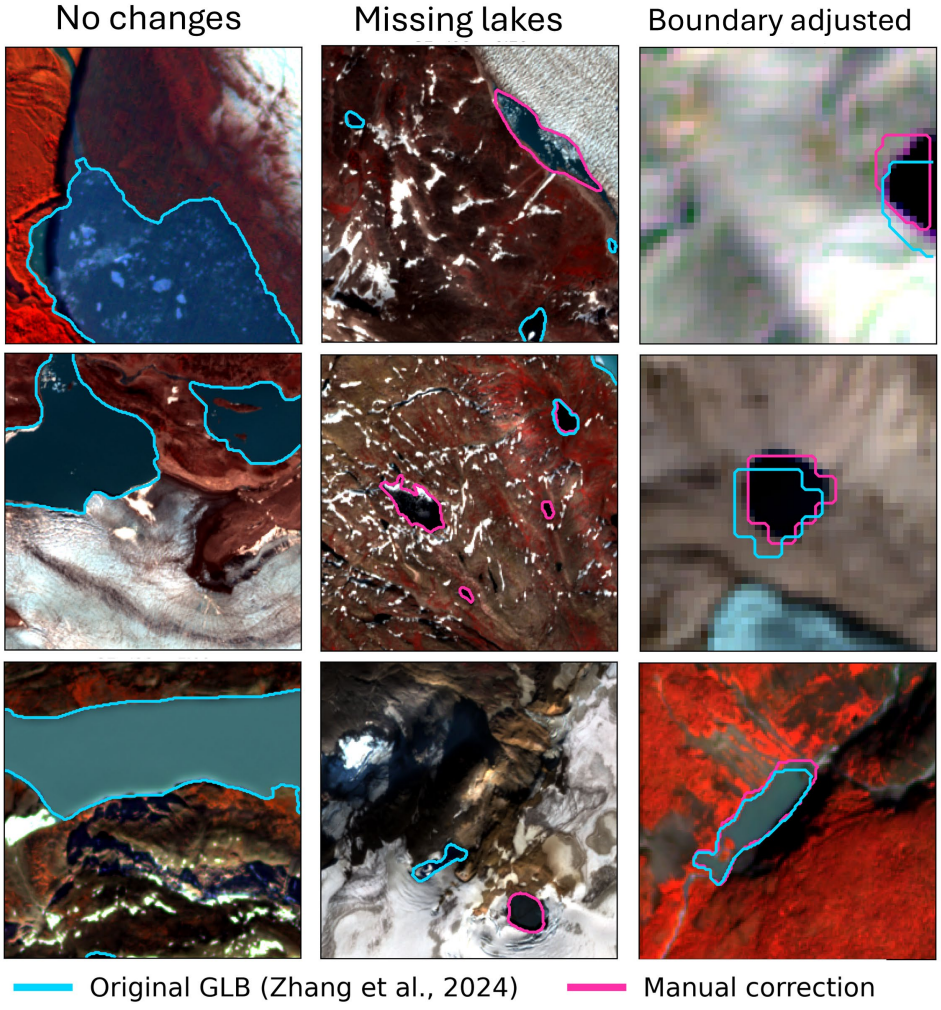
As our labels are derived from Zhang et al. (2024), we quantified the associated uncertainty in the GLB dataset to assess image–label agreement. We selected 50 image–label pairs from each RGI region using stratified sampling, amounting to 900 image–label pairs, which we manually reviewed and corrected to estimate uncertainties. Overall, GLB labels show strong agreement with the manually corrected lake boundaries, achieving a mean Intersection over Union (mIoU) of 0.95, precision of 0.99, and recall of 0.96. In addition, correlation analysis yielded an R^2 of 0.99 (Fig. 4a), indicating excellent consistency between the two. Visual inspection suggests that ~80% of the images required no correction. However, a subset of image chips exhibited missing lakes (omissions) and minor boundary adjustments depending on scene characteristics (Fig. 4b). Furthermore, our analysis reveals substantial regional variability in agreement (Fig. 4b). Agreement is near-perfect in regions with large, optically clear lakes, such as New Zealand, the Southern Andes, Iceland, and Low-Latitude regions (IoU ≥ 0.98) and weakest in high-mountain environments characterized by small, turbid lakes and terrain shadows. These include South Asia West (IoU = 0.83), the Middle East (Caucasus) (area bias −9%), South Asia East, and Scandinavia. This regional variability is physically consistent with the known challenges of mapping small supraglacial and proglacial ponds in steep, shadow-prone High Mountain Asia terrain. Representative examples of the three major cases, no change, missing lakes (omission), and boundary adjustments are shown in Figure 5.



https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

228 **Fig. 4** Uncertainty evaluation of GLB based on manual correction of 900 image-label pairs. A total
229 of 50 samples (image-label pair) were selected from each RGI region using stratified sampling. (a)
230 The overall GLB dataset shows excellent agreement between images and lake boundaries from
231 Zhang et al. (2024), with an R^2 of 0.99. (b) Region-wise fraction of images that required corrections



232

233 **Fig. 5** Examples from visual inspection and manual correction support our quantitative analysis:
234 ~80% of image-label pairs required no changes, while in some cases missing lakes (omissions)
235 and minor boundary adjustments were identified.
236
237

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

2.5. Data Folder Structure and Naming Convention

The dataset is readily available in GeoTIFF format, allowing users to directly integrate the data into deep learning (DL) models without additional preprocessing. The folder structure follows a standard train–validation–test split commonly used in machine learning workflows, ensuring ease of use and reproducibility. Each split contains co-registered image and corresponding label directories, where images and ground-truth lake labels share identical filenames to enable straightforward pairing during model training and evaluation. The accompanying README.txt provides additional details on dataset usage, metadata, and benchmarking recommendations. The naming convention encodes region, sensor, spatial tile, acquisition date, and processing level. For example, AC_S2A_15XVJ_20200810_4_L2A_36 denotes a Sentinel-2A Level-2A image acquired on 10 August 2020 over the Arctic Canada RGI region, mapped to MGRS tile 15XVJ. The full mapping between region names and their codes is documented in the README file.

3. The Baseline

3.1. Baseline Models

We employ four deep learning (DL) models to establish baseline accuracy metrics for the GLB and GLBC datasets. These baseline results serve as a reference for future model development and enable direct comparison with standard DL approaches. To ensure a comprehensive evaluation, we incorporate two recently developed Geo-Foundation Models (GFMs), Prithvi (v2, 600M) and DOFA alongside two widely used convolutional neural network (CNN) architectures, U-Net and DeepLabv3+ (Table 2). Our selection is motivated by different models' architectures, pretraining, and segmentation performance. Prithvi 2.0 was chosen for its large encoder pretrained on 4.2M image chips, strong performance on semantic segmentation tasks within the GEO-Bench framework (Lacoste et al., 2023), and its unique capability to generate spatiotemporal embeddings, which have proven useful in glacial lake mapping (Jiang et al., 2025). Traditional glacial lake mapping algorithms are often tailored to specific remote sensing sensors or data types (e.g., optical or radar), limiting their ability to generalize across diverse datasets. As a result, systematic multi-modal integration remains largely unexplored. In this context, the DOFA GFM offers unique capabilities by adaptively integrating diverse data sources within a single framework through its neural-plasticity-inspired dynamic patch embedding. We compared the performance of DOFA and Prithvi with state-of-the-art remote sensing segmentation models like U-Net (Ronneberger et al.,

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

2015) and DeepLabv3+ (Chen et al., 2018), given their wide applicability in glacial lake mapping in diverse geographies (Dirscherl et al., 2021; Jiang et al., 2025; Tang et al., 2024; Tom et al., 2025). The selection of these models as baseline is motivated by recent study (Tom et al., 2025) emphasizing these two accounts for approximately 60% of DL base glacial lake mapping approaches.

We employed the 600M-parameter version of Prithvi 2.0 (approximately 685M trainable parameters; Table 2) using Terratorch. Prithvi includes a temporal ViT encoder enriched with temporal and geolocation embeddings. This encoder can be paired with various decoders (e.g., UperNet, FCN, Identity) for segmentation tasks and supports both full fine-tuning and frozen encoder configurations. Here, we fully fine-tuned the encoder, passing the generated embeddings (feature masks) to the UperNet decoder (Xiao et al., 2018) to produce segmentation outputs. The input dataset (256 × 256) underwent data augmentation, including vertical and horizontal flips (p=0.5), random rotations, translations, and scale perturbations to improve model robustness and generalizability. Prithvi was originally pretrained on six input channels (Blue, Green, Red, NIR, SWIR-1, and SWIR-2), therefore, we tested multiple band-subset configurations to identify an optimal six band combination for global mapping. Final results are presented using the Blue-Red-NIR-SWIR1-Slope-VV band combination. For semantic segmentation, we used the UPerNet decoder (Xiao et al., 2018), which combines a Feature Pyramid Network (FPN) with convolutional layers to effectively capture multi-scale features.

DOFA was fed with Glacial Lake-Bench dataset (256×256×11), where the 11 channels include Sentinel-2 optical bands (e.g., blue, green, red, near-infrared, SWIR-1, and SWIR-2), NDWI, Sentinel-1 VV and VH polarizations, and DEM-derived slope and elevation. The central wavelength of each spectral band is supplied to DOFA's dynamic patch embedding module, which adapts the model's input weights based on band-specific spectral properties. In case of slope and elevation layer we used green and red band central wavelength as proxy. Consistent with ViT architectures, each 256×256 image is partitioned into 16×16 non-overlapping patches (256 patches total). Each patch (16 × 16 × 11) is flattened and linearly projected (matrix W_e) into the Transformers embedding dimension d (e.g., 768). This operation can be written as: $Z_i = W_e \times X_i$ where $x_i \in \mathbb{R}^{2816}$ is the flattened patch and $z_i \in \mathbb{R}^d$ is the embedded token. This yields a sequence of 256 tokens that DOFA processes using sinusoidal positional encoding and Transformer blocks. For segmentation, the latent representation is reshaped into a 2D grid (e.g.,

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

299 16×16×d) and passed to a UPerNet decoder, which up-samples the features to the original
300 resolution (256×256) to produce the final segmentation mask.

301 Table 2. Selected model's type, inference cost estimated using Giga Floating Point Operations
302 (GFLOPs), and trainable parameters.

Model	Model type	Trainable parameters (m)	Inference cost (GFLOPs)	References
U-Net	CNN	31	97.2	(Ronneberger et al., 2015)
DeepLabv3+	CNN	41	55.78	(Chen et al., 2018)
Prithvi V2	ViT	685	406.83	(Szwarcman et al., 2024)
DOFA-Base	ViT	121.1	41.8	(Xiong et al., 2024)

303

304 ***3.2.Model Training***

305 All computational experiments were conducted using an NVIDIA RTX A6000 GPU. The study
306 was structured in three phases. In Phase 1, we evaluated all deep learning (DL) models using a
307 randomly split Glacial Lake-Bench dataset (Fig. 3(ii)), with 80% of the data used for training, 10%
308 for validation, and 10% for testing. This phase provided an initial benchmark and informed later
309 model selection. In Phase 2, all models were trained on the full Glacial Lake-Bench dataset,
310 excluding tiles with filenames matching those in the Glacial Lake-Bench-Challenge (GLBC) set
311 to avoid leakage. The resulting models were evaluated on the manually curated GLBC dataset. In
312 Phase 3, we assessed the global model applicability by conducting a leave-one-region-out cross-
313 validation (LORO) using the top three performing models from Phase 1 (Prithvi, DOFA, and U-
314 Net). In each iteration, the model was trained on all RGI regions except one, which was held out
315 for testing. We treat LORO as the primary evaluation because it directly tests global transferability
316 and is robust to the spatial autocorrelation that inflates random-split estimates. To ensure
317 consistency across experiments, we fixed the random seed to 42, learning rate 1e-5 with stepLR
318 scheduler, focal loss function, batch size of 16 and 100 epochs. We opted for a full fine-tuning
319 strategy to leverage the full potential offered by GFMs.

320 ***3.3. Evaluation metrics***

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

The model's performance was assessed using a widely used mean Intersection over Union (mIoU) and F1 score (Dice coefficient), and Boundary F1 (BF1) metrics. Mean IoU quantifies the overlap between predicted and ground truth segments for each class, divided by their union, equation (1).

$$mIoU = \frac{1}{k}\sum_{i=1}^{k}\frac{TP}{TP+FP+FN} \quad (1)$$

The F1 Score measures the harmonic mean of precision and recall, reflecting balances between false positives and false negatives, which is particularly informative for imbalanced datasets, such as glacial lakes, that occupy a small proportion of each image.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2)$$

Given the high importance of accurately delineating glacial lake boundaries, we also computed the Boundary F1 (BF1) score, which measures the spatial alignment between predicted and ground-truth boundaries by allowing a tolerance of two pixels.

$$BF1 = \frac{2.|B_P \cap B_G^+(t)|.|B_G \cap B_P^+(t)|}{|B_P|.|B_G|.(\frac{|B_P \cap B_G^+(t)|}{|B_P|}+\frac{|B_G \cap B_P^+(t)|}{|B_G|})} \quad (3)$$

Where, TP = True positive, FP = False positive, B_G = Ground truth boundary, B_P = Predicted boundary, B_G^+ = Dilated ground truth boundary, B_P^+ = Dilated predicted boundary, t = tolerance (2 pixels in our case)

4. Results

4.1. Glacial Lake Bench

Our results from spatially distributed test sites sampled across the globe (Fig. 3) establish Prithvi as the strongest baseline model, achieving a mIoU of 0.85 (std 0.16) and a Dice score of 0.91, outperforming widely used deep learning models for glacial lake segmentation such as U-Net (0.82) and DeepLabv3+ (0.80) (Table 3). Other GFM, DOFA (0.82), also showed competitive performance compared to Prithvi with a drop of 3 percentage points (pp) (Table 3). The distribution of mIoU scores across images reaffirms the stable performance of Prithvi, as indicated by the lowest standard deviation (0.16) compared to U-Net (0.18), highlighting the model's capability to generalize across heterogeneous conditions (Fig. 6a). Boundary accuracy, measured using BF1,

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

also strengthens Prithvi's superiority, achieving a score of 0.65 (std 0.35), the next best model, DOFA, exhibits a drop of 5 pp (Table 3). In comparison U-Net achieved 0.59 (std 0.37) and DeepLabv3+ 0.55 (std 0.36) exhibiting sharp drop of 10 pp in mean boundary F1 (Table 3).

Table 3. Results of spatially distributed testing sites of Glacial Lake-Bench across the globe using different deep learning models.

Glacial Lake-Bench			
Deep Learning Models	**mIoU (std)**	**Dice Score (F1)**	**Boundary F1 (std)**
U-Net	0.82 (0.18)	0.90	0.59 (0.37)
DeepLabv3+	0.80 (0.17)	0.88	0.55 (0.36)
DOFA	0.82 (0.16)	0.89	0.60 (0.36)
Prithvi	**0.85 (0.16)**	**0.91**	**0.65 (0.35)**

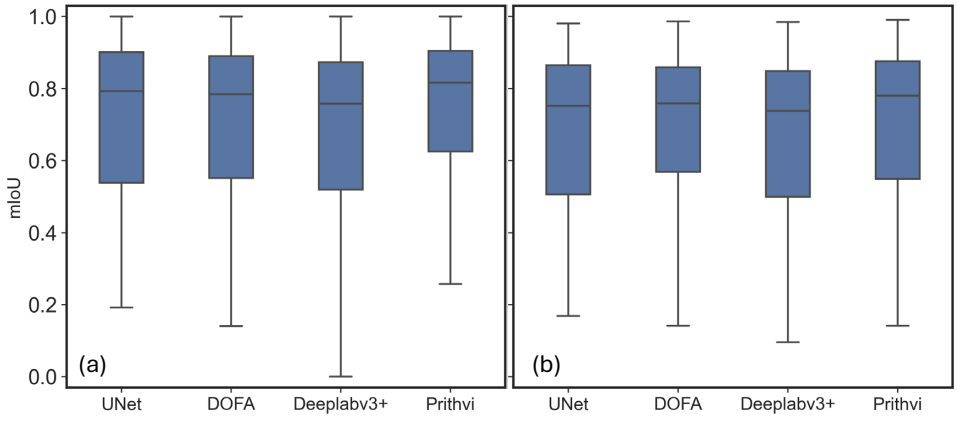


Fig. 6 Comparative evaluation of model performance using mIoU distribution over two datasets: (a) Glacial-Lake-Bench (GLB) and (b) Glacial-Lake-Bench-Challenge (GLBC). On GLB, Prithvi GFM outperformed all other models. On GLBC, which includes more challenging conditions, Prithvi and DOFA GFMs showed comparable performance, effectively tying in their results.

We also evaluated additional statistical measures using the Coefficient of Determination (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Percentage Error (MPE). These metrics, computed on an absolute areal basis, provide insight into each model's ability to capture lake-area variability. Our baseline models reveal slight gains by Prithvi GFM

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

compared to other models, achieving an R^2 of 0.98, MAE of 0.04 km², MSE of 0.01 km², and MPE of -3.3%. These results indicate strong agreement with ground truth, with only a modest underprediction (Fig. 7). In comparison, DOFA GFM achieved an R^2 of 0.96, a slightly higher MAE of 0.05 km², MSE of 0.02 km², and a slightly lower underprediction reflected in MPE of -2.43% (Fig. 7). U-Net exhibited a similar R^2 of 0.96 but a greater average underestimation with an MPE of -4.97%. These results were obtained after removing outliers (z-score > 3), leading to the exclusion of 24 observations from 1,907 samples. U-Net required removal of 41 outliers, indicating higher variability in its predictions.

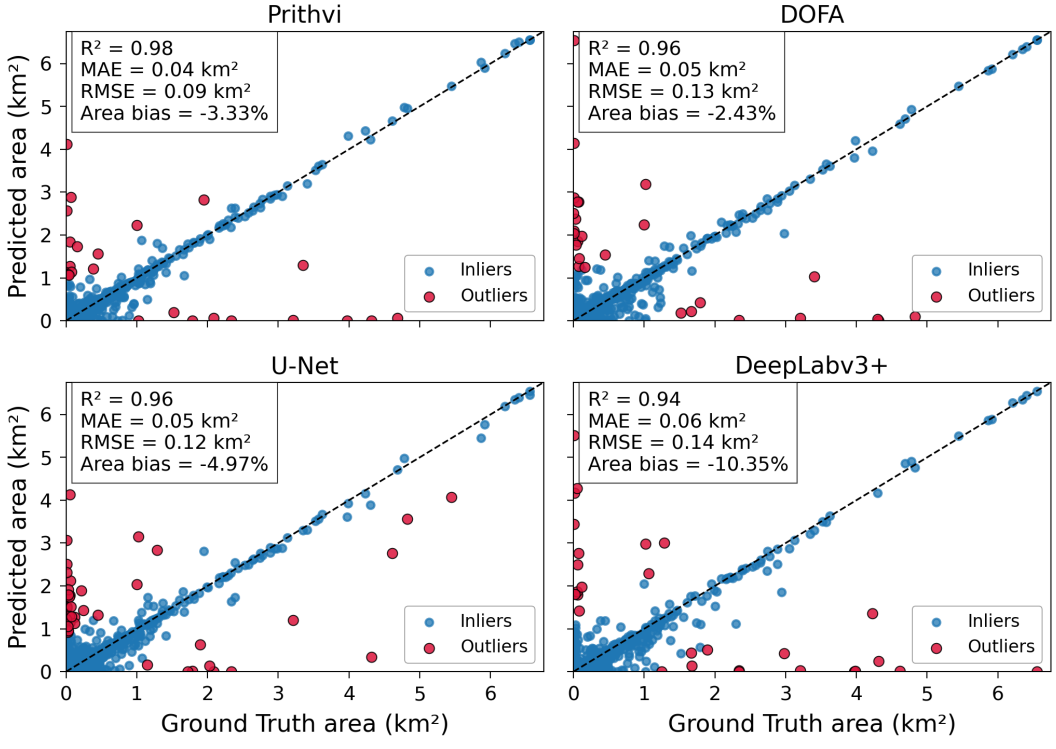


Fig. 7 Comparative evaluation of model performance over Glacial-Lake-Bench using Coefficient of Determination (R^2): indicating model's ability to explain variability in the ground truth data. Mean Absolute Error (MAE): representing the average deviation of the model's predictions from the ground truth, Mean Squared Error (MSE): reflects the average squared deviation, penalizing larger errors and Mean Percentage Error (MPE): provides an estimate of the model's average percentage error.

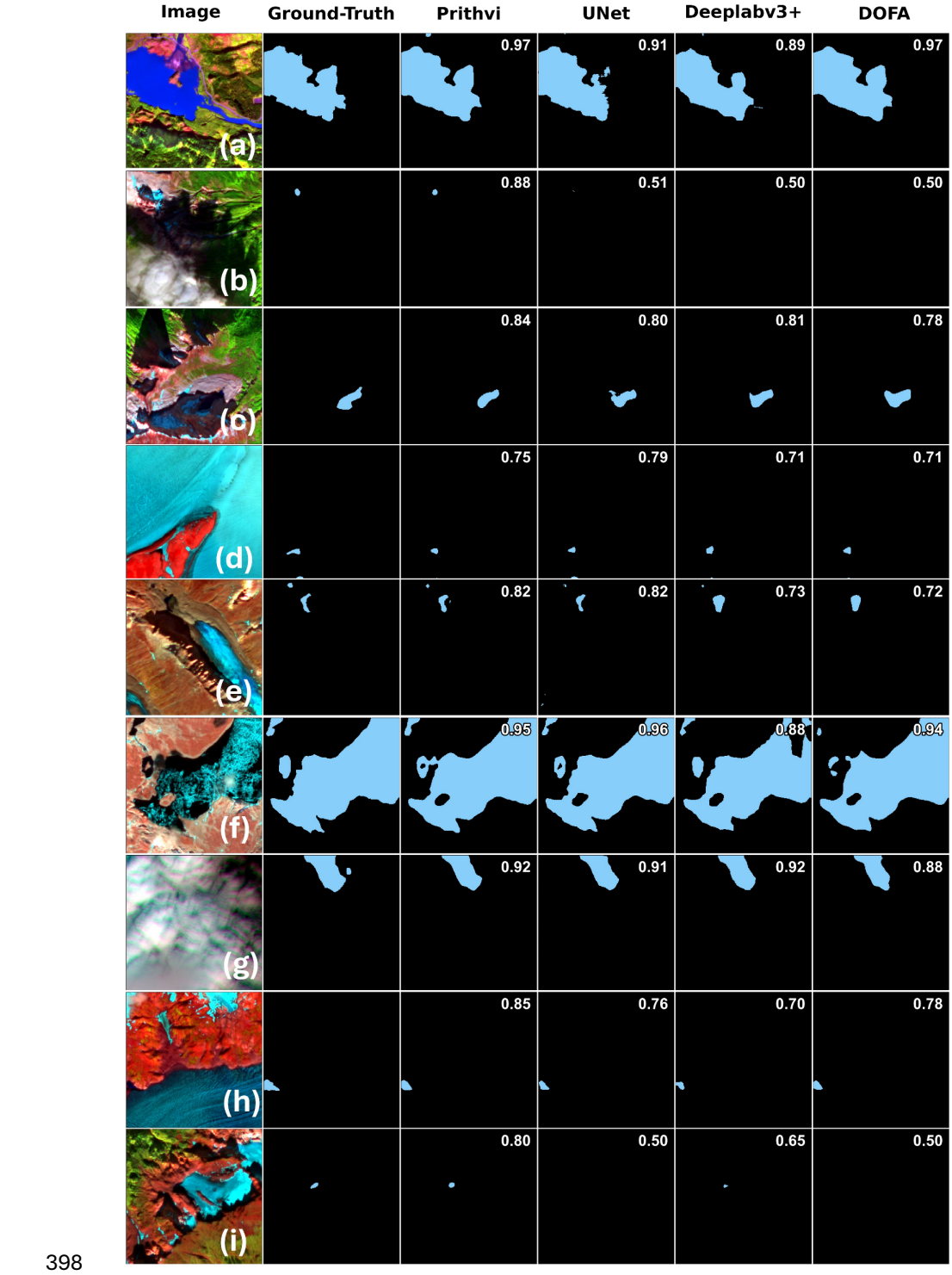
https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

377

378 The visual assessment reaffirms that Prithvi GFM continues to be the strongest baseline
379 model in mapping glacial lakes under the heterogeneous scene conditions, where many previous
380 approaches have struggled (Hu et al., 2024; Kaushik et al., 2022; Tang et al., 2024; Wang et al.,
381 2022). We present the following examples to demonstrate our results: (1) Streams vs lakes spectral
382 confusion, Fig. 8a shows a glacial lake with an outflowing stream, a scenario often misclassified
383 in earlier studies due to similar spectral signatures. With multimodal data, all models correctly
384 distinguished lake from stream; however, Prithvi and DOFA achieved the highest mIoU (0.97),
385 while U-Net (0.93) and DeepLabv3+ (0.89) performed less accurately. (2) Shadow-induced
386 challenges-cloud shadow (Fig. 8b) and mountain shadow (Fig. 8c) show that many models failed
387 to detect small glacial lakes (≤0.05 km²) obscured by shadows, one of the main limitations reported
388 in earlier literature. In contrast, Prithvi consistently detected these lakes. (3) Frozen lakes –
389 mapping frozen lakes remains one of the most persistent challenges (Figs. 8d, 8e, 8h, and 8i). In
390 all such cases, Prithvi consistently outperformed other models, particularly in detecting small
391 frozen lakes (<0.05 km²) under cloudy or shadowed conditions. (4) Dense cloud cover – The most
392 difficult scenario remains dense cloud cover. However, the integration of optical, SAR, and
393 topographic data has enabled accurate lake mapping for many models. In Fig. 8g, Prithvi achieved
394 mIoU of 0.92, matched only by DeepLabv3+, while DOFA lagged slightly with 0.88. (5) Mixed
395 spectral signals (partially frozen lakes or ice-covered surfaces), Fig. 8f, Prithvi, DOFA, and U-Net
396 all performed well, achieving mIoU of 0.95, 0.94, and 0.96, respectively.

397

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.



https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

Fig. 8 Visual assessment of model performance over the Glacial-Lake-Bench (GLB) dataset. These examples illustrate challenging scenarios where previous methods have struggled: (a) running water from a lake, (b) cloud shadow, (c) mountain shadow, (d), (e), (h), and (i) frozen lakes and very small lakes (0.002–0.05km²), (f) mixed spectral signal due to a partially frozen lake, and (g) fully cloudy conditions. Numbers on the top right of each image show mIoU.

4.2. Glacial Lake-Bench Challenge

The results obtained over the Glacial Lake-Bench Challenge (GLBC) show a noticeable decrease in model performance due to the high complexity of the task, indicating difficulty in the dataset. Under the most challenging conditions, Prithvi and DOFA GFMs performed competitively, both achieving an mIoU of 0.79. However, Prithvi slightly outperformed DOFA in terms of BF1, scoring 3% improvement over DOFA (Table 4). The next best-performing model was U-Net, which achieved an mIoU of 0.76 and a BF1 score of 0.50, close to DOFA's 0.51 (Table 4).

The distribution of mIoU per image also shows a slight improvement by Prithvi and DOFA (Figure 6b), reaffirming their relative robustness in handling complex glacial lake mapping scenarios. The evaluation of additional statistical parameters based on direct areal extent reveals that Prithvi and DOFA GFMs perform comparably, both achieving an R^2 of 0.91, MAE of 0.07 km², and MSE of 0.02 km². However, DOFA shows a slightly higher underestimation of 2.13%, compared to Prithvi (Figure 9). In comparison, the next best-performing model, U-Net, achieved an R^2 of 0.85, MAE of 0.08 km², MSE of 0.03 km², and a slightly higher underestimation of 3%.

Table 4. Comparative evaluation of model performance on Glacial-Lake-Bench-challenge.

Glacial Lake-Bench Challenge			
Model	**mIoU (std)**	**Dice Score (F1)**	**Boundary F1 (std)**
U-Net	0.76 (0.17)	0.72	0.50 (0.34)
DOFA	**0.79 (0.16)**	0.75	0.51 (0.31)
DeepLabv3+	0.74 (0.17)	0.67	0.47 (0.34)
Prithvi	**0.79 (0.17)**	**0.76**	**0.54 (0.33)**

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

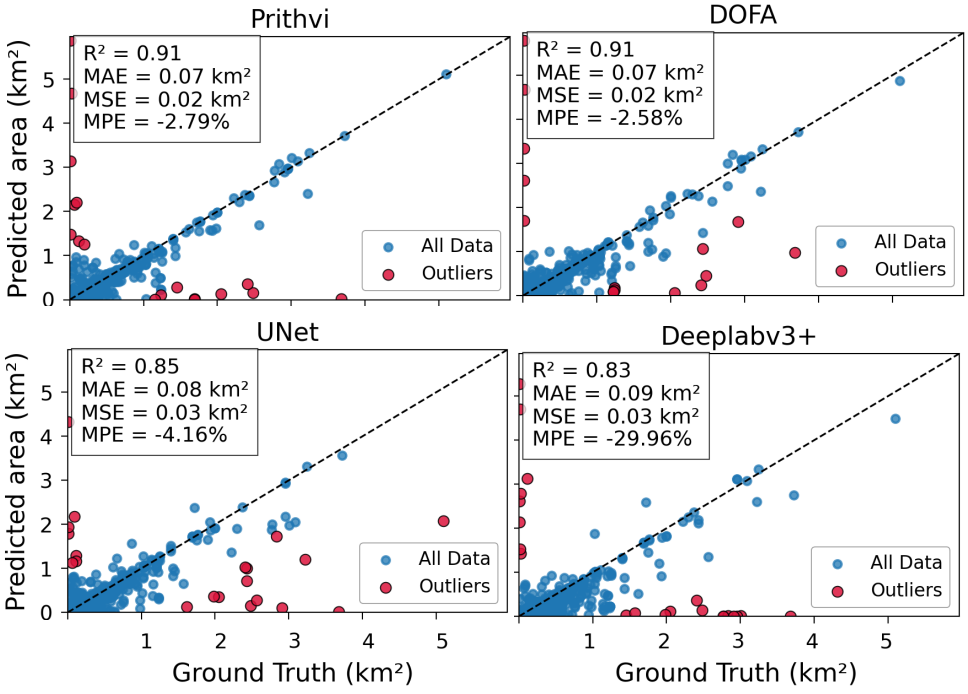


Fig. 9 Comparative evaluation of model performance over Glacial-Lake-Bench-Challenge using Coefficient of Determination (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Percentage Error (MPE).

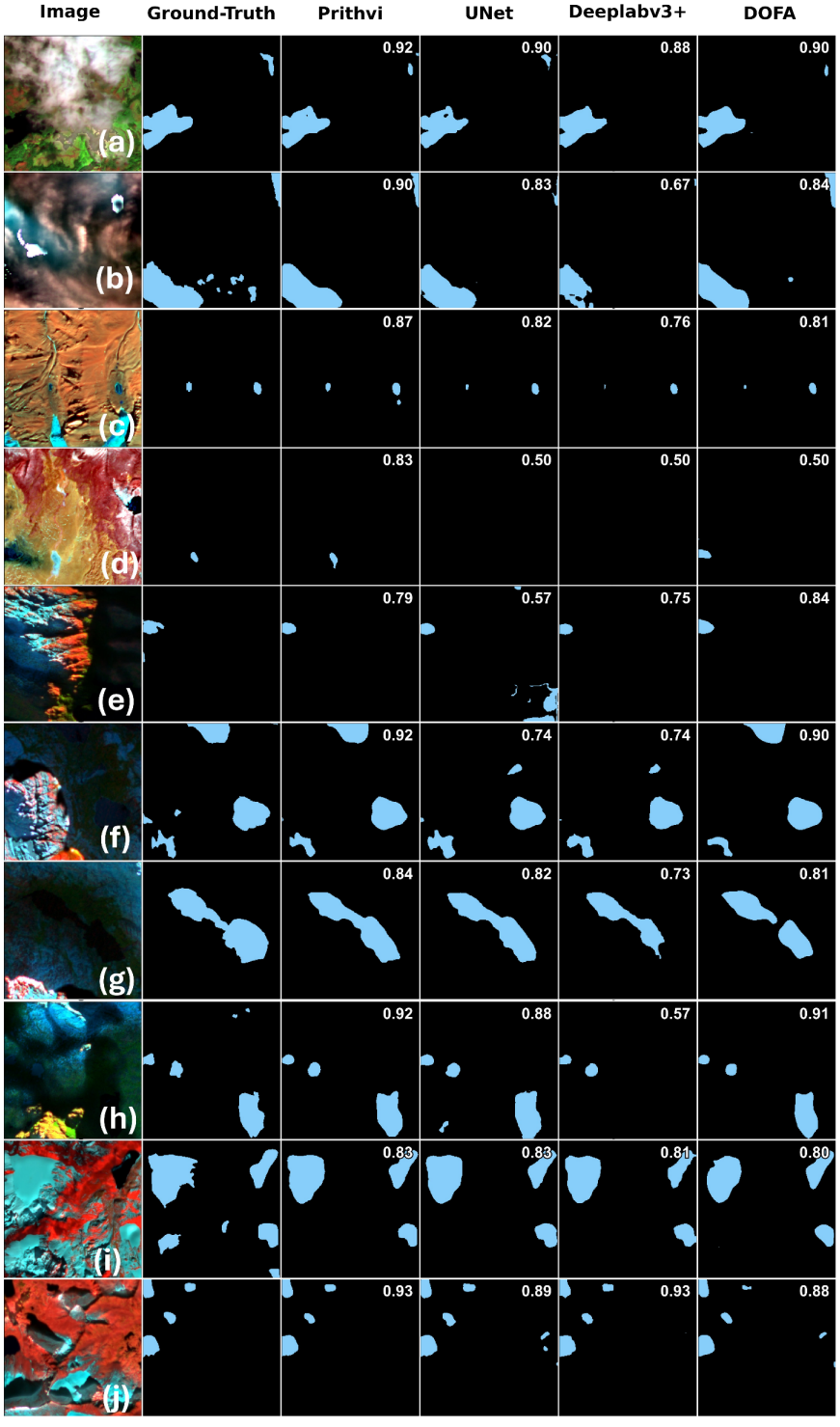
The visual analysis further reinforces the moderate performance gains of Prithvi and DOFA GFMs compared to CNN models (Figure 10). We show the following examples to demonstrate the consistent performance of Prithvi: (1) cloudy conditions- Prithvi achieved mIoU scores of 0.92 (Fig. 10a) and 0.90 (Fig. 10b), while DOFA showed slightly lower performance (by 2–6%), respectively. Other models also performed competitively U-Net achieved 0.90, and DeepLabv3+ reached 0.88 (Fig. 10a). (2) small lakes (≤0.05 km²)- Prithvi showed consistent performance with an mIoU of 0.87, outperforming by DOFA by 6% and U-Net by 5%. In another example of a small lake (Fig. 10d), Prithvi successfully detected the lake, while all other models failed. (3) shadow- model performance remained generally satisfactory. For example, in Fig. 10e, DOFA outperformed Prithvi with an mIoU of 0.84 compared to Prithvi's 0.79, while U-Net and DeepLabv3+ scored 0.57 and 0.75, respectively. In three additional examples (Figs. 10f, 10g, and 10h), Prithvi slightly outperformed other models, achieving mIoU scores of 0.92, 0.84, and 0.92, respectively, compared to DOFA's 0.90, 0.81, and 0.91. For frozen lakes, Prithvi continued to outperform other models.

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

In Fig. 10i, Prithvi and U-Net performed equally well with mIoU scores of 0.83, followed by DOFA at 0.80. A similar observation is seen in Fig. 10j, where Prithvi showed competitive performance with mIoU scores of 0.93 and 0.92, respectively, while DOFA scored 0.88. Overall, these results demonstrate significant advancements in mapping glacial lakes under the most challenging conditions where existing deep learning methods often struggle (Hu et al., 2024; Kaushik et al., 2022; Tang et al., 2024; Wang et al., 2022).

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions



445

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

Fig. 10 Example of successful implementation of Prithvi and DOFA, trained on the Glacial Lake-Bench dataset for mapping lakes under manually curated challenging conditions (i.e., Glacial Lake-Challenge dataset). These conditions include cloud cover (panels a,b), frozen lakes (panels i and j), shadows (panels e-h), and very small lake (0.002–0.01km²: panel c and d) representing a highly imbalanced class.

4.3.Leave one-region-out-cross validation

In the leave-one-region-out cross-validation experiment, Prithvi showed a modest performance advantage, achieving an average mIoU of 0.83, compared to DOFA's 0.80 and U-Net's 0.79 (Table 5). The lowest performance of all three models occurred in Svalbard, where Prithvi achieved 0.66, whereas DOFA and U-Net both achieved mIoU of 0.64. Other regions where the models showed reduced performance included Greenland (Prithvi: 0.81, DOFA: 0.78), Arctic Canada North (Prithvi: 0.70, DOFA: 0.71), and the Southern Andes (Prithvi: 0.75, DOFA: 0.71) (Table 5). In contrast, all three models performed well in regions such as Alaska (Prithvi: 0.83, DOFA: 0.84), Iceland (Prithvi: 0.88, DOFA: 0.86), and North Asia (Prithvi: 0.88, DOFA: 0.87) (Table 5). Prithvi also outperformed DOFA in the Russian Arctic (0.81 vs. 0.73) and South Asia East (0.87 vs. 0.84) (Table 5). Overall, these results demonstrate strong spatial transferability on a global scale. It also confirms that the performance observed in previous experiments is not driven by spatial autocorrelation but rather reflects the robustness of the GLB dataset and the capability of the baseline models for global-scale applications. We would like to highlight that Prithvi and DOFA are pretrained through self-supervised masked image reconstruction and never observe lake labels, which we generated independently. Any residual scene-level overlap is further bound by our from-scratch CNN baselines, which carry no pretraining exposure yet remain within 3 to 4 percentage points of the best GFM on both the random split and the LORO evaluation. The modest GFM gains therefore reflect representation quality rather than memorized test scenes.

Table 5. Results of leave-one-region-out cross validation over Glacial Lake Bench across the RGI regions using DOFA, Prithvi, and U-Net.

RGI Region	DOFA	Prithvi	U-Net
Alaska	0.84	0.83	0.81
Iceland	0.86	0.88	0.72

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

New Zealand	0.83	0.88	0.70
North Asia	0.87	0.88	0.84
Scandinavia	0.83	0.85	0.85
Svalbard	0.64	0.66	0.64
Arctic Canada North	0.71	0.70	0.68
Arctic Canada South	0.83	0.84	0.86
Central Asia	0.76	0.77	0.76
Central Europe	0.90	0.92	0.91
Greenland	0.78	0.81	0.82
Low Latitudes	0.86	0.88	0.86
Middle East *(Caucasus)*	0.84	0.86	0.87
Russian Arctic	0.73	0.81	0.69
Southern Andes	0.71	0.75	0.68
South Asia East	0.84	0.87	0.88
South Asia West	0.82	0.82	0.82
Western Canada and USA	0.81	0.82	0.87
Mean	0.80 (0.06)	**0.83 (0.06)**	0.79

5. Discussion

5.1. Dataset utility and benchmarking

Our compiled datasets, GLB and GLBC, along with recently developed GFMs that leverage pretrained encoders, offer an unprecedented opportunity to advance glacial lake mapping at a global scale and under challenging conditions. We conducted a systematic evaluation of several deep learning models, including CNNs (U-Net and DeepLabv3+), and Vision Transformer based and GFMs (DOFA, and Prithvi).

In the visual analysis (Fig. 8 and Fig. 10), we placed special emphasis on diverse and challenging conditions that have historically limited the effectiveness of state-of-the-art glacial lake mapping methods (Hu et al., 2024; Kaushik et al., 2022; Tang et al., 2024; Tom et al., 2025). Our results demonstrate stable performance over heterogeneous test images representing glacial lake in diverse geographies from clean lake to most challenging conditions. Figure 11 provides detailed visualizations in which the best-performing model (Prithvi) shows consistently accurate segmentation in clean lake conditions (Fig. 11a and b). Figure 11c illustrates dense cloud cover,

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

yet the compiled dataset and proposed model accurately map the lake extent. We would like to highlight that this advancement in mapping capabilities stems from 1) incorporation of SAR as input data and 2) Prithvi's pretrained temporal ViT which improves model's generalizability. For example, when visualizing all bands in Figure 11c (see Supplementary Figure S1), optical bands fail to capture the lake extent, while SAR bands clearly distinguish water from non-water pixels. This complementary information, paired with systematic model selection, enables a comprehensive and robust approach to global glacial lake mapping. Figures 11d and 11e represent conditions involving ice over lake surfaces, which typically produce mixed spectral signals. Figures 11f and 11g show successful detection of very small lakes (0.002–0.01 km²), while Figure 11h presents a frozen lake with spectral characteristics similar to glacial ice, yet the model effectively distinguishes between the two.

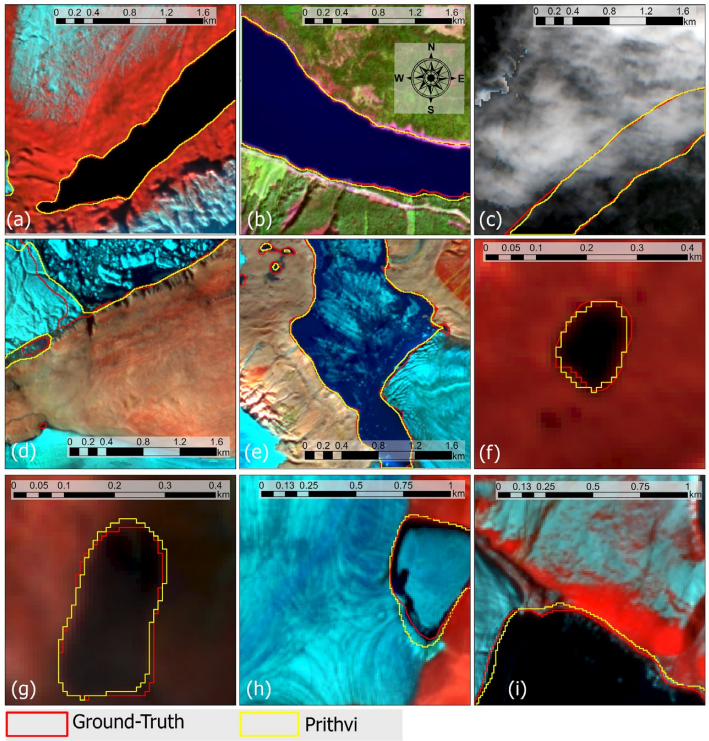


Fig. 11 Snapshots of successful implementation of Prithvi GFM over the Glacial Lake Bench (GLB) dataset under diverse challenging conditions: (a) and (b) clean lakes, (c) dense cloud cover, (d) and (e) full and partially covered with ice, (f) and (g) very small lakes (0.002–0.01 km²), (h) frozen lake.

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

These advances are significant, as existing methods (Kaushik et al., 2022; Ma et al., 2025; Tang et al., 2024; Wang et al., 2022) consistently underperform in such conditions. However, direct comparisons of results are not possible since none of the existing studies compiled a readily available evaluation dataset. Therefore, we establish a community standard for assessing any deep learning model for global glacial lake mapping using GLB, following the same split and leave-one-region-out cross-validation. This approach reduces the high impact of spatial bias, which has been largely overlooked in previous studies. If we compare directly using the Dice score (F1), since it is the most widely reported evaluation metric, we achieved a score of 0.91 over GLB test data. In comparison, Kaushik et al. (2022) reported an average F1 score of 0.84 across four test sites in the Himalayas, Tang et al. (2024b), reported 0.85 for the entire HMA, Wang et al. (2022) reported 0.81 for the Hindu Kush-Himalaya, Ma et al. (2025) reported 0.86 for selected sites in the Himalayas, and Dirscherl et al. (2021) reported a relatively high F1 score of 0.96 for mapping glacial lakes in Antarctica. It is important to note that all these studies were conducted on regional scales, whereas our work is at a global scale, demonstrating the model's capability to generalize globally.

5.2. Application of Glacial-Lake-Bench and Glacial-Lake-Challenge

Given the importance of glacial lakes in understanding climate change, glacier mass wastage, flow velocity, associated GLOF risks, and potential freshwater resources, automated glacial lake mapping has gained wide scientific attention in recent years (Hu et al., 2024; Kaushik et al., 2022; Ma et al., 2025; Tang et al., 2024; Wang et al., 2022). However, most existing methods have focused on regional scales, like High Mountain Asia (Kaushik et al., 2022; Ma et al., 2025; Tang et al., 2024; Wang et al., 2022) or Antarctica (Dirscherl et al., 2021). Notably, none of the previous studies have accounted for spatial autocorrelation while evaluating their models, as testing sites were often selected adjacent to training sites. In this regard, Ma et al. (2025) attempted to compile a Glacial Lake Image dataset; however, the dataset was limited to only parts of the Himalayas and included only optical data, which has shown severe limitations under challenging conditions. Prior work has explicitly recommended globally sampled, multisource remote sensing datasets to support reproducible glacial lake research (Kaushik et al., 2022; Ma et al., 2025).

To address these limitations, our compiled GLB dataset provides a globally distributed, multi-sensor benchmark that enables fair evaluation of newly developed models using a consistent train-

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.


test split and leave-one-region-out cross-validation. This framework reduces spatial bias and allows models to be compared at both regional and global scales. In addition, GLBC, a curated subset, of 1,105 image-label pairs representing challenging conditions, offers a dedicated benchmark for evaluating model robustness in scenarios where traditional methods frequently fail. GLBC therefore provides a standardized and objective means of assessing model applicability under the most demanding real-world conditions. Overall, these datasets are expected to accelerate progress in the glacial lake mapping by bridging long-standing gaps in data availability and providing scientific communities with a consistent evaluation platform for developing and testing deep learning models. Furthermore, this work establishes a community benchmark that can accompany future advances in GFMs. Especially given the rapid development of GFMs, where cryosphere components have not yet been evaluated, we expect the inclusion of cryosphere tasks to become standard when assessing downstream capabilities of these models. The extension of GLB and GLBC datasets will focus on incorporating multi-temporal datasets to assess the GFMs in change detection tasks and quantify rate of lake changes fast and efficiently.

5.3. Usage notes, limitations, and recommendations

We make the following recommendations for use. First, report LORO performance as the primary measure of global transferability, because random splits are subject to spatial autocorrelation between adjacent training and test sites, an effect that has been largely overlooked in earlier work. Second, GLBC should be used to evaluate model capabilities under challenging conditions. Third, account for the inherited label uncertainty of roughly ±3 % in lake area (Zhang et al., 2024), and our independent evaluation reveals ~4% errors of omission in dataset emerging directly through missed lakes in the dataset.

The dataset has known limitations that need to be considered in model comparison. Performance is expected to be lowest for very small lakes (0.002–0.01 km²) under dense cloud (Fig. 12). To further understand this, we visualized each band for examples containing multiple lakes of different sizes beneath cloud cover (Fig. S2). Because optical data provides no usable information under such conditions, lake detection relies entirely on SAR backscatter. For very small lakes, SAR signals are often indistinguishable from surrounding terrain due to speckle, sensor noise, and the lake footprint approaching the spatial resolution of Sentinel-1. In contrast, separation improves for lakes ≥0.01 km². Consequently, our dataset is effective for detecting very

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

small lakes under cloud-free conditions but exhibits reduced accuracy when clouds obscure optical information and SAR alone become insufficient.

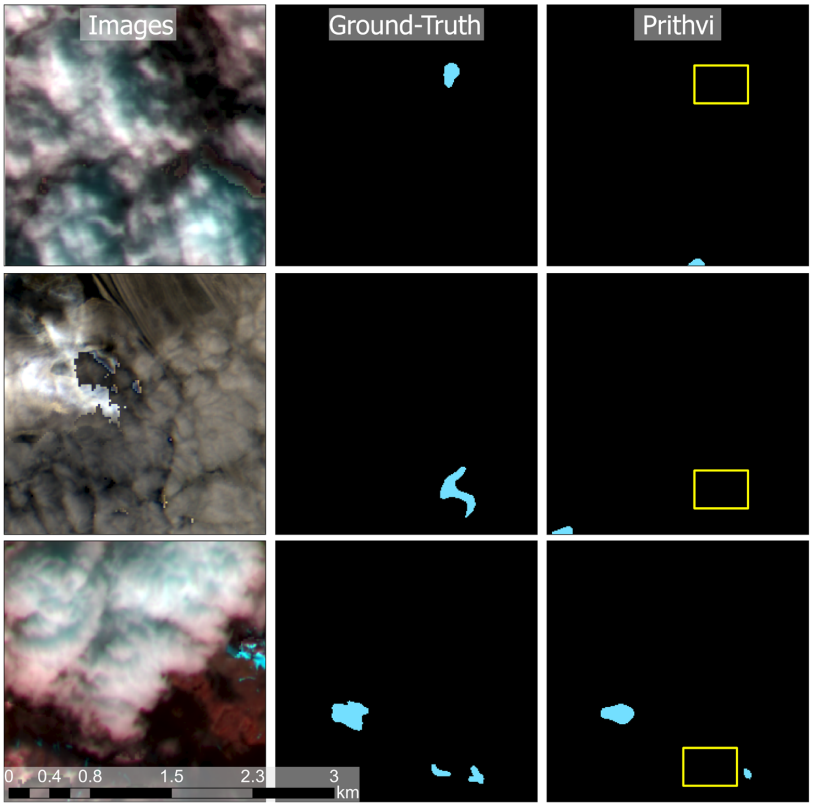


Fig. 12 Our proposed method is prone to miss-classification on very small lakes (0.002–0.01 km²) under cloud cover. Yellow boxes show false negative classification.

Our detailed assessment of dataset with leading Prithvi model confirm significant decline in baseline mIoU (0.70) for small lakes (≤0.05 km²) (Fig. 13a). Comparatively, lakes ranging between 0.05–0.5 km² show performance increases to 0.82 (std 0.14) mIoU. Our LORO evaluation shows decreased model performance in regions such as Southern Andes (mIoU: 0.75, Table 5), Central Asia (mIoU: 0.77, Table 5), and Svalbard (mIoU: 0.66, Table 5). To assess whether model performance in random split follows a similar pattern, we carried out a detailed analysis across each RGI region. Our analysis revealed similar results in Southern Andes, where dataset shows the lowest performance with an mIoU of 0.70 (std 0.19), and Fig. 13b shows a concentration of many

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

observations around mIoU of 0.6. Similarly, the model underperforms in Arctic Canada North with an mIoU of 0.70 (std 0.18), consistent with the LORO evaluation (Table 5). Finally, GLB is temporally limited as we only sampled the dataset for ablation season of 2020 to match the label boundaries. We expect model's temporal transferability within the same season where seasonal snow is at minimum and limited transferability in seasonal change analysis. To extend the presented methodological framework for seasonal analysis we encourage inclusion of sample from different seasons. While evaluating the result please be cautious of inherited ±3% uncertainty in source label dataset (Zhang et al., 2024).

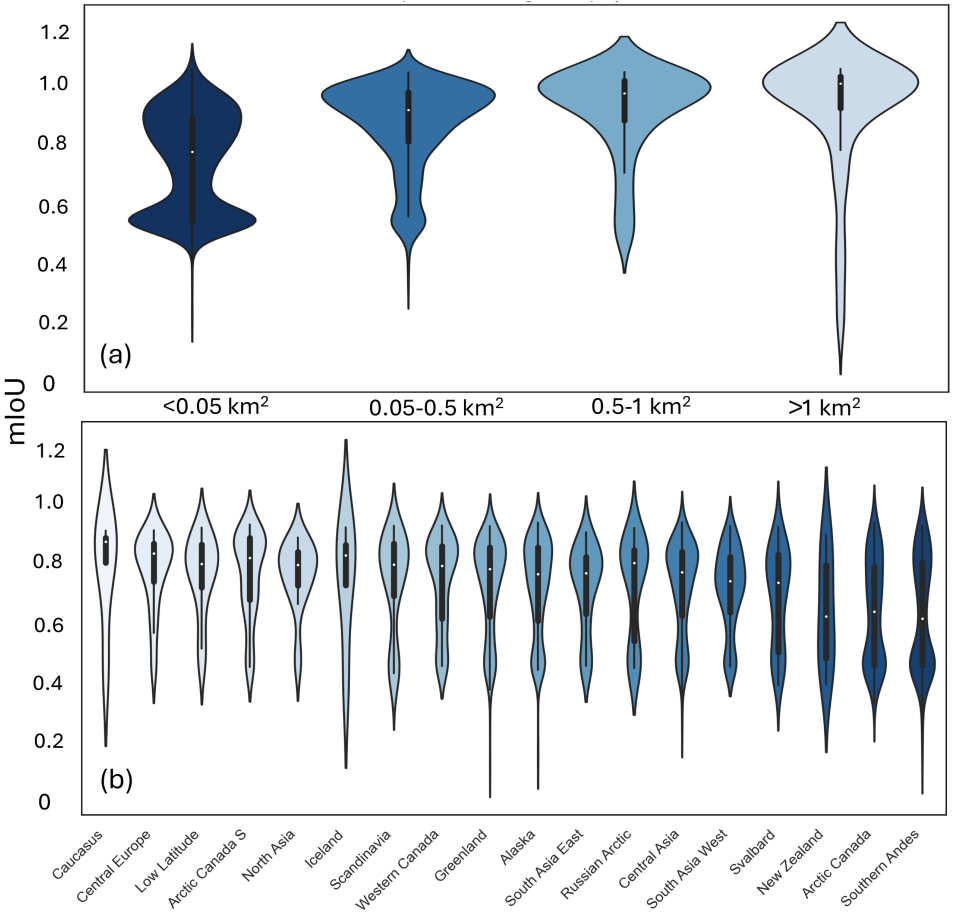


Fig. 13 Evaluation of model's performance for (a) different lake sizes and (b) across geographies.

6. Conclusion

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

589 Fast and efficient glacial lake mapping is of utmost importance to monitor glacial lakes at
590 unprecedented spatial and temporal scales. The unavailability of readily accessible multisource
591 datasets hampers the development and evaluation of deep learning models for globally scalable
592 glacial lake mapping. To address this, we present Glacial-Lake-Bench, a multisource remote
593 sensing dataset consisting of 19,115 image-label pairs sampled across each RGI region. Each
594 image is composed of Sentinel-2, Sentinel-1, slope, and elevation data. This dataset also includes
595 a subset of 1,105 image-label pairs captured under challenging conditions. As the first of its kind,
596 this dataset provides a community standard for evaluating newly proposed deep learning models
597 and promotes fair and rigorous comparison across different approaches. Our analysis reveals ~4%
598 underestimation of lake pixels in GLB. We explicitly mention dataset limitations and recommend
599 usage under diverse evaluation protocol. We used this dataset to train four deep learning models,
600 including U-Net, DeepLabv3+, DOFA, and Prithvi. We provide strong baseline results for future
601 model development and fair comparison. We expect widespread application of the GLB dataset
602 for assessing newly proposed foundational models on one of the most imperative cryosphere
603 features, which is currently completely missing from the downstream task evaluations of recent
604 models.

605 **Data Availability Statement**

606 The generated Glacial-Lake-Bench (GLB) and Glacial-Lake-Bench-Challenge (GLBC) datasets
607 are available at https://zenodo.org/records/17917359 (Kaushik, 2026) under Creative Commons
608 Attribution 4.0 International license. Sentinel-1 and Sentinel-2 data can be downloaded from
609 https://dataspace.copernicus.eu/explore-data/data-collections/sentinel-data/sentinel-2, Prithvi
610 pretrained weights can be found at https://huggingface.co/ibm-nasa-geospatial/Prithvi-EO-2.0-
611 600M

612 **Code Availability**

613 Workflow to generate GLB data can be found at https://github.com/Sk-2103/dl4eo.

614 **Author Contributions**

615 SK: Conceptualization, Data Curation, Software, Methodology, Formal Analysis, Visualization,
616 Writing-Original draft - Review & Editing BT: Conceptualization, Resources, Writing - Review &

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

Editing, IH: Conceptualization, Writing - Review & Editing UKH: Conceptualization, Writing - Review & Editing

References

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. https://doi.org/10.48550/ARXIV.1802.02611

Dirscherl, M., Dietz, A.J., Kneisel, C., Kuenzer, C., 2021. A Novel Method for Automated Supraglacial Lake Mapping in Antarctica Using Sentinel-1 SAR Imagery and Deep Learning. Remote Sensing 13, 197. https://doi.org/10.3390/rs13020197

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/ARXIV.2010.11929

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2021. Masked Autoencoders Are Scalable Vision Learners. https://doi.org/10.48550/ARXIV.2111.06377

Hu, J., Zhang, T., Zhou, X., Yi, G., Bie, X., Li, J., Chen, Y., Lai, P., 2024. A Glacial Lake Mapping Framework in High Mountain Areas: A Case Study of the Southeastern Tibetan Plateau. IEEE Trans. Geosci. Remote Sensing 62, 1–12. https://doi.org/10.1109/TGRS.2023.3349281

Immerzeel, W.W., Lutz, A.F., Andrade, M., Bahl, A., Biemans, H., Bolch, T., Hyde, S., Brumby, S., Davies, B.J., Elmore, A.C., Emmer, A., Feng, M., Fernández, A., Haritashya, U., Kargel, J.S., Koppes, M., Kraaijenbrink, P.D.A., Kulkarni, A.V., Mayewski, P.A., Nepal, S., Pacheco, P., Painter, T.H., Pellicciotti, F., Rajaram, H., Rupper, S., Sinisalo, A., Shrestha, A.B., Viviroli, D., Wada, Y., Xiao, C., Yao, T., Baillie, J.E.M., 2020. Importance and vulnerability of the world's water towers. Nature 577, 364–369. https://doi.org/10.1038/s41586-019-1822-y

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

643 Jiang, D., Li, S., Hajnsek, I., Siddique, M.A., Hong, W., Wu, Y., 2025. Glacial lake mapping using
644 remote sensing Geo-Foundation Model. International Journal of Applied Earth Observation
645 and Geoinformation 136, 104371. https://doi.org/10.1016/j.jag.2025.104371

646 Kaushik, S., Rafiq, M., Joshi, P.K., Singh, T., 2020. Examining the glacial lake dynamics in a
647 warming climate and GLOF modelling in parts of Chandra basin, Himachal Pradesh, India.
648 Science of The Total Environment 714, 136455.
649 https://doi.org/10.1016/j.scitotenv.2019.136455

650 Kaushik, S., Singh, T., Joshi, P.K., Dietz, A.J., 2022. Automated mapping of glacial lakes using
651 multisource remote sensing data and deep convolutional neural network. International
652 Journal of Applied Earth Observation and Geoinformation 115, 103085.
653 https://doi.org/10.1016/j.jag.2022.103085

654 Kaushik, S. Glacial-Lake-Bench: A Global Multi-Sensor Benchmark Dataset for Evaluating Deep
655 Learning Models for Glacial Lake Mapping https://zenodo.org/records/17917359, 2026

656 King, O., Bhattacharya, A., Bhambri, R., Bolch, T., 2019. Glacial lakes exacerbate Himalayan
657 glacier mass loss. Sci Rep 9, 18145. https://doi.org/10.1038/s41598-019-53733-x

658 Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E.D., Kerner, H., Lütjens, B., Irvin, J.A., Dao,
659 D., Alemohammad, H., Drouin, A., Gunturkun, M., Huang, G., Vazquez, D., Newman, D.,
660 Bengio, Y., Ermon, S., Zhu, X.X., 2023. GEO-Bench: Toward Foundation Models for Earth
661 Monitoring. https://doi.org/10.48550/ARXIV.2306.03831

662 Li, L., Long, D., Wang, Y., Woolway, R.I., 2025. Global dominance of seasonality in shaping lake-
663 surface-extent dynamics. Nature 642, 361–368. https://doi.org/10.1038/s41586-025-
664 09046-3

665 Ma, D., Li, J., Jiang, L., 2025. Efficient glacial lake mapping by leveraging deep transfer learning
666 and a new annotated glacial lake dataset. Journal of Hydrology 657, 133072.
667 https://doi.org/10.1016/j.jhydrol.2025.133072

668 Pi, X., Luo, Q., Feng, L., Xu, Y., Tang, J., Liang, X., Ma, E., Cheng, R., Fensholt, R., Brandt, M.,
669 Cai, X., Gibson, L., Liu, J., Zheng, C., Li, W., Bryan, B.A., 2022. Mapping global lake

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

dynamics reveals the emerging roles of small lakes. Nat Commun 13. https://doi.org/10.1038/s41467-022-33239-3

Pronk, J.B., Bolch, T., King, O., Wouters, B., Benn, D.I., 2021. Contrasting surface velocities between lake- and land-terminating glaciers in the Himalayan region. The Cryosphere 15, 5577–5599. https://doi.org/10.5194/tc-15-5577-2021

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

Szwarcman, D., Roy, S., Fraccaro, P., Gíslason, Þ.E., Blumenstiel, B., Ghosal, R., de Oliveira, P.H., Almeida, J.L. de S., Sedona, R., Kang, Y., Chakraborty, S., Wang, S., Gomes, C., Kumar, A., Truong, M., Godwin, D., Lee, H., Hsu, C.-Y., Asanjan, A.A., Mujeci, B., Shidham, D., Keenan, T., Arevalo, P., Li, W., Alemohammad, H., Olofsson, P., Hain, C., Kennedy, R., Zadrozny, B., Bell, D., Cavallaro, G., Watson, C., Maskey, M., Ramachandran, R., Moreno, J.B., 2024. Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications. https://doi.org/10.48550/ARXIV.2412.02732

Tang, Q., Zhang, G., Yao, T., Wieland, M., Liu, L., Kaushik, S., 2024a. Automatic extraction of glacial lakes from Landsat imagery using deep learning across the Third Pole region. Remote Sensing of Environment 315, 114413. https://doi.org/10.1016/j.rse.2024.114413

Taylor, C., Robinson, T.R., Dunning, S., Rachel Carr, J., Westoby, M., 2023. Glacial lake outburst floods threaten millions globally. Nat Commun 14, 487. https://doi.org/10.1038/s41467-023-36033-x

Tom, M., Odermatt, D., David, C.H., Cerbelaud, A., Wade, J., Frey, H., 2025. Monitoring earth's glacial lakes from space with machine learning. Science of Remote Sensing 12, 100277. https://doi.org/10.1016/j.srs.2025.100277

Wang, S., Peppa, M.V., Xiao, W., Maharjan, S.B., Joshi, S.P., Mills, J.P., 2022. A second-order attention network for glacial lake segmentation from remotely sensed imagery. ISPRS Journal of Photogrammetry and Remote Sensing 189, 289–301. https://doi.org/10.1016/j.isprsjprs.2022.05.007

https://doi.org/10.5194/essd-2026-474
Preprint. Discussion started: 9 July 2026
© Author(s) 2026. CC BY 4.0 License.

698 Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified Perceptual Parsing for Scene
699 Understanding. https://doi.org/10.48550/ARXIV.1807.10221

700 Xiong, Z., Wang, Y., Zhang, F., Stewart, A.J., Hanna, J., Borth, D., Papoutsis, I., Saux, B.L.,
701 Camps-Valls, G., Zhu, X.X., 2024. Neural Plasticity-Inspired Multimodal Foundation
702 Model for Earth Observation. https://doi.org/10.48550/arXiv.2403.15356

703 Zhang, G., Bolch, T., Yao, T., Rounce, D.R., Chen, W., Veh, G., King, O., Allen, S.K., Wang, M.,
704 Wang, W., 2023. Underestimated mass loss from lake-terminating glaciers in the greater
705 Himalaya. Nat. Geosci. 16, 333–338. https://doi.org/10.1038/s41561-023-01150-1

706 Zhang, T., Wang, W., An, B., 2024. Heterogeneous changes in global glacial lakes under coupled
707 climate warming and glacier thinning. Commun Earth Environ 5, 374.
708 https://doi.org/10.1038/s43247-024-01544-y

709 Zheng, G., Allen, S.K., Bao, A., Ballesteros-Cánovas, J.A., Huss, M., Zhang, G., Li, J., Yuan, Y.,
710 Jiang, L., Yu, T., Chen, W., Stoffel, M., 2021. Increasing risk of glacial lake outburst floods
711 from future Third Pole deglaciation. Nat. Clim. Chang. 11, 411–417.
712 https://doi.org/10.1038/s41558-021-01028-3

713

714

715