AGU ADVANCING EARTH AND SPACE SCIENCES

Earth and Space Science

RESEARCH ARTICLE
10.1029/2025EA004197

DL4GAM: A Multi-Modal Deep Learning-Based Framework for Glacier Area Monitoring, Trained and Validated on the European Alps

Codruț-Andrei Diaconu[1,2], **Harry Zekollari**[3,4], **and Jonathan L. Bamber**[1,5]

[1]School of Engineering and Design, Technical University of Munich, München, Germany, [2]Earth Observation Center, German Aerospace Center (DLR), Cologne, Germany, [3]Department of Water and Climate, Vrije Universiteit Brussel, Brussels, Belgium, [4]Laboratory of Hydraulics, Hydrology and Glaciology (VAW), ETH Zürich, Zürich, Switzerland, [5]Bristol Glaciology Centre, University of Bristol, Bristol, UK

Key Points:

- We introduce DL4GAM, a Deep Learning-based framework for Glacier Area Monitoring, leveraging ensemble learning and optimal image selection
- A curated glacier mapping data set for the European Alps and the DL4GAM predictions are released, to facilitate future research
- We estimate an Alpine-wide glacier area loss rate of about two percent per year over 2015–2023

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

C.-A. Diaconu,
codrut-andrei.diaconu@dlr.de

Citation:

Diaconu, C.-A., Zekollari, H., & Bamber, J. L. (2025). DL4GAM: A multi-modal deep learning-based framework for glacier area monitoring, trained and validated on the European Alps. *Earth and Space Science*, *12*, e2025EA004197. https://doi.org/10.1029/2025EA004197

Received 20 JAN 2025
Accepted 28 AUG 2025

Author Contributions:

Conceptualization: Codruț-Andrei Diaconu, Harry Zekollari, Jonathan L. Bamber
Data curation: Codruț-Andrei Diaconu
Formal analysis: Codruț-Andrei Diaconu
Methodology: Codruț-Andrei Diaconu, Harry Zekollari, Jonathan L. Bamber
Software: Codruț-Andrei Diaconu
Supervision: Harry Zekollari, Jonathan L. Bamber
Writing – original draft: Codruț-Andrei Diaconu, Harry Zekollari, Jonathan L. Bamber
Writing – review & editing: Codruț-Andrei Diaconu, Harry Zekollari, Jonathan L. Bamber

Abstract Glaciers play a critical role in our society, impacting everything from sea-level rise and access to clean water to the tourism industry. Their accelerated melt represents a key indicator of the changing climate, highlighting the need for efficient monitoring techniques. The traditional way of assessing glacier area change is by rebuilding glacier inventories. This often relies on manual correction of semi-automated outputs from satellite imagery, which is time-consuming and susceptible to human biases. However, recent advancements in Deep Learning have enabled significant progress toward fully automatic glacier mapping. In this work, we introduce DL4GAM: a multi-modal Deep Learning-based framework for Glacier Area Monitoring, available open-source. It includes uncertainty quantification through ensemble learning and a procedure to identify the imagery with the best mapping conditions independently for each glacier. DL4GAM is trained and evaluated on the European Alps, a region for which experts estimated an annual change rate of around −1.3% over 2003–2015. We use DL4GAM to investigate the glacier evolution from 2015 to 2023 using Sentinel-2 imagery and elevation (change) maps. By employing geographic cross-validation, our models, based on U-Net ensembles, demonstrate strong generalization capabilities. We then apply the models on 2023 data and estimate the area change at both the glacier and regional levels. Regionally, we estimate an area change rate of $-1.90 \pm 1.26\%$ per year. We provide quality-controlled individual estimates over 2015–2023 for about 900 glaciers, covering around 70% of the region. Debris-covered regions remain the most uncertain.

Plain Language Summary Glacier melt is accelerating. To keep track of the evolution of glacier surface in an automated manner, we created a new tool called DL4GAM. This tool automatically selects satellite images suitable for delineating the glaciers and then uses an ensemble of Neural Networks to analyze the selected images and measure changes in glacier size. By studying the glaciers in the European Alps, for which a previous study estimated an annual change rate of −1.3% over 2003–2015, we found that they were shrinking at an increased rate of about 2% per year over 2015–2023. This increased glacier area loss confirms the impact of on-going climate change.

1. Introduction

Glaciers represent a critical component of the Earth system, playing various roles in our society, from sea level rise (Edwards et al., 2021) and water security (Immerzeel et al., 2020) to tourism (Salim, 2023). Moreover, glaciers are unique indicators of climate change (Hock & Huss, 2021), and have been classified as an Essential Climate Variable under the Global Climate Observing System (GCOS) (Bojinski et al., 2014). Due to accelerated melt observed over the last decades (Hugonnet et al., 2021), frequent updates of the glacier outlines inventory are needed for change assessment, their surface being one important parameter that can now be tracked at large scale using satellite imagery. The most recent inventory for the glaciers in the European Alps was published in 2020 by Paul et al. (2020) based on images dating (mainly) back to 2015. Compared to the previous regional inventory for the region, built using imagery from 2003 (Paul et al., 2011), an estimated loss of ca. 300 km^2 was found, which translates into a shrinkage rate of 1.3% y^{-1} over 2003–2015.

Another crucial indicator of a glacier's "health" is its Mass Balance (MB), defined as the total sum of all the accumulation (e.g., snow, freezing rain, avalanches) and ablation (e.g., melting, calving, sublimation) across the entire glacier over a certain period (e.g., a year). To estimate it, one common approach is through Digital

Elevation Model (DEM) differencing. This involves co-registering two DEMs and calculating the elevation difference between them, that is, the volume change. By making certain density assumptions, this volume change can then be converted to mass change, a technique known as the geodetic method (Berthier et al., 2023). During this process, it is also essential to account for potential changes in glacier area, especially when calculating the so-called specific MB (total mass balance per unit area). Many studies assume here a constant area and rely on single-dated glacier outlines (Berthier et al., 2023). However, this assumption can introduce significant biases, particularly if there has been considerable glacier shrinkage between the DEM acquisitions, with errors reaching up to 19% (Florentine et al., 2023).

Building a glacier inventory requires intense manual editing, as standard automated mapping methods (e.g., band-ratio thresholding) fail in many cases, for example, for debris-covered glaciers (Paul et al., 2013). The presence of debris makes glacier mapping a challenge even for experts, where the interpretation can be subjective, sometimes leading to errors in the order of 10%–20% for small glaciers (Paul et al., 2020). This makes the glaciers in the European Alps particularly challenging, as around 16.4% of the glacierized area is estimated to be debris-covered (Herreid & Pellicciotti, 2020), much higher than the globally estimated fraction of 7.3%. For the glaciers in the Swiss Alps, which represent approximately one-half of the glacierized area of the European Alps, Linsbauer et al. (2021) estimate 11% debris-coverage. However, despite having a significant debris-coverage fraction compared to other regions, the fraction of clean ice remains significantly higher. Therefore it is worth investigating whether glacier area change rates could be accurately estimated using the standard band-ratio thresholding, which we analyze in this study.

A promising approach toward fully-automated glacier mapping has relatively recently evolved, based on Deep Learning (DL) models. One major advantage of these models lies in their capacity to ingest multiple data modalities and automatically extract the necessary features. Xie et al. (2020) were the first to employ a fully convolution neural network on this task, showcasing the potential to map debris covered glaciers. Since then, many other methods have been proposed, for example, Xie et al. (2022), Tian et al. (2022), Peng et al. (2023), Thomas et al. (2023), Maslov et al. (2025) (see Diaconu, Heidler, et al. (2025) for a detailed overview of the existing works). However, to the best of our knowledge only two studies have investigated whether these DL-based methods can be used for glacier area change assessment. Roberts-Pierel et al. (2022) trained a segmentation model to detect the glaciers in Alaska and then apply it on biannual image composites from Landsat over 1985–2020. They estimated a change of 8,425 km^2 (−13%) from a total initial area of 64,077 km^2, the equivalent of an annual rate of −0.37% y^{-1}. Rajat et al. (2022) performed a similar analysis for a small Himalayan region (Himachal Pradesh), also based on Landsat imagery, and estimated that glaciers in this region shrank from a total of 4,027 km^2 in 1994 to only 2198.5 km^2 in 2021, that is, an annual rate of −1.68% y^{-1}. Our current work continues on this line of research, focused on the European Alps, and makes the following contributions:

- We build a Deep Learning-based automatic framework for Glacier Area Monitoring (DL4GAM), which includes uncertainty quantification through ensemble learning and a procedure to automatically identify the imagery with the best mapping conditions independently for each glacier. DL4GAM is available open-source at https://github.com/dcodrut/dl4gam_alps.
- We release the curated and processed training data set, ready to be used by other Machine Learning researchers and glaciologists, together with the predictions of DL4GAM which can be visualized at https://dcodrut.github.io/dl4gam_alps.
- We assess the glacier area change for the European Alps over recent years (2015–2023), highlighting the benefit of using elevation change maps as complementary inputs for the DL models to further improve the mapping of debris-covered areas. At the same time, we acknowledge that the DL-based predictions also have limitations and may not yet meet the quality standards expected for glacier inventory production without additional corrections.

2. Data

2.1. Optical Data

To estimate glacier area changes, we collected two Sentinel-2 images for each glacier, as far in time as possible (to increase the signal-to-noise ratio). To download the data independently for each glacier we used the `geedim` (Leftfield Geospatial, 2021) Python library. The first image is always from the same year as in the glacier

Table 1
Glacier Inventory Description

Year	Count	Area
2015	3,063 (69.7%)	1420.1 km^2 (78.6%)
2016	1,260 (28.7%)	367.8 km^2 (20.4%)
2017	72 (1.6%)	18.0 km^2 (1.0%)
All	**4,395**	**1805.9 km^2**

Note. Number of glaciers and their total area, separated by the imagery acquisition year, based on the inventory from Paul et al. (2020), which is used for training and evaluating our models. The last row, in bold, shows the regiowide totals.

inventory we use to obtain the training labels (details in the next section). For the second image, we decided to use the year 2023, as currently, this is the most recent year with good glacier mapping conditions (more details in Section 2.1.2).

2.1.1. Data Matching the Glacier Inventory (ca. 2015)

Our analysis is based entirely on the glacier inventory built by Paul et al. (2020) for the European Alps. The inventory was produced using (mainly) Sentinel-2 images by manually correcting the band ratio thresholding method results, using the Red and Short-Wave InfraRed (SWIR) bands. The final product contains 4,395 glaciers ≥ 0.01 km^2, covering a total area of 1805.9 km^2. The authors aimed to use mainly imagery from August 2015 whenever the mapping conditions were satisfactory, that is, cloud-free and without seasonal snow. However, when this was not possible, they employed imagery from 2016 or, in some cases, 2017 (see Table 1 for details). For simplicity, throughout the text we will consider 2015 when referring to the inventory date. In the particular cases where more clarity is needed, we will use "inventory" years to refer to the exact dates.

Since we mainly rely on Sentinel-2 data, which comes at a 10 m Ground Sample Distance (GSD), we decided to drop glaciers smaller than 0.1 km^2, which corresponds to having at least 1000 pixels. Even if some of the small glaciers are still detectable, we will face difficulties when estimating the area change since the signal is probably dominated by noise.

Moreover, while visualizing our preliminary results, we observed that some glaciers still have poor-quality imagery, mainly because of cloud cover, despite being manually selected by the experts who built the inventory (Paul et al., 2020). This issue is due to choosing the best data at the sub-regional level, not per individual glacier. To address this issue, we decided to impose a maximum cloud/shadow coverage of 30% relative to the surface of the glacier with a 50 m buffer around it (to make sure we increase the chances that the edges are visible), computed using `geedim` (Leftfield Geospatial, 2021). Additionally, we identified a few glaciers with too much seasonal snow or completely covered in cast shadows. To avoid discarding all of them, we first tried to find imagery with better mapping conditions while keeping the same year. Since we had to visualize all the glaciers during this process, we identified cases where we could find better data for some glaciers that already had satisfactory ones. Ultimately, we manually removed 53 glaciers and replaced the imagery of 157 (40 that otherwise would have been removed, plus another 117 because the data was better). The spatial extent of the

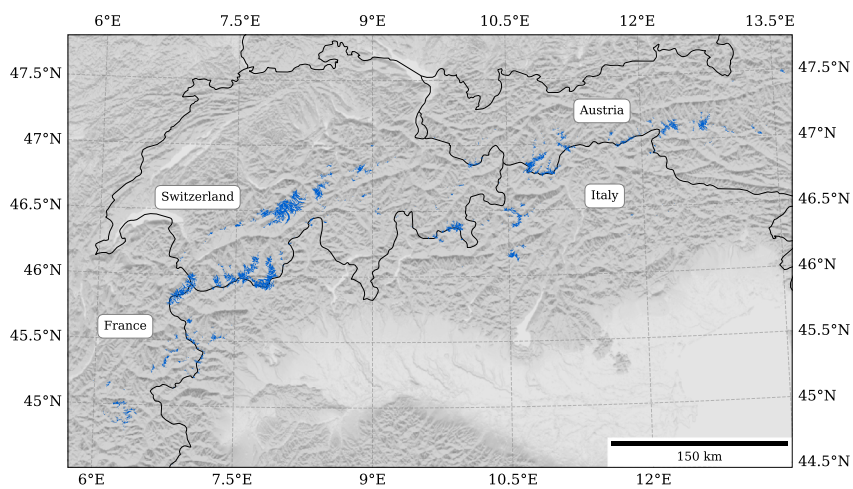

Figure 1. Overview of the study region. The figure shows the outlines of the glaciers included in our study ($n = 1593$, see also Table 2) based on the inventory from Paul et al. (2020), together with the corresponding countries (background: Copernicus GLO-30 DEM hillshade).

Table 2
Data Filtering Steps

Step	Count	Area
Before filtering	4,395	1805.9 km^2
After filtering by area ($\geq$ 0.1 km^2)	1646 (37.5%)	1706.2 km^2 (94.5%)
After manual quality check	1593 (36.2%)	1684.7 km^2 (93.3%)

Note. Number of glaciers and their total area after filtering, first by imposing a minimum area of 0.1 km^2, then by manually removing the glaciers with poor-quality imagery (for which alternative dates in the same year could not be found).

glaciers and the statistics of the final data set are shown in Figure 1 and Table 2, respectively. Note that we still cover more than 93% of the total glacierized region despite including only ca. 36% of the glaciers.

2.1.2. Recent Data (2023)

We aimed to collect the most recent data to increase the length of the covered period. For 2024, however, heavy snowfall in early September made the glacier mapping conditions unsuitable, while the data from late August still shows some seasonal snow. This is also supported by the data from the Glacier Monitoring service in Switzerland (GLAMOS), which shows much less negative glacier mass balances for 2024 compared to, for example, 2022 or 2023, with some glaciers being close to equilibrium (GLAMOS, 2024).

According to the GLAMOS data, 2023 was a strong melt year (GLAMOS, 2024), which makes the mapping conditions more favorable. This allows a more relaxed acquisition time window, increasing the chance of having multiple good-quality images per glacier. Moreover, since 2017, two Sentinel-2 satellites have been available, and therefore, the amount of data available in 2023 is greater compared to earlier years. Across all glaciers, if considering all the Sentinel-2 tiles with less than 75% cloud coverage from 2023-08-01 to 2023-10-15, we obtain ∼30 k candidate images, significantly more compared to 2015 or 2016, with ∼10 k and ∼15 k candidates, respectively.

To ensure an efficient image selection process, we employed an automated pipeline that incorporates the following steps:

1. We acquire all the images from 2023-08-01 until 2023-10-15, independently for each glacier, but only considering tiles with less than 75% cloud coverage. We include a buffer of 1.33 km around the glacier outline which will be needed when sampling the training patches (detailed in Section 3.4). The resulting 30 k candidate images span 357 unique Sentinel-2 tiles. On average, each glacier is covered by around 18 images (range: 9 to 25, due to orbital overlaps).
2. After clipping each image to the buffered glacier extent, we compute four scores:

(a) The percentage of valid pixels within the scene (to exclude images where the glacier lies near a tile boundary)
(b) The cloud coverage percentage (including cloud shadows) over the glacier surface plus a 50 m buffer;
(c) The average Normalized Difference Snow Index (NDSI), using the cloud-free non-glacier pixels within a 50 m buffer;
(d) The average albedo, using the cloud-free glacier pixels plus a 50 m buffer, computed with the following RGB-based approximation: $albedo = 0.5621 \cdot B + 0.1479 \cdot G + 0.2512 \cdot R + 0.0015$ (Wang et al., 2016).

3. For each glacier, we keep only the images with more than 90% valid pixels and less than 30% cloud coverage. This leaves ∼17 k candidates (i.e., around 10 per glacier) across 282 unique tiles.
4. We sort the rest by cloud coverage and, separately, by the NDSI, after we round these two metrics to the nearest integer percentage. Based on the position in the sorted list, an image will get two scores (1 to n), one for each of the two criteria. Finally, the two scores are averaged and the images are sorted based on this. If two images get the same final score, the estimated albedo (without any rounding) is used as a tie-breaker, choosing the image with the smallest albedo value.

With this procedure, we aim to minimize both the amount of clouds and the amount of snow, so we give equal weight to these two criteria, the sorting-based scores being used as a normalization step. Therefore, we rely on the assumption that for each glacier, there is at least one image with good mapping conditions. We then aim to retrieve the best image with our procedure, the advantage of which is that we don't have to set absolute thresholds. We visualized the selected images ($n = 1593$), and with minor exceptions (mainly caused by bad-quality cloud masks), the overall quality was good. To further validate the procedure, we also ran it for the same years as the inventory ones, to check whether the automatically selected images match those from our 2015 data set and found an agreement in more than 98% of the cases (details are provided in the Supporting Information S1).

To validate the temporal generalization capabilities of our models, we built a small (reference) data set for 2023, consisting of 130 glaciers, which we manually delineated using very-high resolution data from swisstopo (2024a, 2024b) (details are provided in the Supporting Information S1).

2.2. Auxiliary Data

2.2.1. Surface Elevation

One of the major challenges in glacier mapping, assuming we already have good-quality imagery, is posed by debris-cover. The percentage of debris can vary significantly, with some glaciers being completely covered. Since it is difficult to distinguish the parts of the glacier covered by debris from the surrounding terrain, a DEM is usually provided as input (Maslov et al., 2025; Peng et al., 2023; Thomas et al., 2023; Tian et al., 2022; Xie et al., 2020, 2022). This is based on the assumption that there is a link between elevation and debris coverage (e.g., there is a much higher chance of having debris at the tongue). Additional features extracted from DEM are presented in Section 3.1.

As data source, we use the freely available Copernicus DEM GLO-30 (Release 2023_1), with a 2010–2015 acquisition phase, at a 30 m GSD (Copernicus - ESA, 2023). We also performed some experiments with the EEA-10 version, which provides 10 m elevation data for European countries but with restricted usage access. Since we did not observe any significant improvement, we will rely on the GLO-30 version. Note that the same DEM, along with the features derived from it, will be used for both the 2015 and 2023 imagery due to limitations in data availability. However, we expect that although the surface may lower over time, the relative topographic patterns remain stable enough to be informative. Therefore, we assume that the features derived from the DEM are still representative of both the glacier surface and adjacent terrain, capturing the stable topographic context of the glacierized valleys. We also note that the DEM-derived features play only a supporting role in the model, and our experiments indicate that the model's overall performance is not strongly dependent on them.

2.2.2. Surface Elevation Change

For the first time, we investigate the use of elevation change (dh/dt) maps as complementary input data to alleviate the problem posed by the presence of debris. Even though a thick enough layer of debris reduces the glacier melt (Rounce et al., 2021), almost all the glaciers in the world lost volume over 2000–2019 (Hugonnet et al., 2021). Consequently, an dh/dt map could also capture the debris-covered portions when contrasting them with the surrounding topography, where no change is expected. Based on the product of Hugonnet et al. (2021), we use the 2010–2014 and 2015–2019 dh/dt maps as additional inputs to the 2015 and 2023 rasters, respectively.

3. Methodology

3.1. Pre-Processing

We use all the 10 m bands (i.e., R, G, B and Near-InfraRed (NIR)) and one of the SWIR bands (i.e., B11). The latter helps distinguish snow and ice from clouds and is also used for calculating the NDSI. Since the SWIR band comes originally at 20 m resolution, we resampled it to 10 m using bilinear interpolation to align it the other input bands.

We also derived two sets of features following previous studies on glacier mapping (Peng et al., 2023; Thomas et al., 2023; Xie et al., 2020):

- Three from the optical bands: NDSI, Normalized Difference Water Index (NDWI) and Normalized Difference Vegetation Index (NDVI).
- Five from the DEM: slope (Horn, 1981) and aspect (using its sine and cosine as input), terrain ruggedness index (Riley et al., 1999), planform and profile curvatures (Zevenbergen & Thorne, 1987). All are computed using the `xDEM` Python library (xdem contributors, 2021).

To summarize, we obtain a 16-dimensional input after stacking all the features:

- Five optical bands + three derived features
- One DEM + six derived features
- One elevation change map.

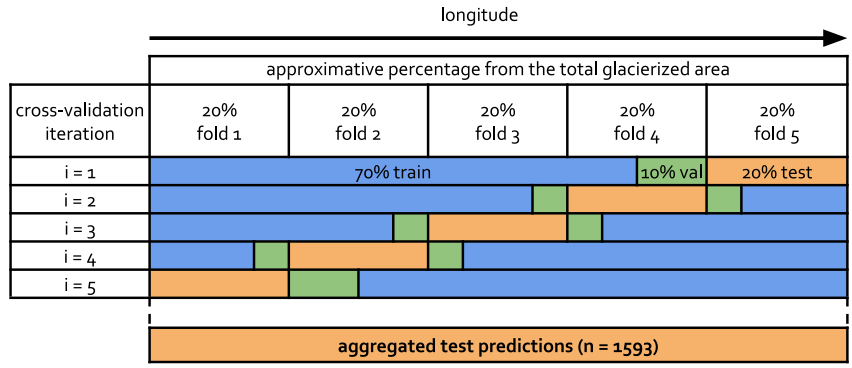

Figure 2. Cross-validation scheme with a geographic split. The figure displays how we split the data set iteratively (five-fold cross-validation) into training-validation-test folds geographically (using the longitude of the glaciers' centroids). To ensure no overlap between training and testing, the validation fold is used as a buffer between them. Note that the split into the three sub-sets is based on the glaciers' areas. Therefore, the number of glaciers in the test fold varies significantly across iterations (see Table 3). Finally, once the 10 models are trained, by collecting the inferences on the test folds, we obtain one estimate for each of the 1593 glaciers considered in our study.

3.2. Segmentation Model

Our method is based on the U-Net architecture (Ronneberger et al., 2015), using the implementation from the `smp` library (Iakubovskii, 2019). U-Net is a Convolutional Neural Network (CNN) architecture specifically designed for image segmentation tasks. Its name is based on its U-shaped architecture, which consists of a contracting path (encoder, for which we use another well-known CNN as feature extractor, i.e., ResNet34 (He et al., 2016)) and an expansive path (decoder). The contracting path captures context information through a series of convolutional and pooling layers, while the expansive path upsamples the features to generate a detailed segmentation map. A key feature of U-Net is skip connections, which allow the network to preserve spatial information and fine-grained details. This architecture has proven highly effective in many Computer Vision tasks, including in previous studies on glacier mapping (Xie et al., 2021).

3.3. Geographic Cross-Validation

Given that the models we train are not error-free, we want to at least increase the chances that these errors are not significantly affecting our glacier area change rate estimation. One way to support this is by using the inferences on both the 2015 and 2023 images, such that, at least, the systematic errors cancel each other when computing the difference. Furthermore, to increase the chances that the predictions are not biased toward the inventory, as it is also used for training, we will only refer to the images falling into the testing subset. However, since our final goal is to provide glacier area change rates for as many glaciers as possible, ideally covering the entire European Alps region, using only the testing subset is insufficient. To address this, we train in a five-fold cross-validation scheme with a geographic split (see Figure 2). This not only prevents any data leakage but also allows us to obtain a test inference on all the glaciers covered in the data set. The statistics corresponding to the test fold glaciers of each cross-validation iteration are shown in Table 3.

Table 3
Geographic Cross-Validation

Subregion	Lon range	#Glaciers	Area
R_1	10.7°–13.6°E	426 (25.9%)	322.9 km^2 (18.9%)
R_2	8.4°–10.8°E	401 (24.4%)	343.1 km^2 (20.1%)
R_3	7.8°–8.4°E	176 (10.7%)	342.6 km^2 (20.1%)
R_4	7.3°–7.9°E	236 (14.3%)	337.9 km^2 (19.8%)
R_5	6.0°–7.3°E	354 (21.5%)	338.2 km^2 (19.8%)

Note. The table shows the statistics corresponding to the test fold of each cross-validation iteration (see Figure 2 for a graphical description of the splitting procedure).

3.4. Training Scheme

We employ a custom patch-based sampling strategy to ensure that our model captures the relevant features and avoids being biased toward the majority class (i.e., non-glacierized areas). For each glacier stack (of varying sizes), we generate all possible patch locations using a patch size of 256 × 256 pixels (i.e., 2.56 km × 2.56 km) and a sampling stride of 32 pixels in each direction. From these, we keep only the patches that have the center pixel within a 50 m buffer of the glacier. Additionally, to guarantee representation for each glacier, we include one patch centered on the glacier centroid. Depending on

the cross-validation fold, this results in 15,160–15,739 patch locations. These locations are stored and used to extract patches on-the-fly during training (in batches of 16), after randomly subsampling them to approximately half the original number (see Section 3.5.3).

We train the models using PyTorch (Paszke et al., 2019) and PyTorch Lightning (Falcon & The PyTorch Lightning team, 2019). We use the focal loss (Lin et al., 2020), which was found to have a positive impact on the model's calibration (Mukhoti et al., 2020). Pixels without optical data or covered by clouds/shadows are excluded when computing the loss. The models are trained using Adam (Kingma & Ba, 2015), with a starting learning rate of 1e−4. Before inputting the data to the network, the slope (in degrees) is divided by 90, and all the variables that are not already in the unit scale (i.e., the optical indices and the aspect's (co)sine) are normalized to zero mean and unit standard deviation using the statistics of the training fold glacier stacks.

We initialize the ResNet34 encoder of the U-Net with the ImageNet weights (He et al., 2016) and then train each model for 30 epochs. We save the model with the best validation performance measured as follows: for each glacier in the validation fold, we compute the average Intersection Over Union (IOU) over all its patches and then compute the overall IOU average, weighted by glacier area (favoring larger glaciers). While uniform patch sampling would implicitly give more weight to larger glaciers, our glacier-centric sampling strategy introduces some imbalance across the region. This explicit weighting step helps restore the original distribution of glacier areas, making the metric more representative of total ice coverage and more comparable across experiments with different sampling settings.

3.5. Glacier Area (Change) Estimation

The first step for producing change estimates is to process the predictions from the segmentation models and estimate glacier areas. We then evaluate the performance of our models using various metrics w.r.t. to both the 2015 inventory and our small reference data set for 2023. Because debris cover, shadows, and snow conditions can lead to erroneous delineations, we quantify the uncertainty of each prediction using an ensemble of (10) models. These uncertainties are then used to identify and exclude unreliable results. Finally, we scale up the filtered area change rates to the full glacier inventory, obtaining a robust assessment of regional glacier area change.

3.5.1. Glacier Segmentation and Area Estimation

Once the models are trained, we want to use them to estimate the area of each glacier, first in 2015 and then in 2023. We build (in memory), glacier by glacier, all the patches with a sampling step of 32 pixels (so doubling the overlap used in training). Then, all these patches are mosaicked while the overlapping predictions are averaged. If some pixels were masked out because they had missing data or were covered by clouds/shadows, we fill them in using the average predicted probability of the closest 30 pixels. A binary mask is then computed by applying a threshold of 0.5 on the pixel-wise probabilities. Finally, these binary masks are used to estimate the glacier areas.

Paul et al. (2020) estimates that the uncertainty of the glacier outlines is between one and two pixels (i.e., 10 and 20 m, respectively), depending on the degree of debris cover. Thus, since we are interested in glacier-wide area change and not in “discovering” new glaciers in the region, we only use the predictions on the pixels within the inventory outlines plus a 20 m buffer. Note that a standard buffer may then include the pixels of another neighbor glacier but these regions are ignored, that is, equivalent to keeping the same ice divides. We also use the same buffer in 2023, thus assuming that glaciers do not grow, which is expected given the recent negative estimates for mass balance or volume change (GLAMOS, 2024; Hugonnet et al., 2021). In this way, we can track the area of each glacier individually, even if some glaciers may disintegrate into multiple parts over time. Finally, for each glacier, the difference between the area estimation for 2023 and the one for the inventory year is divided by the number of years in between (i.e., 8, 7 or 6 years), to obtain a glacier-specific area change rate. An illustration of the entire process for a single glacier is shown in Figure 3.

3.5.2. Performance Evaluation Metrics

We first use the standard evaluation procedure from Computer Vision, that is, computing various performance metrics over the testing samples and reporting their average. We first compute the number of True Positives (TP),

10.1029/2025EA004197

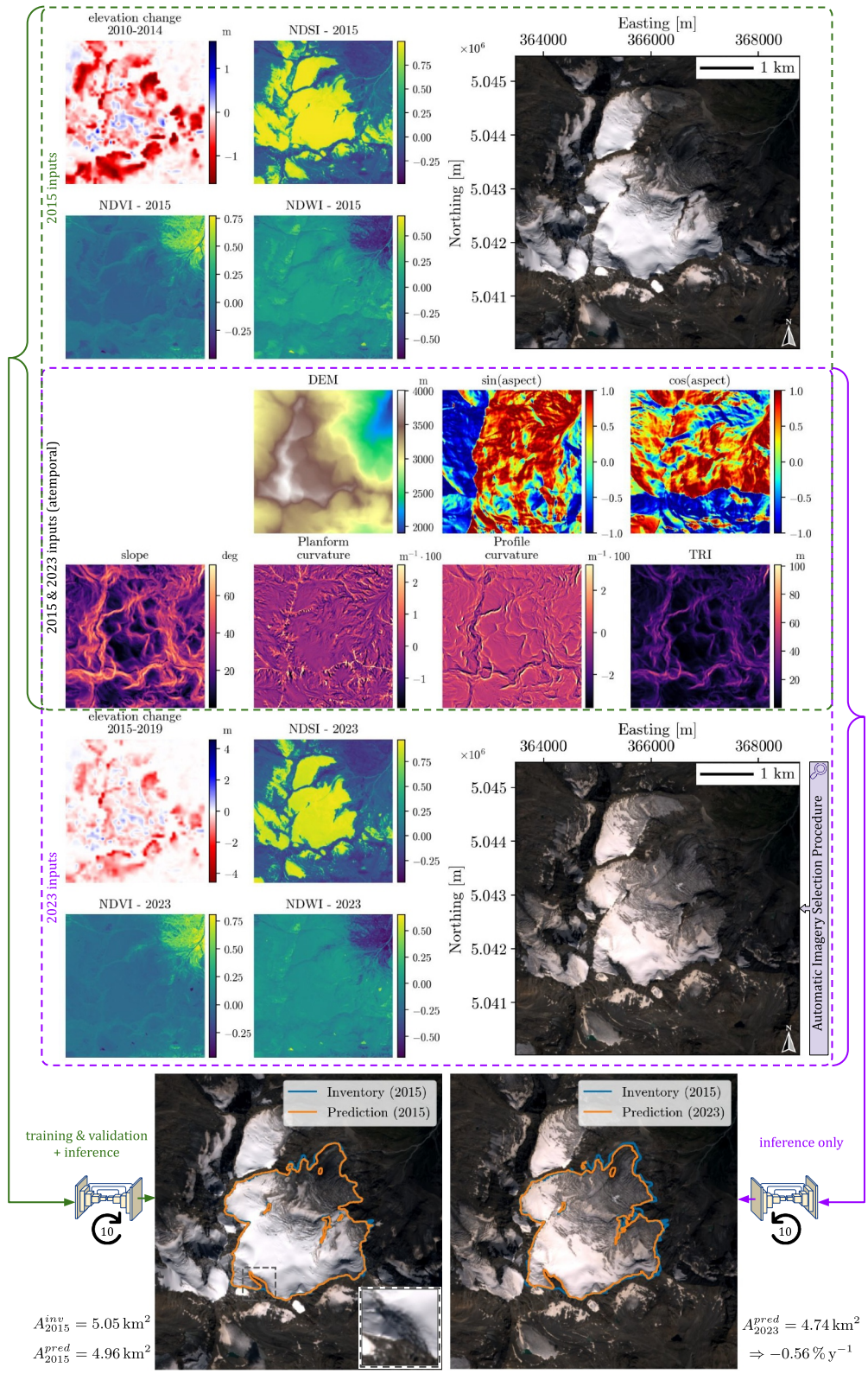

Figure 3. DL4GAM exemplified for a glacier (Tribolazione, IT - 45.52°N, 7.28°E). Note: we show only the RGB bands (Copernicus Sentinel-2) and we clipped the extreme 1% values for the dh/dt, planform, and profile curvatures. The inset in the 2015 image shows a close-up on an area where the prediction and the inventory disagree (probably due to snow-cover).

False Positives (FP), True Negatives (TN), and False Negative (FN). For the FP, we consider only a 50 m buffer around the reference outlines to reduce sensitivity to irrelevant FPs far from glacierized areas. We then derive the following metrics:

- **Accuracy** = $(TP + TN)/(TP + FP + TN + FN)$
- **IoU** = $TP/(FP + TP + FN)$
- **Precision** = $TP/(TP + FP)$
- **Recall** = $TP/(TP + FN)$
- $\mathbf{F_1} = 2 \cdot TP/(2 \cdot TP + FP + FN)$ (i.e., the harmonic mean of Precision and Recall)

In addition to the standard metrics above, we report a second set of more interpretable, area-based metrics, which will be aggregated across all the glaciers of a similar size:

- **Positive area:** $P = TP + FN$, that is, the true glacierized area according to the inventory outlines.
- **True Positive Rate (TPR):** Same as recall, that is, the proportion of the reference glacier area correctly predicted.
- **False Negative Rate (FNR):** $FNR = FN/P = 1 - TPR$, that is, the proportion of the reference area that was not captured (omission rate).
- **Negative area:** $N = FP + TN$, that is, the area outside the reference outlines (within a 50 m buffer), used for computing false positives.
- **False Positive Rate (FPR):** $FPR = FP/N$, that is, the proportion of non-glacier area incorrectly labeled as glacier (commission rate).
- **Relative uncertainty:** $(FN + FP)/P$, a normalized error metric representing the total segmentation error magnitude relative to the true glacierized area.

3.5.3. Uncertainty Quantification

Estimating the uncertainties in DL models remains a challenging but highly important task. Uncertainty estimates are necessary for assessing the reliability of the predictions before interpreting them. Initially used in Random Forests (Breiman, 2001), one of the classical methods that remains a robust approach is bagging, that is, training multiple models on a different sample of the original data set and ensembling them. The average predictions usually also outperform the individual members, but here we are mainly interested in the spread of the ensemble, which provides a measure of predictive uncertainty. This idea has also been exploited in Deep Neural Networks by Lakshminarayanan et al. (2017), showing that ensembles provide high-quality uncertainty estimates and became a gold standard in probabilistic machine learning (Wimmer et al., 2023). For a broader overview of ensemble methods in deep learning, we refer the reader to Gawlikowski et al. (2023), who highlight their widespread use and strong empirical performance in uncertainty estimation, and also discuss methods for making ensembles more efficient.

We built an ensemble of ten U-Nets for each cross-validation split. Before training, we sample 7,488 patches (i.e., 468 batches) from the corresponding training set, which is approximately half of all the generated ones, to increase the diversity among the ensemble members. Additionally, the decoder and segmentation head weights are re-initialized with a different random seed. Once the models are trained, we will have 10 predictions for each pixel. To ensure that the predicted probabilities are well-calibrated, we applied temperature scaling (Guo et al., 2017) to each ensemble member, using pixel-wise predictions and ground truth labels from the validation set. The optimal temperature for each model was chosen to minimize the Expected Calibration Error (ECE), which we compute and report. Based on these calibrated values, we will first compute the average, which will be used as the final prediction, and their standard deviation, which will be used to derive the lower and upper bounds of the glacier area.

Following the work of Tollenaar et al. (2024), we initially defined area uncertainty by subtracting from (or adding to) the ensemble average prediction one standard deviation, then applying a fixed 0.5 threshold to compute the lower and upper bounds. However, to address limitations of this heuristic—particularly in cases where ensemble members may exhibit low spread but still be uncertain—we extended it by incorporating an area-level calibration step. First, we introduce a glacier-specific parameter $\tau \in [0, 0.5]$ and compute the glacier area bounds as a function of τ as follows:

- $A_{\text{lb}} = a_{\text{px}} \sum_{i \in \mathcal{B}_{20m}} \mathbf{1}\left[\left(\mu_i - \sigma_i\right) \geq 1 - \tau\right]$,
- $A_{\text{ub}} = a_{\text{px}} \sum_{i \in \mathcal{B}_{20m}} \mathbf{1}\left[\left(\mu_i + \sigma_i\right) \geq \tau\right]$,

where μ_i and σ_i are the ensemble mean and standard deviation at pixel i, τ is the calibrated threshold, and $\mathcal{B}_{20m}$ denotes the set of pixels within a 20 m buffer around the target glacier outline (same as at inference time), excluding any pixels belonging to adjacent glaciers. The factor a_{px} accounts for the area of a single pixel (i.e., 1×10^{-4} km^2 for Sentinel-2). This works as follows: by setting the lower bound threshold to $1 - \tau$ and the upper bound threshold to τ, we allow the bounds to adjust asymmetrically to the ensemble predictions: if the confidence is low (i.e., μ is close to 0.5), a small standard deviation is sufficient for a pixel to cross the threshold; conversely, a high-confidence prediction (with μ near 0 or 1) paired with high ensemble disagreement (i.e., a large σ) also makes it easier to cross the threshold. This approach ensures that both the central prediction μ and its uncertainty σ are appropriately balanced in determining the bounds. The optimal value of τ is then selected to achieve a target coverage of 68%, corresponding to the probability mass within one standard deviation under a normal distribution. This means that the true glacier area is expected to fall within the predicted lower and upper bounds (i.e., the 1-standard-deviation confidence interval) in 68% of the validation cases (we optimize τ on the glaciers from the validation set for each cross-validation iteration). This method offers two main advantages: (a) it decouples the bounds from a fixed decision threshold, allowing dynamic calibration in cases where all ensemble members predict 0.5 (i.e., no spread), thus preventing the underestimation of uncertainty; and (b) it calibrates area-level uncertainty in a data-driven way, ensuring that the resulting bounds behave as proper confidence intervals under Gaussian assumptions.

To select τ, we sweep from 0 to 0.5 with a step of 0.001. For each candidate τ, and for each glacier in the validation set, we compute the corresponding prediction interval $[A_{\text{lb}}, A_{\text{ub}}]$ using the equations above. We then choose the τ that yields 68% empirical coverage of the true areas. Under our five-fold geographic cross-validation, this produces five fold-specific thresholds, which we apply to the held-out test folds to assess calibration performance. When aggregating all test-fold intervals, however, we observed an overall under-coverage of around 9%, indicating that a single τ per fold does not generalize perfectly across subregions. To address this issue, we replaced the fixed-τ approach with a glacier-specific threshold estimated via quantile regression. Noting that the previously described procedure is equivalent to (a) computing, for each validation glacier, the maximum τ that still covers its true area, then (b) taking the 0.32 quantile of those values, we fit a quantile-regression model at $q = 0.32$. We also use a logit link to enforce $2 \cdot \tau$ in [0, 1]. The model predicts τ from four predictors (plus an intercept):

- mean of the pixel-wise ensemble predictions μ_i
- root-mean-square of the Bernoulli variance $\mu_i \cdot \left(1 - \mu_i\right)$
- root-mean-square of the ensemble pixel-wise standard deviations σ_i
- logarithm of the predicted area (using a 0.5 threshold).

We fit one such model per validation fold and then predict glacier-specific τ values for the corresponding test fold. These thresholds are also applied to the 2023 inferences.

Finally, when computing the 2015–2023 area change rates, we do so on a glacier-by-glacier basis and assume the uncertainties from the two epochs are independent, summing them in quadrature. We acknowledge that this assumption contradicts our earlier suggestion (in Section 3.3) that systematic errors might cancel when comparing 2015 and 2023 predictions. A proper analysis of temporal error correlation would require an expert-made glacier inventory for 2023, which is currently unavailable. We therefore leave this for future work. That said, assuming independence leads to conservative uncertainty estimates—i.e., potentially overestimating the actual uncertainty—which we consider preferable in the absence of a definitive ground truth.

3.5.4. Quality-Control Filtering

While our method is not intended to produce a new glacier inventory, it is well-suited for change assessment, where relative temporal consistency is more critical than absolute per-image accuracy. To further reduce the risk of consistent misclassification (e.g., missing a glacier in both years), we incorporate ensemble-based uncertainty filtering to automatically discard predictions with low confidence—trading some spatial coverage for increased robustness.

Given the two area estimates, we can compute the corresponding annual area change rate for each glacier. However, since DL4GAM can still fail to identify debris-covered areas, we apply some filtering on the estimated

change rates. As a first step, we use DL4GAM's uncertainties as follows: we impose that the ratio between the estimated annual area change rate and its corresponding uncertainty—i.e., Signal-to-Noise Ratio (SNR)—is larger than 1. As a second step, we try to eliminate the cases where DL4GAM fails to identify a glacier and for which the ensemble-based uncertainty did not capture the error, which can still happen especially for (small) fully debris-covered glaciers. As such, we impose a minimum recall of 90% for 2015 to avoid such cases. We then assume that the performance of the models will remain similar for 2023, an assumption that should be met since we followed a strict evaluation strategy using the geographic cross-validation split, and therefore the 2015-based evaluation metrics should also be representative for 2023.

3.5.5. Regional Extrapolation

To avoid any bias introduced by filtering out the glaciers smaller than 0.1 km^2, we first upscale the estimated rates to the entire inventory using a second-order polynomial fit weighted by the estimated uncertainties (details in the Supporting Information S1).

3.6. Band-Ratio Thresholding

For building the inventory used in our study, Paul et al. (2020) initially computed the R/SWIR ratio and applied a manually selected threshold such that good results are obtained for the shadowed areas, with the risk of making false detections. To reduce them, a second threshold on the Blue band was used, as it has been shown in a previous study that this reduces misclassified rock in shadow (Paul et al., 2016). We investigate whether applying this classical method separately on the 2015 images and the 2023 ones would provide similar glacier area change rates.

Instead of manually choosing thresholds for each day or S2 tile, such that they are adjusted to scene conditions (Paul et al. (2020)), we aim to maximize the performance of this method by automatically selecting the best values for the two thresholds using a simple grid search over [0.1, 5.5], with a 0.1 step, and [0, 1500], with a step of 25, respectively. The best threshold pair is chosen similarly to the validation phase of the DL models, that is, by maximizing the average IOU (see Section 3.4). Furthermore, we implemented two versions of calculating these thresholds:

1. **(sub)regional** band-ratio method: for each of the cross-validation splits, we combine all the glaciers from the training and validation folds and choose the pair that provides the best IOU (weighted by the glacier areas). We thus end up with five threshold pairs, one for each cross-validation iteration.
2. **Glacier-wise** band-ratio method: instead of choosing a single parameter of the entire sub-region, we find the pair that yields the best average IOU for each glacier independently.

4. Results and Discussion

4.1. Model Performance Evaluation

Before extracting the glacier areas based on our predictions, we first evaluate the quality of the predictions from the five U-Net ensembles using standard metrics from Computer Vision segmentation problems (introduced in Section 3.5.2). The results are displayed in Table 4, separately for each subregion in the geographic split. An F_1 around 90% on average indicates a good quality of the predictions, and therefore, we can trust them to further compute the total glacier areas.

4.2. Glacier Area (Change) Estimation

Before estimating the corresponding annual change rates, we first investigate how well our area estimates match both the 2015 inventory and our 2023 reference data set. A size-dependent analysis provides detailed insights into the strengths and limitations of our method. Table 5 summarizes the area estimation performance metrics grouped by glacier size class. Our results indicate that relative omission errors tend to be larger for smaller glaciers—reflecting higher sensitivity to boundary misclassification—whereas the false positive rate (commission error) shows the opposite trend.

For our 2023 reference data set, we observe a slight deterioration in recall (TPR), which may be explained by our inclination to be more inclusive when delineating glaciers in the very-high-resolution imagery. The FPR

Table 4
Segmentation Performance Metrics

R	Accuracy (%)	IoU (%)	Precision (%)	Recall (%)	F_1 (%)
R_1	85.2 ± 10.8	75.4 ± 20.5	98.1 ± 3.2	76.7 ± 20.9	84.5 ± 16.8
R_2	90.6 ± 7.9	85.7 ± 13.2	94.3 ± 4.8	90.6 ± 13.8	91.8 ± 9.7
R_3	92.4 ± 5.5	88.9 ± 8.3	94.1 ± 5.4	94.1 ± 6.8	93.9 ± 5.2
R_4	91.8 ± 8.0	88.1 ± 12.9	94.5 ± 4.8	92.7 ± 12.7	93.3 ± 8.7
R_5	88.5 ± 9.4	82.0 ± 16.8	94.4 ± 5.3	86.5 ± 17.9	89.1 ± 13.2
All	89.1 ± 9.3	82.8 ± 16.6	95.3 ± 4.9	86.7 ± 17.5	89.7 ± 12.8

Note. Glacier-level testing results for each of the five geographic cross-validation splits (see Table 3), evaluated using standard segmentation metrics. We report the mean and one standard deviation across glaciers within each corresponding subregion.

decreases slightly for the 2023 data, likely due to the larger buffer of deglaciated pixels (outside the 2015 outlines but within the 50 m buffer).

It is important to note that the inventory outlines themselves are not perfectly accurate. As mentioned in Section 3.5.1, Paul et al. (2020) estimated an uncertainty of 1–2 pixels, and our inference pipeline allows the model to predict glacier presence slightly beyond the inventory outlines (up to +20 m). As a result, part of what is labeled as false positive (FP) may, in practice, fall within the plausible glacier extent. Similarly, some apparent boundary errors may reflect differences in interpretation rather than true misclassification, since pixels at these edges often contain a mixture of glacier ice and rock—especially at 10 m resolution. These effects should be considered when interpreting pixel-level error metrics, as they may overestimate the actual disagreement. For example, when focusing only on a 20–50 m buffer, the FPR shows noticeable differences, as detailed in Table 7.

Next, a glacier-level evaluation is made in Figure 4, showing a Mean Absolute Percentage Error (MAPE) of 12.51% for 2015, followed by 15.95% for 2023. Overall, more than half of the results are not at the level

Table 5
Area Estimation Performance Metrics

Size class	#	P (km^2)	TP (km^2)	TPR (%)	FN (km^2)	FNR (%)	N (km^2)	FP (km^2)	FPR (%)	$\frac{FN+FP}{P}$(%)
Relative to the 2015 inventory (Paul et al., 2020)										
[0.1, 0.2)	509	72.3	58.7	81.1	13.7	18.9	57.7	4.8	8.3	25.6
[0.2, 0.5)	504	159.0	137.6	86.5	21.5	13.5	89.3	7.1	7.9	17.9
[0.5, 1)	235	162.5	146.4	90.1	16.1	9.9	65.9	5.8	8.7	13.4
[1, 2)	175	245.5	229.8	93.6	15.7	6.4	72.6	7.2	9.9	9.3
[2, 5)	101	314.8	298.6	94.8	16.2	5.2	68.0	7.0	10.3	7.4
[5, 10)	48	322.9	310.1	96.0	12.9	4.0	55.5	7.0	12.6	6.2
[10, 20)	16	211.8	204.1	96.4	7.7	3.6	31.0	4.7	15.2	5.9
≥20	5	195.5	190.6	97.5	4.9	2.5	25.1	3.7	14.6	4.4
All	1593	1684.4	1575.8	93.6	108.6	6.4	465.2	47.3	10.2	9.3
S_{2023}	130	167.1	161.4	96.6	5.7	3.4	38.6	5.07	13.1	6.4
Relative to our 2023 reference data set (S_{2023})										
S_{2023}	130	151.8	143.4	94.4	8.4	5.6	53.9	5.10	9.5	8.9

Note. We report several metrics that assess how suitable the final ensemble-averaged predictions are for glacier area estimation. Metrics are grouped by glacier size class, with the final row reporting aggregated results across all inventory glaciers. For each class, the values, expressed as areas, represent the sum over all glaciers in that class. Finally, S_{2023} represents the set of 130 glaciers from our 2023 reference data set.

10.1029/2025EA004197

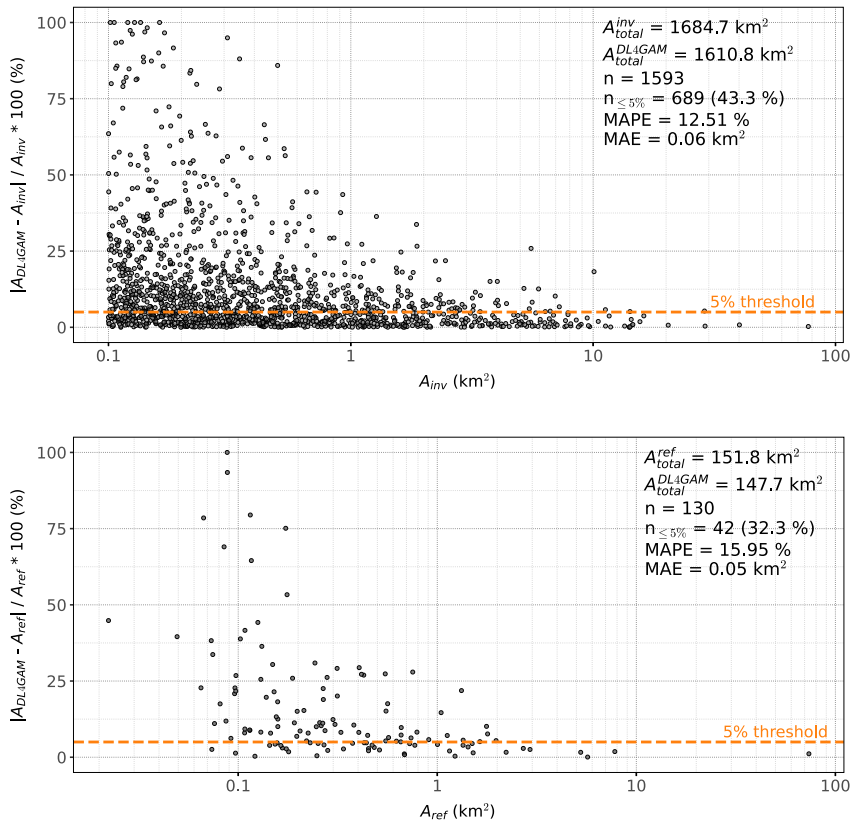

Figure 4. Glacier area comparison against the inventory and our 2023 reference subset. In the upper panel we show the relative absolute percentage errors of the DL4GAM's area estimates versus the inventory ones from Paul et al. (2020), and, below, the results using our small reference data set for 2023. The 5% threshold is the one recommended by GCOS (Global Climate Observing System (GCOS), 2022) as acceptable uncertainties. The text shows how many glaciers pass this threshold. Note also the slight increase in the estimated glacier area compared to Table 5, as we now allow the model to make inferences within the 2015 outlines + a 20 m buffer, to allow for potential errors in the inventory.

recommended by GCOS (Global Climate Observing System (GCOS), 2022), that is, 5%. However, some of these errors are consistent over time, thus canceling out when computing area changes.

In Figure 5 we selected four glaciers (from the subset which we manually delineated for 2023) to illustrate three sources of biases:

- Differences in scene illumination: first row and the third show an increase in the amount of ice covered by shadows; in the first one DL4GAM manages to overcome the change, whereas in the third case it missed a segment;
- Seasonal snow: second and last cases show much more snow in 2015 compared to 2023; this not only creates a base for false positives, but the fact that the amount of snow changes over time contradicts our assumption that errors are systematic;
- Increased debris cover: in all cases except the second the debris cover increases; if for the third case, the models perform well, as there is enough clean ice nearby for context, in the first and last cases, the network misses completely the debris segments.

Similar cases but with very-high-resolution data were illustrated in our previous work (Diaconu, Heidler, et al., 2025)—see Figures 6 and 7. For future work, our automatic data selection procedure should be improved to at least try to eliminate differences in illumination and seasonal snow. Alternatively, we could train the models with multiple images from the same season, as a temporal augmentation. Debris cover, however, remains a challenge, as the current results are not yet at a human annotator level.

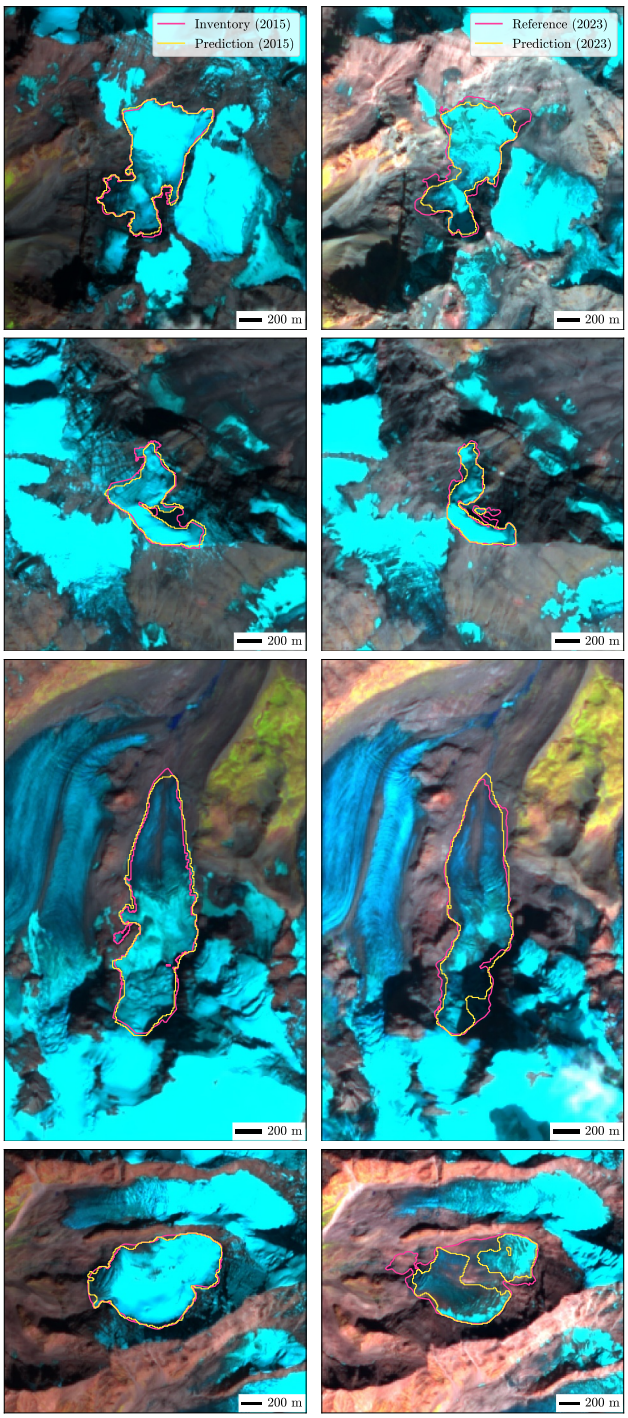

Figure 5. Predictions for the inventory year and our 2023 reference subset. We selected four glaciers for which the surfaces conditions change over time, to illustrate an important source of bias. Imagery: SWIR-NIR-R bands from Copernicus Sentinel-2.

4.2.1. Uncertainty Quantification

As detailed in Section 3.5.3, we then derive the uncertainties in the predictions from the spread of a ten-member ensemble and the confidence of the ensemble average, after applying two calibration steps.

At the pixel level, before calibration, the ensemble members already show good calibration on the test set, with an average ECE of 0.96%. This relatively low value is likely due to two factors: (a) in segmentation tasks, many pixels are easy to predict, so the bin of very confident and correct predictions dominates the ECE computation; and (b) the use of focal loss, which has been shown to improve calibration (Mukhoti et al., 2020). After calibrating each ensemble member on its corresponding validation set, the average ECE is further reduced to 0.52%, with the maximum improvement reaching 1.21%.

The second calibration step, intended to provide area-level confidence intervals, provides a probability decision threshold (τ) for each glacier. When we aggregate the resulting intervals across all test folds ($n = 1\,593$ glaciers), the empirical coverage is 65.2%, slightly below the nominal 68%, with a substantial spatial variation in τ which indicates underlying heterogeneity across glaciers and regions. Importantly, since the reference outlines may themselves contain small delineation errors, the resulting uncertainties are likely conservative. Evaluating the quality of uncertainty estimates remains inherently difficult, as a completely error-free "ground truth" for glacier outlines is not available.

Figure 6 displays the uncertainty buffer and computed area bounds for several glaciers, including failure cases. It is therefore important to note that the effectiveness of these calibrated uncertainties ultimately depends on the quality of the underlying models. While area-level calibration step ensures that area bounds capture errors correctly on average, it cannot account for cases where all ensemble members are confidently and consistently wrong, for example, due to systematic misclassification in fully debris-covered regions. Such failure modes remain a limitation of our approach—and more generally, of uncertainty quantification methods in deep learning (Ovadia et al., 2019).

When comparing our uncertainty estimates to prediction errors with respect to the inventory areas (Figure 7), we observe a significant Spearman correlation of 78%, suggesting that the uncertainty magnitudes do relate meaningfully to error. Figure 7 contains another quality assessment of our uncertainties, where we compare them with debris areas for a subset of 288 Swiss glaciers. We use this sample because we have access through GLAMOS to a high quality debris-cover product, which is part of the Swiss Glacier Inventory (SGI2016), and that roughly matches the period covered in our data set (Linsbauer et al., 2021). We find a significant correlation (47%) but with a high spread, suggesting that the uncertainties also be affected by other components (e.g., cloud/shadow coverage). Overall, we can conclude that DL4GAM's uncertainty can be reliably used when interpreting the estimated glacier area change rates.

4.2.2. Comparison With the Round-Robin Experiment From SGI2016

To illustrate the versatility of DL4GAM, we ran it for 13 Swiss glaciers from the SGI2016 inventory (Linsbauer et al., 2021). These 13 glaciers are a subset of the 15 glaciers used by the authors in a round-robin experiment designed to estimate the uncertainty of the product (two were dropped because one had an area smaller than

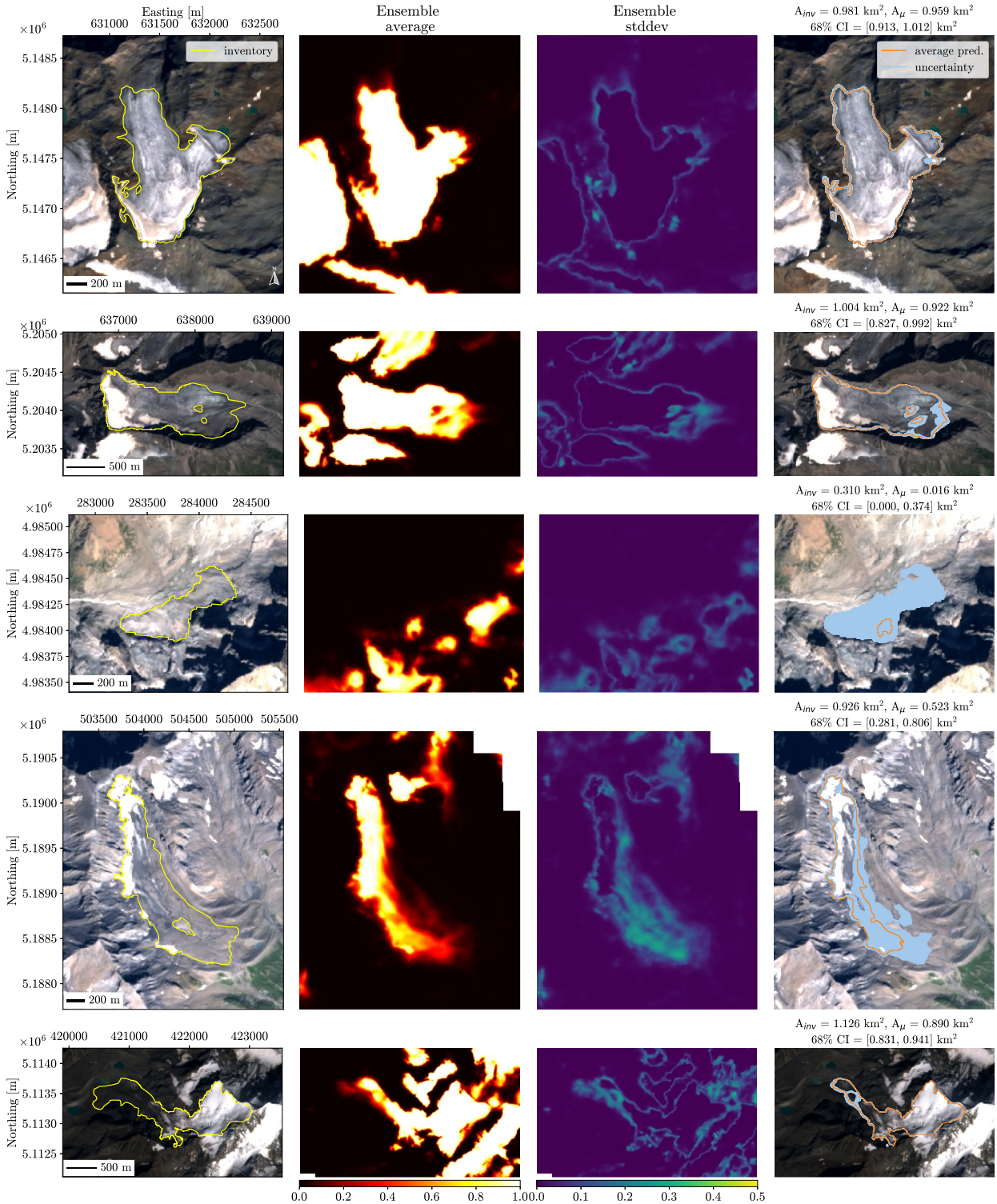

Figure 6. Ensemble-based Uncertainty Quantification. We illustrate here, for several glaciers with different degrees of debris coverage, the process of obtaining lower and upper bounds for the predicted glacier area starting from the ensemble's pixel-wise average prediction (second column) and its standard deviation (third column). The titles from the last column show the inventory area (A_{inv}), DL4GAM's final predicted area (A_μ) and the corresponding 68%-CI bounds (A_{lb} and A_{ub}) which depend on the estimated τ (see Section 3.5.3). For the first glacier, which is almost debris-free, the predictive buffer is relatively narrow. For the second glacier, the uncertainty is slightly higher, capturing well the presence of debris on the glacier tongue. For the third glacier, nearly fully debris covered, the prediction is poor and the upper bound covers almost entirely the allowed 20 m buffer as a consequence of a very low estimated τ (≈ 0.003). For the last two cases, the network misses significant parts of the tongue and the uncertainty buffer is too small, illustrating limitations of our predictive uncertainty (we discuss them in Section 4.2.1). Imagery: Copernicus Sentinel-2.

10.1029/2025EA004197

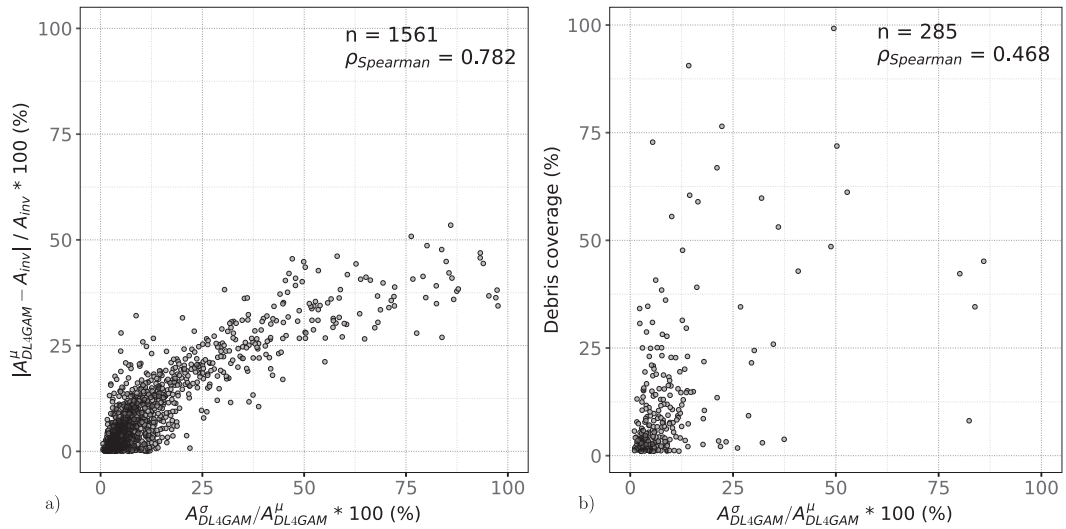

Figure 7. Qualitative checks of the estimated uncertainties. Panel (a) displays a comparison between the estimated relative uncertainties (derived from the ensemble) w.r.t. to the average prediction, and the relative absolute errors w.r.t. the inventory areas. In panel (b), we compare the uncertainties against debris percentages for a small sample of the Swiss glaciers, which are debris-covered (i.e., > 1%) according to the inventory from Linsbauer et al. (2021). Note that we dropped a few glaciers with very small (or zero) predicted areas, that is, $n = 32$ for (a) and $n = 3$ for (b) so we can compute the ratio. In both cases, the (Spearman) correlation is significant.

0.1 km^2 and the other had the outline dated back to 2013, therefore before the Sentinel-2 era). Moreover, having access to area estimates from five different experts provides the opportunity to evaluate the DL4GAM's predictions against reliable ground truth by referring to the average area. Furthermore, the overall uncertainty of the SGI2016 inventory is already lower than that of the inventory we use, estimated at around 2.5%, which is probably a consequence of using very-high resolution (25 cm) optical data. At the same time, we can investigate whether DL4GAM's uncertainties match the human perceptual uncertainty, as this would also indicate a good quality of the predictive uncertainties. Figure 8 shows the two comparisons. For eight out of these 13 glaciers we show the predictions in the Supporting Information S1.

4.2.3. Estimation of Glacier Area Change Rates Over 2015-2023

The benefit of DL4GAM can now be shown for 2023, for which a glacier inventory is not available yet. To validate whether we produce reliable change rates, we compared in Figure 9 the change rates derived using the inventory and our reference data set to those obtained using DL4GAM, by referring only to its predictions. Note that since we did not allow our annotations to be larger than the 2015 inventory, these reference rates are always negative. On the other hand, DL4GAM wrongly estimates positive change rates for two (debris covered) glaciers (only one is shown in the scatter). In general, we observe that DL4GAM overestimates the changes but the overall agreement is good (Spearman correlation: 68.9%). Furthermore, we also check whether our predicted uncertainties do capture the observed errors and we find a prediction interval coverage of 50.8% (nominal: 68%), which we find acceptable given that the reference values themselves have (unquantified) errors. This suggest we can trust the uncertainties in the data filtering step, while using another recall-based one for eliminating cases like the one in Figure 6.

The impact of the two filters is relatively small in terms of total glacier area covered but significant in terms of the number of glaciers. Table 6 shows these coverage statistics after applying each of the two steps. Figure 6 includes a (fully debris-covered) glacier which was filtered out by the second, recall-based, filter.

We summarize the remaining 880 annual change rates by glacier size class in Figure 11. We note that most of the glaciers have a significantly negative change rate but with variation among glaciers, especially for the smaller classes (although for these, the uncertainties are also relatively higher). If we then summarize the glacier area change rates in cells of 10 km × 10 km, we can visualize how they are distributed along the Alps (see Figure 10).

10.1029/2025EA004197

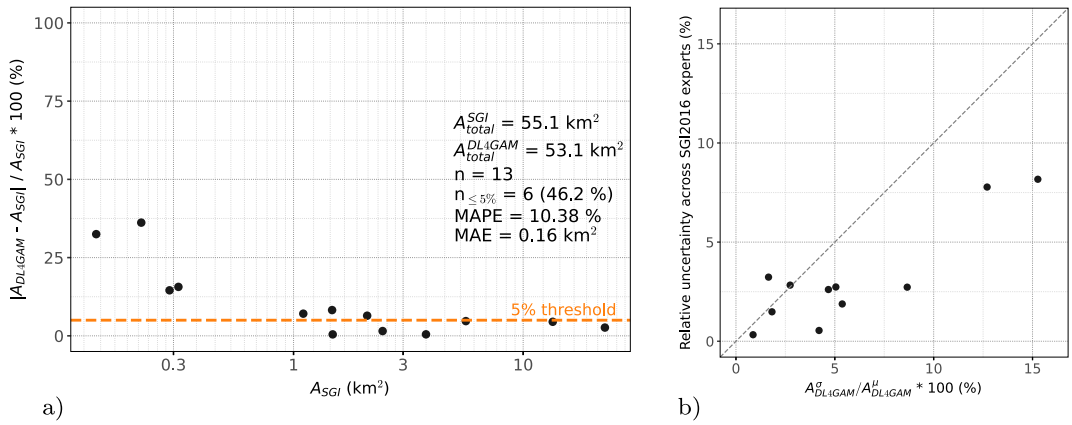

Figure 8. Comparison with the round-robin experiment from SGI2016. The figure shows a comparison between DL4GAM and the results of the round-robin experiment on multiple digitizations of glaciers by five experts from the SGI2016 inventory (Linsbauer et al., 2021). Panel (a) shows the accuracy of the DL4GAM predicted areas against the average across experts, whereas (b) compares DL4GAM's estimated relative uncertainty, derived from the ten-member U-Net ensemble, against the relative deviation of the area across the five experts.

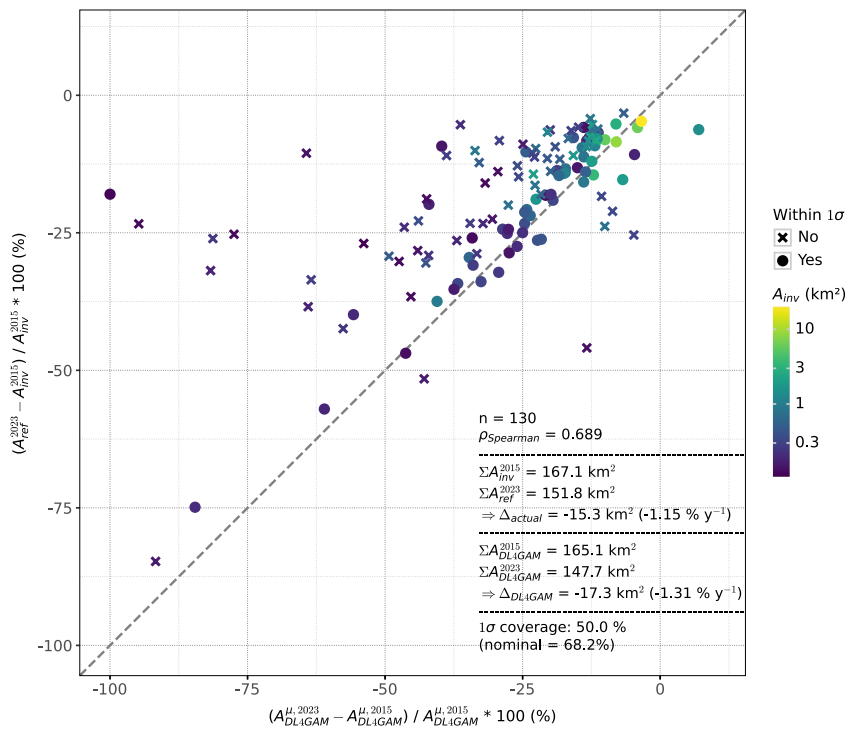

Figure 9. Estimated annual area change rates against the reference. We compare the relative changes in glacier area (expressed as percentage change over the entire 8 years period) between reference measurements (using the 2015 inventory and our small reference data set for 2023), and DL4GAM predictions from both years. The color represents the glacier area in 2015 (with a $\log_{10}$ transformation applied for scale and clipped to 20 km^2 for contrast enhancement), and point shapes denote whether the reference change falls within the 1σ uncertainty interval of the predicted change. The overlaid text box summarizes key statistics: the total number of observations (n), the Spearman correlation coefficient between predicted and measured relative changes, the aggregated glacier areas for 2015 and 2023 (both for reference and DL4GAM estimates), the absolute change in area, and the percentage of reference change values within the predicted 1σ range (with a nominal coverage of 68%). Note that we dropped one outlier from our predictions (with a wrongly estimated change of +41%) for improving visibility, but was included in the statistics.

Table 6
Regional Coverage After Results Filtering

Step	Count	Area
Inventory	4,395 (100%)	1805.9 km^2 (100%)
Our data set	1,593 (36.2%)	1684.7 km^2 (93.3%)
After uncertainty-based filtering (SNR $= \frac{\mu}{\sigma} \geq 1$)	1,136 (25.8%)	1308.0 km^2 (72.4%)
After filtering by recall (≥90%)	880 (20.0%)	1212.5 km^2 (67.1%)

Note. The table shows the coverage impact of the two filters (first by uncertainties, second by 2015 recall) applied on the estimated annual glacier area change rates. Despite reducing the number of glaciers (second column) by another 400, the total area (third column) covered by the estimates remains significant.

4.3. Region-wide Glacier Area (Change) Assessment

Since the 880 estimates cover a significant part of the total glacierized area in the European Alps (i.e., ca. 67%, see Table 6), we can confidently produce a regional estimate. However, rather than throwing away the borderline cases, we instead weighted the predicted area change rates by the estimated uncertainties before applying the size-dependent model described in Section 3.5.5.

For 2015, we estimate 1796.1 ± 91.69 km^2, followed by 1532.6 ± 146.79 km^2 in 2023. After taking into account the period between the two acquisitions of each glacier, which varies depending on the exact inventory year, we obtain a regional annual area change rate of −34.20 ± 22.68 km^2, that is, −1.90 ± 1.26% y^{-1}. Our estimate is therefore more negative compared to 2003–2015, that is, around −1.3% y^{-1} (Paul et al., 2020). However, this is probably an overestimate given the biases that we discussed which would result in an overestimate for 2015 and underestimate for 2023. For instance, if we compute the area change rate using our 2023 reference data set (n = 130), we obtain a rate of −1.31% y^{-1} if we use our models and −1.15% y^{-1} if we compare the data sets directly. Although the difference is relatively small, this confirms the risk of overestimation. The value is also significantly smaller than what we estimate in the end for the entire region which would require further analyses. A second, more recent expert-made inventory would be of real value in elucidating some of these uncertainties.

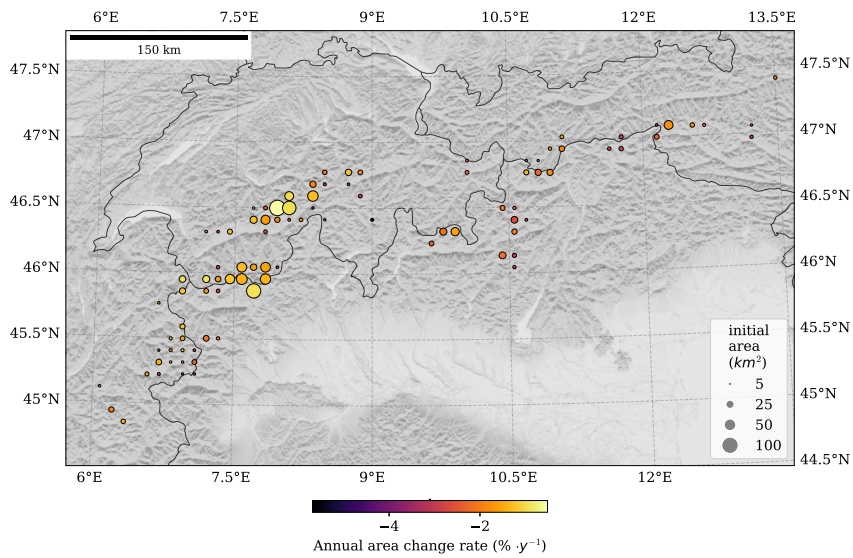

Figure 10. Estimated annual area change rates (gridded). Here, we display (by color) our estimates of annual glacier area change rates after aggregating them in cells of 10 km × 10 km based on the coordinates of the glaciers' centroids. The disks scale with the initial glacierized area in 2015. To enhance the contrast, we only display the cells with an initial area larger than 2 km^2. Country borders are shown in black (see also Figure 1), with Copernicus GLO-30 DEM hillshade in the background. See also the Supporting Information S1 where we show the same plot but for a single glacier area size class.

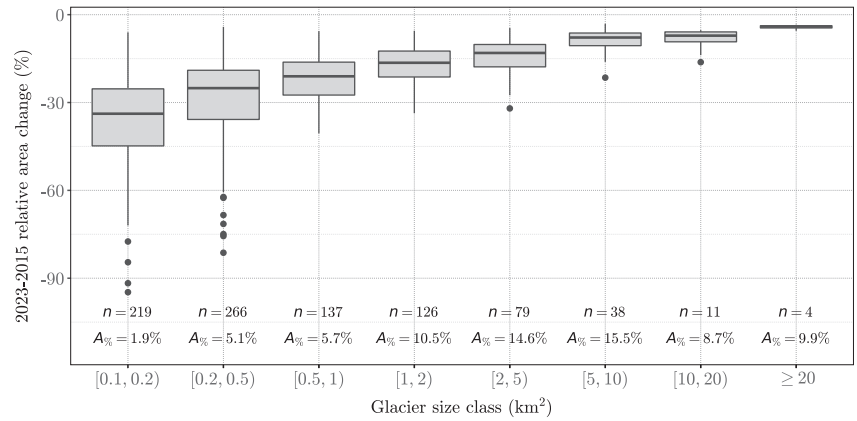

Figure 11. Distribution of glacier-specific area change rates. After splitting the glaciers into the eight different classes used by Paul et al. (2020), we show the distribution of the corresponding annual area change rates estimated in this work. Below each boxplot, we display the number of glaciers (n) falling into that class and the corresponding fraction ($A_{\%}$) of the sum of their areas relative to the total area of 1684.7 km^2.

4.4. Comparative Analysis

4.4.1. Comparison to Clean-Ice-Only Change Rates

The results of the two band-ratio methods described in Section 3.6 are summarized in Table 7 (see also the Supporting Information S1 for an example of how we choose the two thresholds by maximizing the IOU). Based on these results, we make the following observations:

- Despite fitting a much larger number of parameters in the glacier-wise version (i.e., the number of glaciers × 2 vs. only the number of cross-validations splits × 2), the recall stays approximately constant (88%) and only the FP rate decreases from 9% to ca. 7%.
- Given the focus of our work, the most important is the difference between the glacier area change rate estimates, which are significantly more negative for the band-ratio method. One possible explanation could be the negative correlation between glacier areas and debris-cover fractions (Herreid & Pellicciotti, 2020). Knowing also that glaciers are melting rapidly (GLAMOS, 2024; Hugonnet et al., 2021), a shrinkage in their surface is then expected, which will translate into a (relatively) higher percentage of debris in 2023 compared to 2015.
- Additionally, the differences may also come from the calibration procedure: thresholds were optimized based on the 2015 inventory, potentially biasing the model to favor higher recall in 2015 scenes (e.g., by tuning thresholds to include shadowed areas). If the scene conditions in 2023 differ significantly—such as reduced

Table 7
Comparison to Band-Ratio Method and the Impact of dh/dt

	Inventory				2023	
Method	Total area predicted (km^2)	Recall[a] (%)	Recall[b] (%)	FP[c] (%)	Total area predicted (km^2)	Annual area change rate (% y^{-1})
DL4GAM	1610.8	93.5	83.1	4.7	1391.0	−1.78
DL4GAM (no dh/dt)	1593.5	92.6	74.3	4.1	1335.9	−2.10
for clean-ice only: Band-ratio (v1)	1519.1	87.7	–	8.1	1216.8	−2.59
Band-ratio (v2)	1534.4	88.8	–	7.5	1220.0	−2.67

Note. The table shows the results obtained using our U-Net ensemble and two versions of the band-ratio (R/SWIR) thresholding method (v1 and v2 denote the (sub)regional and glacier-wise versions, respectively; see Section 3.6). The second line shows the impact of dropping the dh/dt maps from the inputs. The results are based on the aggregated estimates for all the 1593 glaciers covered in our data set, without any outlier filtering. [a]Here we are only taking into account the predictions strictly within the glacier inventory boundaries (i.e., the 20-m buffer is not used). We then compute the recall by referring to the total inventory area of the covered glaciers, that is, 1684.7 km^2 (see Table 2). [b]The recall of the debris is computed using only the glaciers from Switzerland and with a debris coverage percentage larger than 1% (n = 288, total area = 594.45 km^2), based on the inventory from Linsbauer et al. (2021). The resulting total debris area is 56.04 km^2. Note that the band-ratio methods do not have the capability to retrieve debris-covered pixels so we do not report the recall for them. [c]The FP rate is computed by referring to the non-glacierized pixels within a 20–50 m buffer, which results in a total area of 262.4 km^2.

snow cover or different illumination—those same thresholds may no longer be optimal, resulting in lower recall and a consequent overestimation of shrinkage.

4.4.2. The Importance of Elevation Change Maps

As discussed in the Introduction, debris-cover glaciers remain one of the biggest challenges in fully automatic glacier mapping. To the best of our knowledge, no other DL study investigated the use of dh/dt maps as complementary input data. To show the benefit of including them, we re-trained the U-Net ensemble but discarded the dh/dt inputs. The results, summarized in Table 7, show that including the dh/dt improves the overall recall by around 1%, and by 8.8% when considering only the debris-covered areas, both while increasing the FP rate only by 0.6%.

One important issue with this approach, however, is that the dh/dt product does not have a perfect temporal match with the images—nor do the DEM and its derived features. While this may be less critical for the DEM, given that surface elevation does not change drastically over time (in relative terms), the temporal mismatch can be more problematic for the dh/dt input. Ideally, we would use a dh/dt product that (a) captures elevation changes over a shorter time period and (b) is temporally close to the image acquisition date—e.g., covering the current hydrological year or even a shorter interval, depending on vertical accuracy. This would ensure that the elevation change signal reflects the most recent spatial extent and condition of the glacier. Without such alignment, there is a risk that the dh/dt input reflects outdated glacier geometry or surface processes, which could introduce biases into the final area estimates. In addition, the spatial resolution of the dh/dt product (100 m) is substantially coarser than that of the Sentinel-2 optical imagery (10 m), which may lead to mismatches near glacier boundaries and reduce the effectiveness of this input in capturing small-scale features. However, high-resolution, short-term elevation change products are not (yet) available.

4.5. Computational and Data Scalability

4.5.1. Runtime and Computational Resources

The models were trained on one node from JUWELS Booster (hosted by Jülich Supercomputing Center), equipped with an AMD EPYC 7402 processor, 512 GB RAM and 4 NVIDIA A100 GPUs. Training a single model takes up to 1 hr. Since we use five geographic splits and train an ensemble of 10 models per split, this would total around 50 hr. However, by fitting two models per GPU in parallel, we reduced the effective training time to approximately 10 hr.

Once trained, inference on a single glacier using one model typically takes less than 2 s ($\mu = 1.79$ s), depending on glacier size—with the largest glaciers taking up to 1 min. Since we use an ensemble of 10 models, each glacier is processed 10 times per year, resulting in a total of 20 inferences per glacier (2015 and 2023). Across 1593 glaciers, 2 years, and 10 ensemble models, this results in 31,860 glacier-wide model evaluations. An additional 7,240 inferences were performed on glaciers from the validation folds of 2015, for calibration purposes. Although this would take approximately 19 hr if run sequentially, parallelization reduced total inference time to under 5 hr. Processing glaciers one at a time reduces GPU efficiency, particularly for small glaciers with only a few patches. We could further reduce the time by grouping patches across multiple glaciers, thus improving throughput by building larger batch sizes during inference.

4.5.2. Design Constraints and Data Volume

While DL4GAM achieves high-quality results and robust uncertainty estimates through deep ensembles, this comes with a non-negligible computational cost. In our case, the relatively small size of the study region kept this cost manageable, especially with parallelized training and inference. However, when scaling to larger or global glacierized regions, computational efficiency could become increasingly important. As an intermediate step, reducing the ensemble size to five members—which Ovadia et al. (2019) found to still yield robust uncertainty estimates under distribution shift—could reduce the computational cost, though the quality of the uncertainty estimates would need to be carefully reassessed. Future work may explore more scalable alternatives to ensembles, for example, deterministic uncertainty estimation methods that provide uncertainty estimates from a single model.

5. Conclusions

This study demonstrates the potential of deep learning techniques for accurately and efficiently monitoring glaciers, at individual level. The DL4GAM framework is based on an ensemble of ten U-Net models, which are trained and tested using a five-fold geographic cross-validation scheme. This has the advantage that we can concatenate the predictions on the testing folds and use them in the change analysis, which minimizes the biases toward the training set. In addition to validating the model against the inventory used for training, we applied DL4GAM on a small set of glaciers from the Swiss glaciers inventory (SGI2016). We showed that our results align well with their round-robin experiment, demonstrating high accuracy in the estimated areas and reliable uncertainty estimates. Once the models are trained on the 2015 data, they are applied on the most recent Sentinel-2 images, from 2023. These images are automatically selected using a procedure that aims to minimize both the cloud coverage and the seasonal snow to ensure good mapping conditions. Finally, we provide annual area change rates over 2015–2023 for ca. 900 glaciers, covering around 70% of the region. Based on these, we estimate a regional change rate of $-1.90 \pm 1.26\%$ per year, with significant inter-glacier variability. We also compared the DL4GAM regional estimates with those obtained based on the band-ratio thresholding (therefore capturing only the clean ice), showing that the latter would overestimate the glacier area shrinkage rates even more, probably caused by the increasing debris coverage. The potential for using our results (outlines or change rates) in geodetic mass balance calculations depends on the scale of analysis. While we are confident in the regional estimates, as systematic errors tend to balance out over large areas, their applicability at the sub-regional scale remains promising but requires further validation. At the individual glacier level, the variability is higher, and only results that pass quality control can be considered reliable. Future work could explore whether mass balance estimates improve by using these quality-controlled outlines while applying an average change rate to the remaining glaciers.

A few challenges remain unresolved and require additional study. First, our models still struggle to accurately identify debris-covered glacier regions, even after incorporating the elevation change maps as inputs, leading to the exclusion of almost 40% of the glaciers in our quality control process. While our final estimates still cover a significant portion of the glacierized area in the region, ideally, we aim to monitor every individual glacier. Second, the effectiveness of the dh/dt input is limited by both its coarse spatial resolution and temporal mismatch with the imagery, which may introduce biases in the estimated glacier extents. Future work could investigate the impact of these limitations more systematically and conduct a more detailed ablation study to evaluate the performance of the method without such inputs, or in regions characterized by minimal elevation change. Third, regarding the assumption made during the image acquisition phase—that at least one image with good mapping conditions exists within the specified period—we aim to explore techniques for evaluating the suitability of automatically selected images, and inform the user when none is available. In addition to addressing these limitations, as future work we also plan to extend our study to additional regions to assess whether the models generalize effectively or require further training. In the latter case, progress will be constrained by the availability of training labels which have to be dated post-2015, unless we rely on different sensors. Finally, the current use of deep ensembles for uncertainty estimation, while effective, may become computationally demanding when scaling to larger regions or operational applications. Future work could explore lighter-weight alternatives that estimate uncertainty directly from a single model, offering a more scalable solution. In addition, we aim to investigate whether incorporating expert-derived uncertainty estimates (e.g., inter-annotator variability) into the calibration process could improve the interpretation of model uncertainty, particularly in cases where the reference outlines are themselves imperfect.

To summarize, our method serves as an intermediate solution for glacier monitoring, enabling reliable estimates of regional glacier area change and, where quality thresholds are met, glacier-wide change rates. While the analysis is currently limited to glaciers larger than 0.1 km^2 and those passing our uncertainty-based filtering, the approach supports near-real-time updates with minimal manual effort.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The processed data set is available through HuggingFace: https://huggingface.co/datasets/dcodrut/dl4gam_alps (Diaconu et al., 2025a). The source code for processing the data and generate the results is available on GitHub (https://github.com/dcodrut/dl4gam_alps) or Zenodo (Diaconu et al., 2025b). The predicted outlines can be visualized at https://dcodrut.github.io/dl4gam_alps/ and downloaded as shapefiles from https://huggingface.co/datasets/dcodrut/dl4gam_alps/tree/main/outlines. The weights of the trained models can also be found on HuggingFace (https://huggingface.co/dcodrut/dl4gam_alps). The raw data is also openly available from the original sources: the Sentinel-2 imagery is provided by the European Union/ESA/Copernicus through for example, Google Earth Engine (Gorelick et al., 2017) (we used `geedim` (Leftfield Geospatial, 2021)); the dh/dt maps from Hugonnet et al. (2021) are available at https://doi.org/10.6096/13; the Copernicus GLO-30 DEM is available at https://doi.org/10.5069/G9028PQB.

Acknowledgments

Codruț-Andrei Diaconu is supported by the Helmholtz Association through the joint research school Munich School for Data Science - MuDS (Grant HIDSS-0006). Harry Zekollari received funding from the European Research Council (ERC) under the European Union's Horizon Framework research and innovation programme - "ICE^3" project (Grant 101115565) and from the research foundation—Flanders (FWO) through an Odysseus Type II project - "GlaciersMD" (Grant G0DCA23N). Jonathan Bamber was supported by the European Union's Horizon 2020 research and innovation programme through the project Arctic PASSION (Grant 101003472) and the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (Grant 01DD20001). We gratefully acknowledge the computing time granted by the Helmholtz Association's Initiative and Networking Fund on the HAICORE@FZJ partition. We would like to thank Frank Paul and an anonymous reviewer for their constructive comments, which greatly improved the clarity and quality of this paper. We are also grateful to Prof. Kristy Tiampo for coordinating the review process. Open Access funding enabled and organized by Projekt DEAL.

References

Berthier, E., Floricioiu, D., Gardner, A. S., Gourmelen, N., Jakob, L., Paul, F., et al. (2023). Measuring Glacier mass changes from space–A review. *Reports on Progress in Physics*, *86*(3), 036801. https://doi.org/10.1088/1361-6633/acaf8e

Bojinski, S., Verstraete, M., Peterson, T. C., Richter, C., Simmons, A., & Zemp, M. (2014). The concept of essential climate variables in support of climate research, applications, and policy. *Bulletin of the American Meteorological Society*, *95*(9), 1431–1443. https://doi.org/10.1175/BAMS-D-13-00047.1

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Copernicus - ESA. (2023). COP-DEM_GLO-30-DGED. https://doi.org/10.5270/ESA-c5d3d65

Diaconu, C.-A., Heidler, K., Bamber, J. L., & Zekollari, H. (2025). Chapter 13 - Multi-sensor deep learning for glacier mapping. In S. Saha (Ed.), *Deep learning for multi-sensor Earth observation* (pp. 287–333). https://doi.org/10.1016/B978-0-44-326484-9.00024-5

Diaconu, C. A., Zekollari, H., & Bamber, J. L. (2025a). DL4GAM: A multi-modal Deep Learning-based framework for Glacier Area Monitoring, trained and validated on the European Alps (Version 644cb0e) [Dataset]. *Hugging Face*. https://doi.org/10.57967/hf/6289

Diaconu, C. A., Zekollari, H., & Bamber, J. L. (2025b). DL4GAM: A multi-modal Deep Learning-based framework for Glacier Area Monitoring, trained and validated on the European Alps (Version v1.0.2) [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.16809567

Edwards, T. L., Nowicki, S., Marzeion, B., Hock, R., Goelzer, H., Seroussi, H., et al. (2021). Projected land ice contributions to twenty-first-century sea level rise. *Nature*, *593*(7857), 74–82. https://doi.org/10.1038/s41586-021-03302-y

Falcon, W., & The PyTorch Lightning team. (2019). PyTorch lightning. https://doi.org/10.5281/zenodo.3828935

Florentine, C., Sass, L., McNeil, C., Baker, E., & O'Neel, S. (2023). How to handle glacier area change in geodetic mass balance. *Journal of Glaciology*, *69*(278), 1–7. https://doi.org/10.1017/jog.2023.86

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., et al. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, *56*(S1), 1513–1589. https://doi.org/10.1007/s10462-023-10562-9

GLAMOS. (2024). The Swiss glaciers 1880-2018/19, glaciological reports no 1-140, yearbooks of the cryospheric Commission of the Swiss Academy of Sciences (SCNAT), published since 1964 by VAW/ETH Zurich. https://doi.org/10.18752/glrep_series

Global Climate Observing System (GCOS). (2022). *Implementation Plan for the Global Observing System for Climate: GCOS-245 (2025th ed.; Tech. Rep. No. WMO-No. 1267)*. World Meteorological Organization. (see p. 211 for the 5% area-change threshold).

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

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321–1330).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition* (pp. 770–778). CVPR. https://doi.org/10.1109/CVPR.2016.90

Herreid, S., & Pellicciotti, F. (2020). The state of rock debris covering Earth's glaciers. *Nature Geoscience*, *13*(9), 621–627. https://doi.org/10.1038/s41561-020-0615-0

Hock, R., & Huss, M. (2021). Chapter 9 - Glaciers and climate change. In T. M. Letcher (Ed.), *Climate change* (3rd ed., Vol. 176, p. 157). https://doi.org/10.1016/B978-0-12-821575-3.00009-8

Horn, B. (1981). Hill shading and the reflectance map. *Proceedings of the IEEE*, *69*(1), 14–47. https://doi.org/10.1109/PROC.1981.11918

Hugonnet, R., McNabb, R., Berthier, E., Menounos, B., Nuth, C., Girod, L., et al. (2021). Accelerated global glacier mass loss in the early twenty-first century. *Nature*, *592*(7856), 726–731. https://doi.org/10.1038/s41586-021-03436-z

Iakubovskii, P. (2019). Segmentation models pytorch. *GitHub*. Retrieved from https://github.com/qubvel/segmentation_models.pytorch

Immerzeel, W. W., Lutz, A. F., Andrade, M., Bahl, A., Biemans, H., Bolch, T., et al. (2020). Importance and vulnerability of the World's water towers. *Nature*, *577*(7790), 364–369. https://doi.org/10.1038/s41586-019-1822-y

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd international conference for learning representations (ICLR)*. https://doi.org/10.48550/arXiv.1412.6980

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *No. arXiv: 1612.01474)*. https://doi.org/10.48550/arXiv.1612.01474

Leftfield Geospatial. (2021). geedim. Retrieved from https://geedim.readthedocs.io/en/stable/

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826

Linsbauer, A., Huss, M., Hodel, E., Bauder, A., Fischer, M., Weidmann, Y., et al. (2021). The New Swiss Glacier Inventory SGI2016: From a topographical to a glaciological dataset. *Frontiers in Earth Science*, *9*, 704189. https://doi.org/10.3389/feart.2021.704189

Maslov, K. A., Persello, C., Schellenberger, T., & Stein, A. (2025). Globally scalable glacier mapping by deep learning matches expert delineation accuracy. *Nature Communications*, *16*(1), 43. https://doi.org/10.1038/s41467-024-54956-x

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. H. S., & Dokania, P. K. (2020). Calibrating deep neural networks using focal loss. In *Proceedings of the 34th international conference on neural information processing systems* (pp. 15288–15299).

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., et al. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, *32*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Paul, F., Barrand, N. E., Baumann, S., Berthier, E., Bolch, T., Casey, K., et al. (2013). On the accuracy of glacier outlines derived from remote-sensing data. *Annals of Glaciology*, *54*(63), 171–182. https://doi.org/10.3189/2013AoG63A296

Paul, F., Frey, H., & Le Bris, R. (2011). A new glacier inventory for the European Alps from Landsat TM scenes of 2003: Challenges and results. *Annals of Glaciology*, *52*(59), 144–152. https://doi.org/10.3189/172756411799096295

Paul, F., Rastner, P., Azzoni, R. S., Diolaiuti, G., Fugazza, D., Le Bris, R., et al. (2020). Glacier shrinkage in the Alps continues unabated as revealed by a new glacier inventory from Sentinel-2. *Earth System Science Data*, *12*(3), 1805–1821. https://doi.org/10.5194/essd-12-1805-2020

Paul, F., Winsvold, S. H., Kääb, A., Nagler, T., & Schwaizer, G. (2016). Glacier remote sensing using Sentinel-2. Part II: Mapping Glacier extents and surface facies, and comparison to Landsat 8. *Remote Sensing*, *8*(7), 575. https://doi.org/10.3390/rs8070575

Peng, Y., He, J., Yuan, Q., Wang, S., Chu, X., & Zhang, L. (2023). Automated glacier extraction using a Transformer based deep learning approach from multi-sensor remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *202*, 303–313. https://doi.org/10.1016/j.isprsjprs.2023.06.015

Rajat, S., Singh, B. R., Prakash, C., & Anita, S. (2022). Glacier retreat in Himachal from 1994 to 2021 using deep learning. *Remote Sensing Applications: Society and Environment*, *28*, 100870. https://doi.org/10.1016/j.rsase.2022.100870

Riley, S., Degloria, S., & Elliot, S. (1999). A terrain ruggedness index that quantifies topographic heterogeneity. *Internation Journal of Science*, *5*, 23–27.

Roberts-Pierel, B. M., Kirchner, P. B., Kilbride, J. B., & Kennedy, R. E. (2022). Changes over the last 35 years in Alaska's glaciated landscape: A novel deep learning approach to mapping glaciers at fine temporal granularity. *Remote Sensing*, *14*(18), 4582. https://doi.org/10.3390/rs14184582

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *MICCAI* (Vol. 9351, pp. 234–241). https://doi.org/10.1007/978-3-319-24574-4_28

Rounce, D. R., Hock, R., McNabb, R. W., Millan, R., Sommer, C., Braun, M. H., et al. (2021). Distributed global debris thickness estimates reveal debris significantly impacts glacier mass balance. *Geophysical Research Letters*, *48*(8), e2020GL091311. https://doi.org/10.1029/2020GL091311

Salim, E. (2023). Glacier tourism without ice: Envisioning future adaptations in a melting world. *Frontiers in Human Dynamics*, *5*, 1137551. https://doi.org/10.3389/fhumd.2023.1137551

swisstopo, F. O. o. T. (2024a). swissALTI3D, the high precision digital elevation model of Switzerland. Retrieved from https://www.swisstopo.admin.ch//en/height-model-swissalti3d

swisstopo, F. O. o. T. (2024b). SWISSIMAGE, the digital color Orthophotomosaic of Switzerland. Retrieved from https://www.swisstopo.admin.ch/en/orthoimage-swissimage-10

Thomas, D. J., Robson, B. A., & Racoviteanu, A. E. (2023). An integrated deep learning and object-based image analysis approach for mapping debris-covered glaciers. *Frontiers in Remote Sensing*, *4*, 1161530. https://doi.org/10.3389/frsen.2023.1161530

Tian, S., Dong, Y., Feng, R., Liang, D., & Wang, L. (2022). Mapping mountain glaciers using an improved U-Net model with cSE. *International Journal of Digital Earth*, *15*(1), 463–477. https://doi.org/10.1080/17538947.2022.2036834

Tollenaar, V., Zekollari, H., Pattyn, F., Rußwurm, M., Kellenberger, B., Lhermitte, S., et al. (2024). Where the white continent is blue: Deep learning locates bare ice in Antarctica. *Geophysical Research Letters*, *51*(3), e2023GL106285. https://doi.org/10.1029/2023GL106285

Wang, Z., Erb, A. M., Schaaf, C. B., Sun, Q., Liu, Y., Yang, Y., et al. (2016). Early spring post-fire snow albedo dynamics in high latitude boreal forests using Landsat-8 OLI data. *Remote Sensing of Environment*, *185*, 71–83. https://doi.org/10.1016/j.rse.2016.02.059

Wimmer, L., Sale, Y., Hofman, P., Bischl, B., & Hüllermeier, E. (2023). Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Proceedings of the thirty-ninth conference on uncertainty in artificial intelligence* (pp. 2282–2292).

xdem contributors. (2021). xDEM: Robust analysis of DEMs in Python. *Zenodo*. https://doi.org/10.5281/zenodo.4809698

Xie, Z., Asari, V. K., & Haritashya, U. K. (2021). Evaluating deep-learning models for debris-covered glacier mapping. *Applied Computing and Geosciences*, *12*, 100071. https://doi.org/10.1016/j.acags.2021.100071

Xie, Z., Haritashya, U. K., Asari, V. K., Bishop, M. P., Kargel, J. S., & Aspiras, T. H. (2022). GlacierNet2: A hybrid multi-model learning architecture for alpine glacier mapping. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102921. https://doi.org/10.1016/j.jag.2022.102921

Xie, Z., Haritashya, U. K., Asari, V. K., Young, B. W., Bishop, M. P., & Kargel, J. S. (2020). GlacierNet: A deep-learning approach for debris-covered glacier mapping. *IEEE Access*, *8*, 83495–83510. https://doi.org/10.1109/ACCESS.2020.2991187

Zevenbergen, L. W., & Thorne, C. R. (1987). Quantitative analysis of land surface topography. *Earth Surface Processes and Landforms*, *12*(1), 47–56. https://doi.org/10.1002/esp.3290120107