

MULTI-SENSOR DEEP LEARNING FOR GLACIER MAPPING

Codruț-Andrei Diaconu^{1, 2, *}, Konrad Heidler², Jonathan L. Bamber^{2, 3}, Harry Zekollari^{4, 5, 6}

¹Earth Observation Center, German Aerospace Center (DLR), Germany

²School of Engineering and Design, Technical University of Munich, Germany

³Bristol Glaciology Centre, University of Bristol, United Kingdom

⁴Department of Water and Climate, Vrije Universiteit Brussel, Belgium

⁵Laboratory of Hydraulics, Hydrology and Glaciology (VAW), ETH Zurich, Switzerland

⁶Laboratoire de Glaciologie, Université Libre de Bruxelles, Belgium

*codrut-andrei.diaconu@dlr.de

ABSTRACT

The more than 200,000 glaciers outside the ice sheets play a crucial role in our society by influencing sea-level rise, water resource management, natural hazards, biodiversity, and tourism. However, only a fraction of these glaciers benefit from consistent and detailed in-situ observations that allow for assessing their status and changes over time. This limitation can, in part, be overcome by relying on satellite-based Earth Observation techniques. Satellite-based glacier mapping applications have historically mainly relied on manual and semi-automatic detection methods, while recently, a fast and notable transition to deep learning techniques has started.

This chapter reviews how combining multi-sensor remote sensing data and deep learning allows us to better delineate (i.e. map) glaciers and detect their temporal changes. We explain how relying on deep learning multi-sensor frameworks to map glaciers benefits from the extensive availability of regional and global glacier inventories. We also analyse the rationale behind glacier mapping, the benefits of deep learning methodologies, and the inherent challenges in integrating multi-sensor earth observation data with deep learning algorithms.

While our review aims to provide a broad overview of glacier mapping efforts, we highlight a few setups where deep learning multi-sensor remote sensing applications have a considerable potential added value. This includes applications for debris-covered and rock glaciers that are visually difficult to distinguish from surroundings and for calving glaciers that are in contact with the ocean. These specific cases are illustrated through a series of visual imageries, highlighting some significant advantages and challenges when detecting glacier changes, including dealing with seasonal snow cover, changing debris coverage, and distinguishing glacier fronts from the surrounding sea ice.

Keywords

Earth Observation · Glacier Extent Mapping · Calving Front Detection · Deep Learning · Multisensor · Multimodal

1 Introduction

There are ca. 275,000 glaciers around the world (RGI 7.0 Consortium, 2023), which act as important contributors to sea-level rise (Edwards et al., 2021; Hugonet et al., 2021), water resources (Immerzeel et al., 2020), triggers of natural hazards (Veh et al., 2023), biodiversity regulators (Bosson et al., 2023), and touristic attractions (Salim, 2023). Around 0.1 percent of these glaciers have a reasonably long record of in-situ observations of mass change (WGMS, 2024; Zemp et al., 2015). Therefore, satellite-based Earth Observation (EO) is the only feasible way to obtain representative

This article will be a chapter of the book *Deep Learning for Multi-Sensor Earth Observation*, to be published by Elsevier.

sampling across the broad latitudinal and altitudinal range occupied by glaciers. Glacier-specific observations are essential because i) individual glaciers can respond to the climate in complex ways that depend on their geometry, morphology and other boundary conditions (Brun et al., 2019) while ii) as an ensemble, glaciers are valuable and unique indicators of integrated climate change over multi-decadal to centennial timescales (Marzeion et al., 2018). Numerous satellite missions and sensors can, and have been, used to measure glacier changes, often with conflicting results (Gardner et al., 2013; Wouters et al., 2019; Zemp et al., 2019).

Machine Learning (ML), in particular multi-modal approaches based on Deep Learning (DL), offer tremendous potential. These methods can extract information, patterns and trends that have proven to be challenging to obtain through "conventional" approaches. While DL methods in EO, in particular for land classification, are well-developed fields with a rich history, application of DL methods in glacier (change) detection only recently started to emerge in the literature. Despite being a recent field of research, the application of DL for glacier mapping has rapidly progressed. We believe this rapid evolution is, in part, due to the following reasons: i) the glacier mapping task is mainly a segmentation task, and, as a consequence, many Computer Vision methods are directly applicable (Xie et al., 2021); ii) there are many existing regional glacier inventories, including a global one — Randolph Glacier Inventory (RGI) (RGI 1.0 Consortium, 2012; RGI 7.0 Consortium, 2023) — which provide the necessary training labels, and iii) the level of complexity of glacier mapping is less than for other glacier-related problems (e.g. see Section 5.2 where we also shed light on DL developments related to glacier evolution modelling) where the input-output relations are more complicated, often requiring additional physical constraints, and typically suffering from a lack of extensive, high-quality training data.

Here, we discuss how recent glacier mapping efforts build on segmentation algorithms, emphasizing its particular challenges. We specifically focus on the glaciers outside the ice sheets while also including ice-sheet outlet glaciers.

Motivation

A review of Deep Learning applications for cryospheric studies, including, e.g. glaciers, ice sheets, permafrost, and snow, is provided by Liu (2021). Their review provides a broad perspective on the field, focusing on a few selected studies. Here, we instead focus on a single application area, i.e. glacier mapping, which is rapidly growing, thereby developing maturity from a DL perspective.

To our knowledge, this is the first study that provides an overview of the glacier mapping literature based on DL. Here, we describe some of the significant studies published so far and summarize which data sources have been used, highlighting that glacier mapping is, to a large extent, a multi-modal task.

Structure

In Section 2 we provide an introduction for glacier mapping: we explain why this is an essential problem in cryospheric studies, then discuss the benefits of relying on DL and finally provide an overview of the associated challenges. Section 3 is dedicated to the data modalities used in the studies included in our literature review (Section 1), which covers two major topics in glacier mapping, detecting i) the full extent of a glacier (Section 4.1) and ii) calving fronts (Section 4.2). Next, we provide a discussion in Section 5 followed by a summary (Section 6). Additionally, in Section 7 we include a list of resources (mainly databases) that could be exploited with DL.

2 Glacier Mapping with Deep Learning

Glacier Mapping

To study various properties of glaciers, e.g. area, hypsometry, we first need to know where they are located. Various regional glacier inventories have been produced, leading to the first global glacier inventory, the RGI (RGI 1.0 Consortium, 2012) that was initially derived from Global Land Ice Measurements from Space (GLIMS), which is a multi-temporal glacier database. The RGI was initially developed as part of the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (Stocker et al., 2013) and was designed to be a snapshot of all the glaciers in the world at the beginning of the 21st century. There are, however, significant variations among the subregions, as RGI is a compilation of regional inventories from various sources, most of which are based on satellite imagery from 1999–2010. In the latest version (RGI 7.0 Consortium, 2023), the RGI contains close to 275,000 glaciers, with a minimum area of 0.01 km², covering a total surface of ~707,000 km², with an estimated ±5% error. Using RGI, we

can derive the glacier area distribution with elevation, an essential input in many glacier evolution models (Marzeion et al., 2020; Zekollari et al., 2022), thus making RGI a critical dataset, being constantly enhanced.

Glaciers have been defined by Bojinski et al. (2014), under the Global Climate Observing System, as an Essential Climate Variable (ECV), given that glaciers are sensitive and unique indicators of climate change (Hock et al., 2021). Glacier area, elevation change and mass change/balance are listed as ECV products. Various studies have shown a significant glacier retreat in different parts of the world. Such studies usually imply (re)creating the glacier outlines at two different points in time and analyzing the differences. For instance, Paul et al. (2020) rebuilt a glacier inventory for the European Alps and compared the total glaciated area with the previous inventory (from RGI), showing a regional area change rate of ca. -1.2% per year from 2003 to 2015. A more recent study by Ali et al. (2023) used Object-Based Image Analysis (OBIA) to produce the outlines of ca. 2200 glaciers in the Arctic from four different regions at three time periods: 1985–1989, 2000–2002, and 2019–2021. Their results show an accelerated loss in the second period, e.g. from $-36.1 \text{ km}^2 \text{ yr}^{-1}$ to $-41 \text{ km}^2 \text{ yr}^{-1}$ for Novaya Zemlya and from $-3.7 \text{ km}^2 \text{ yr}^{-1}$ to $-8 \text{ km}^2 \text{ yr}^{-1}$ in Kenai. They also show that 73 glaciers have entirely disappeared. When imposing the same ice divides on the glacier outlines, each glacier can be individually tracked over time, allowing to capture the heterogeneity caused by the complex interactions between climate and local topography. Tarca et al. (2023) tracked five small Alpine glaciers individually over 2017–2021, revealing an annual area loss rate between 0.98% and 3.4%. Since glaciers leave behind distinct landforms (e.g. moraines) as they move across the landscape, various studies reconstructed the glacier extents during the Little Ice Age (an advanced glacier position, typically in the second half of the 19th century) and compared these to the present state, showing that glaciers decreased in size in all regions, with some glaciers that disappeared (Parkes et al., 2018; Paul et al., 2019).

Glacier mapping is also an essential task due to its direct link with Mass Balance (MB) estimation. Glacier MB¹ is 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 (usually a year or a season). One way to estimate the MB of the glacier is through Digital Elevation Models (DEMs) differencing: two DEMs are co-registered, and then the elevation difference (i.e. volume change) is converted into mass change by making certain density assumptions. This technique is usually referred to as the geodetic method (Berthier et al., 2023), the resulting geodetic MBs being an essential indicator of the glacier status. The dataset from Hugonnet et al. (2021) provides geodetic MBs with global coverage over the 2000–2019 period, based on DEM time-series from ASTER. When two (or more) DEMs are differenced to estimate the elevation change, it is also important to consider the potential change in the glacier area. When estimating the specific MB (i.e. the total mass balance divided by the area), many studies assume no area change and, therefore, use a single-dated glacier outline (Berthier et al., 2023). However, this can induce significant biases, especially when the time gap between the two DEMs acquisitions is large, and the glaciers retreated significantly between the two DEMs considered. A recent study focused on five North American glaciers shows that a fixed area assumption can cause a MB underestimation of up to 19% (Florentine et al., 2023). Recent global studies on MB estimation (Hugonnet et al., 2021; Zemp et al., 2019) have used regional correction factors to account for glacier area changes. Ideally, this correction should be performed independently for each glacier, but this would require inventories that temporarily match the DEMs acquisitions, which are usually unavailable at large scale. In general, in geodetic MB studies, the glacier outlines are updated for relatively small regions, e.g. the Swiss Alps (Mannerfelt et al., 2022), the European Alps (Sommer et al., 2020), a subregion from Svalbard (Geyman et al., 2022), Northern & Southern Patagonian Icefields (Abdel Jaber et al., 2019). In contrast, for large regions, the outlines are typically static, e.g. for High-Mountain Asia (HMA) (Brun et al., 2017; Shean et al., 2020) or the Andes (Braun et al., 2019; Dussaillant et al., 2019). Similar glacier outline mismatches can also occur for in-situ MB estimation when point-wise measurements have to be extrapolated to the entire glacier, which, therefore, also requires updating glacier outlines (Huss et al., 2012).

Earth's large polar ice sheets in Antarctica and Greenland contain massive glaciers, storing the majority of global freshwater (Siegert, 2005). Besides most alpine glaciers, these glaciers are usually marine-terminating, meaning they terminate into the ocean, calving off icebergs. Fundamental differences such as the presence of open water and sea ice set apart mapping these glaciers from glacier mapping in lower latitudes. Therefore, calving front detection is often categorized separately from alpine glacier mapping, with the recent work on global glacier mapping by Maslov et al. (2024) suggesting the use of dedicated approaches for mapping the calving fronts. Consequently, we will treat calving front detection in this chapter.

¹For a more comprehensive set of definitions of glacier-related terms, see Cogley et al. (2011).

Deep Learning

Previous studies have shown that DL can provide a significant performance improvement compared to classical approaches for a wide range of EO tasks, including for instance, land-cover classification, vegetation parameters estimation (e.g. height, biomass) and precipitation down-scaling or now-casting (Yuan et al., 2020). Consequently, DL was also adopted in cryospheric studies, e.g. for sea-ice concentration forecasting (Andersson et al., 2021), glacier evolution modelling (Bolibar et al., 2022; Jouve et al., 2022), modelling the ice thickness of Antarctica (Leong et al., 2020), estimating the mass balance of ice sheet and its various components (Roda Husman et al., 2024; Van Der Meer et al., 2023) and detecting blue ice in Antarctica (Tollenaar et al., 2024), to name a few examples.

A particular area of cryospheric sciences where DL is rapidly evolving is the domain of glacier mapping. First, given the large number of glaciers, fully automating the task of mapping glacier outlines is of high interest. Historically, semi-automatic methods were used for glacier mapping, which usually consist of thresholding band ratios, e.g. Red / Short-Wave InfraRed (SWIR), as exemplified in Figure 1, or the Normalized Difference Snow Index (NDSI). This thresholding step is usually followed by manual corrections, especially for debris-covered glacier parts or those subject to shadowing (Paul et al., 2002, 2020). Similarly, rough estimates of calving fronts can be obtained from thresholding reflectance or backscatter values, as the ocean will appear much darker than ice and snow in imagery. However, such approaches are easily confounded by the presence of sea ice (Liu et al., 2004). A particular difficulty of these semi-automatic methods relates to the fact that, in many cases, thresholds need to be calibrated to each different scene, e.g. to account for varying atmospheric conditions or to capture shadowing effects. From this perspective, DL has the added value of potentially exploiting local features and automatically finding the appropriate threshold. Second, even though DL models have associated errors, one can assume that these errors are systematic in space and time (assuming an in-distribution testing scenario). In contrast, the quality of the manual glacier outlines can significantly vary depending on the individuals ('experts') performing the glacier labelling (Paul et al., 2013). To reduce these errors related to manual labelling in newer inventories, a preparation phase is usually implemented to ensure a consistent and homogeneous quality among the human annotators (Linsbauer et al., 2021). Third, DL can learn from multiple data sources simultaneously, which is particularly helpful for more complex classification cases. Such an example is classifying debris-covered glaciers (Xie et al., 2020) that are visually very difficult to distinguish from their (non-glaciated) surroundings, which classical band-ratio methods can only (partly) circumvent through time-consuming and error-prone manual corrections (Paul et al., 2020).

Despite the enormous potential of DL to be applied in the field of glacier mapping, some substantial challenges exist depending on the scientific question of interest. First, despite continuous improvements, the glacier inventories needed for training DL architectures suffer from considerable uncertainties. Although DL can still deal with "noisy" labels to a certain extent and still learn the underlying patterns (Arpit et al., 2017; Zlateski et al., 2018), these uncertainties certainly affect the performance metrics used when evaluating the quality of the predictions, making it more challenging to compare different methods. Additionally, the uncertainties in the inventories also vary from one region to another, thus potentially introducing biases when training global models. Moreover, in many cases, it is difficult (and often even entirely impossible) to ensure a perfect temporal match between the (optical) input data and the glacier inventory, which adds to the uncertainties. Lastly, glacier mapping remains a challenging task even for experts, especially for debris-covered glaciers where the interpretation can be subjective, sometimes leading to errors in the order of 10%-20% for small glaciers (Paul et al., 2020). These limitations should be accounted for when developing adequate DL frameworks for glacier mapping, and associated uncertainties in DL model predictions should always be quantified (Maslov et al., 2024).

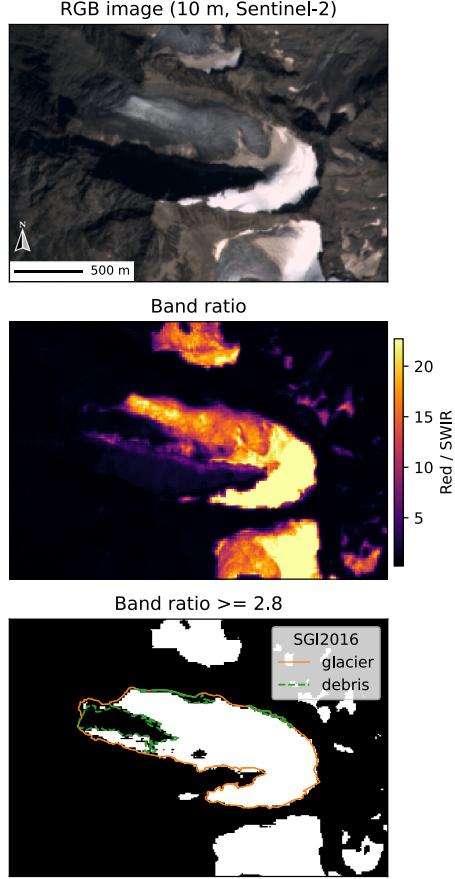


Figure 1: Band-ratio method. Example of the Band-ratio method being applied to Vadret da Misau, a glacier in Switzerland (46.42° N , 9.89° E). The upper panel uses the RGB bands from the Sentinel-2 acquisition on 26/08/2015 (courtesy of EU Copernicus program). We then apply the classical band-ratio method (central panel) using the Red and SWIR bands, with a threshold of 2.8, to delineate glacier areas (lower panel). This threshold was manually chosen to extract shadowed pixels. When comparing to the glacier & debris outlines from the Swiss Glacier Inventory (SGI) (Linsbauer et al., [2021]), we notice that the clean-ice and snow are accurately extracted, but this approach does not capture the debris-covered parts.

3 Data Modalities

This section summarises the data modalities exploited by various DL models from the literature, surveyed in Section 4, or used by experts when building glacier inventories.

Optical (multi-spectral) imagery

Optical data² is by far the most commonly used type of observation for glacier mapping, primarily because in many circumstances (e.g. for clean-ice glaciers), optical data can already support glacier delineation on its own, without the need of additional data sources. The importance of optical data relative to some other modalities (i.e. Synthetic Aperture Radar (SAR) backscatter intensities, Interferometric Synthetic Aperture Radar (InSAR) coherence, DEM, thermal imaging) was empirically evaluated using ablation studies in the works of Peng et al. (2023) and Maslov et al. (2024). Optical data alone helps building very accurate glacier outlines, especially for debris-free glaciers where the classical band-ratio method or NDSI thresholding yields robust results. The band-ratio method (e.g. Red / SWIR) can separate the very low spectral reflectance of ice and snow in the shortwave infrared versus the high reflectance in the visible spectrum (Paul et al., 2013, 2015). From a DL perspective, optical data offers many advantages: i) numerous data sources available, including many with open-access policies, e.g. Sentinel-2 or Landsat-9 (or older), ii) most of the

²Note that we here mainly refer to medium resolution imagery (10 to 30 m Ground Sample Distance (GSD))

DL architectures were designed for optical data (usually RGB), with many pre-trained models publicly available, and iii) it is relatively easy to visualize and understand this data, even for non-experts, either by extracting the RGB bands or by using false colour images. However, optical data comes with some disadvantages, i.e. i) many glaciers are often covered by clouds, limiting the amount of data, ii) optical data over glaciers often has decreased visibility due to shadows (either caused by clouds or surrounding topography) or suffers from the absence of sunlight (e.g. the case of glaciers in the polar areas during the Polar Night), iii) optical data is relatively sensitive to atmospheric interference, iv) optical data cannot, in general, distinguish between debris-covered glacier segments and the off-glacier surrounding topography, and v) typically data from the end of summer is needed to reduce the presence of seasonal/perennial off-glacier snow.

However, some of these disadvantages can be compensated by using Very High-Resolution (VHR) optical data, if available. In Figure 2 we show optical data over a single glacier at three different spatial resolutions, i.e. 30 m, 10 m and a VHR one, 25 cm. From this, we can, for instance, observe that the debris-covered segments become distinguishable in the VHR image due to the presence of crevasses, which are hardly visible in the other (lower resolution) products. Additionally, we can notice in the VHR data that the segments under shadow are still relatively visible compared to the other, lower-resolution sensors. Paul et al. (2013) compare the glacier outlines from multiple analysts on two resolutions, i.e. 30 m and 1 m, and suggest that the interpretation of debris-covered segments is mainly independent of the resolution. It is, however, hard to conclude from such studies whether DL models would perform differently, and we did not find any study that analyzes the role of spatial resolution from this perspective. None of the existing DL methods discussed in Section 4 has utilized VHR data, suggesting the limited availability and high costs of such high-resolution data. However, as the volume of (open) data continues to increase over time, we anticipate a growing number of studies exploring these avenues in the future.

Synthetic Aperture Radar (SAR)

Although not widely adopted yet, various types of SAR datasets have been explored in the glacier mapping literature. Compared to optical sensors, they offer various advantages: i) SAR uses longer wavelengths that have the potential to "see" through clouds, an important benefit for mountain regions; ii) being an active remote sensing method, it has day/night capability; iii) it can penetrate the snow (to some extent), which can be beneficial when transient snow obscures the boundaries of glaciers. SAR, however, also has some disadvantages, primarily due to the side-looking geometry, resulting in radar shadowing and foreshortening/layover, which are especially problematic in (steep) mountain regions. Additionally, for DL practitioners, SAR is more difficult to interpret than optical data and requires more intensive preprocessing. Lastly, SAR speckle patterns may cause additional challenges to the traditional DL architectures that are usually designed for optical (RGB) data (Zhu et al., 2021). An example of a SAR intensity image with its optical correspondent is displayed in Figure 3.

While optical data is often sufficient for debris-free glaciers to distinguish glacier outlines under cloud- and snow-free conditions, SAR can play an important role for the debris-covered segments and, by extension, for rock glaciers. First, SAR backscatter intensities correlate with the surface roughness, which can help distinguish the debris parts from the surrounding terrain. Second, by combining two (or more) SAR acquisitions, InSAR can reveal glacier motion or deformations, an indicator of glacier segments covered by debris or active rock glaciers. Previous studies have employed either InSAR displacement maps using the (unwrapped) interferogram or InSAR coherence (Maslov et al., 2024; Robson et al., 2020).

Digital Elevation Model

The surface elevation is usually provided as an additional input for glacier mapping efforts since it helps the models to extract further information related to topography, e.g. the glacier flow direction, which can be a valuable source of information, especially for calving glaciers. Additionally, using the elevation information, the DL models can "understand" where the glacier terminus is, which can be covered by debris as opposed to the accumulation areas. However, obtaining a large-scale DEM from satellite data requires specialized sensors and processing techniques, with additional challenges in mountain regions. For instance, two standard techniques are stereo photogrammetry and InSAR. If the former is affected by cloud and snow coverage or topographic shadowing (Hugonnet et al., 2021), the latter is challenged by the steep terrain and can yield significant ice-penetration biases (Berthier et al., 2023; Dehecq et al., 2016). Studies that rely on DEM for glacier mapping often need to account for a mismatch in timing between the DEM and the considered visual imagery. Despite this limitation, DEM information can still provide helpful information

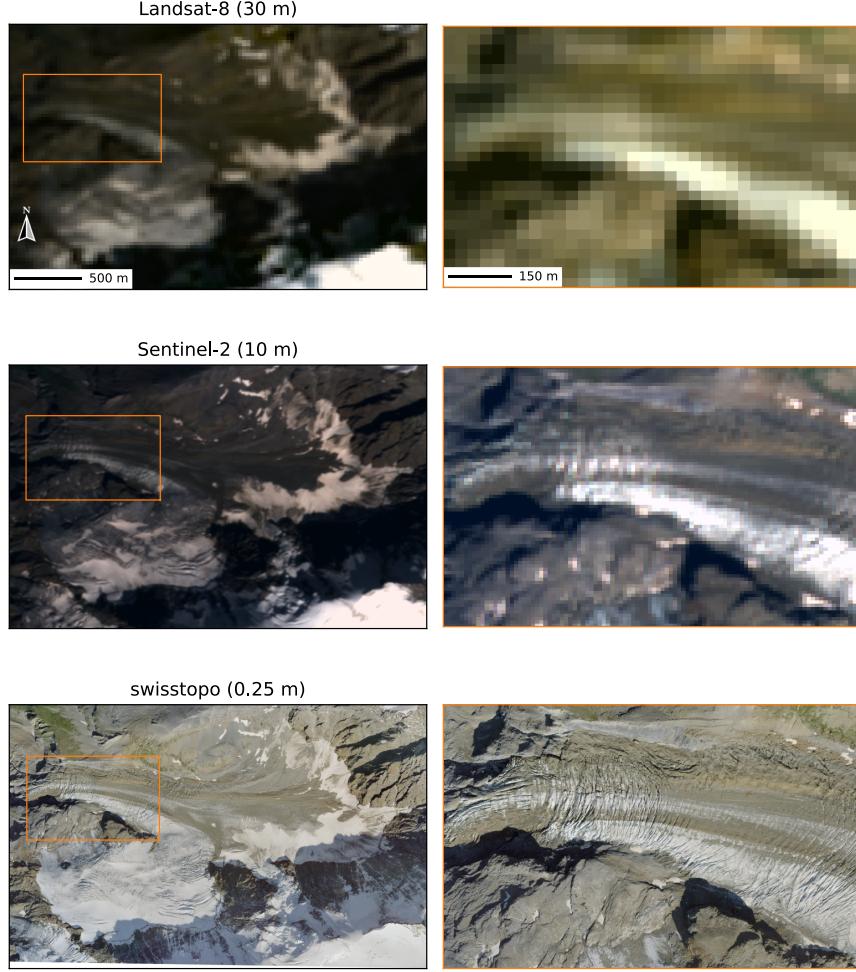


Figure 2: Spatial resolution comparison. The effect of spatial resolution on visual products is here illustrated for Rottalglletscher, a glacier in Switzerland (46.52° N, 7.95° E). We use the RGB bands from the Landsat-8 acquisition on 22/08/2018 (image courtesy of the U.S. Geological Survey), from Sentinel-2 on 20/08/2018 (courtesy of EU Copernicus program) and the VHR aerial image from swisstopo (2024b) (flight year = 2018). As resolution increases, an increasing level of detail can be observed, particularly pronounced for the debris-covered and shadowed glacier parts. The right panels provide a zoomed-in view of the glacier tongue, illustrating, among others, how the crevasses, a distinct feature of glaciers, become visible with increasing spatial resolution.

about the glacier topography and its surroundings. A few standard and openly available DEM choices are Copernicus GLO-30 DEM (Cop30DEM), Shuttle Radar Topography Mission (SRTM) DEM or its improved version - NASADEM, and ALOS World 3D - 30m (AW3D30). Figure 4 displays a DEM at two different spatial resolutions, i.e. 30m and 0.5m, for the tongue of the glacier from Figure 2. While both DEMs roughly capture the valley in which the glacier flows, crevasses become partially visible in the VHR, which can help to better identify the debris-covered parts. This comparison again suggests that spatial resolution could affect the models' performance.

When at least a pair of DEMs is available, it is possible to derive a surface elevation change map. For instance, a negative elevation change rate was observed over the last two decades (2000-2019) for almost all the glaciers in the world (Hugonnet et al., 2021). As a result, DEMs differences can help capture the glacier extent by contrasting it to the surrounding topography where typically no change is expected. This information can supplement image classification approaches, especially for debris-covered glaciers, as these are difficult to classify using optical data alone. However, the availability of global products is somewhat limited: the only global product that is based on the same data source (i.e. ASTER) from Hugonnet et al. (2021) comes at 100 m GSD. Moreover, for short timescales, high-precision DEMs are needed, both vertically and horizontally, to distinguish between potentially small glacier surface changes and off-glacier



Figure 3: Calving fronts - optical and SAR. Two calving fronts that belong to the Kangiata Nunaata Sermia and Akullersuup Sermia glaciers, in Greenland (64.33° N, -49.65° E), are observed using optical data (upper panel, RGB bands from the Sentinel-2 acquisition on 24/07/2023), and SAR data (lower panel, false-color composite extracted from the Level-1 Ground Range Detected of the Sentinel-1 acquisition on 06/07/2023). Courtesy of EU Copernicus program.

changes due to noise. In Figure 5 we highlight how a VHR elevation change map helps identify glaciers that are entirely covered by debris.

Ideally, we want to track glacier area change over time, e.g. at an annual to sub-decadal timescale, to capture how glaciers respond to changes in climatic conditions. Therefore, temporal resolution also plays a vital role in all the aforementioned data sources. One difficulty arises from the regions that undergo deglaciation early in the period spanned by two acquisitions: for these regions, which are of particular interest when tracking glacier areas over time, without a time-series, it is not feasible to identify the segments that become deglaciated early-on during the covered period, thereby hindering the analysis on the involved response times.

In summary, we have discussed various data sources for glacier mapping, each offering its pros and cons. This comparison emphasizes the potential of integrating multi-sensor data into DL models. As we explain in the next section’s literature overview, most studies utilize at least two data sources.

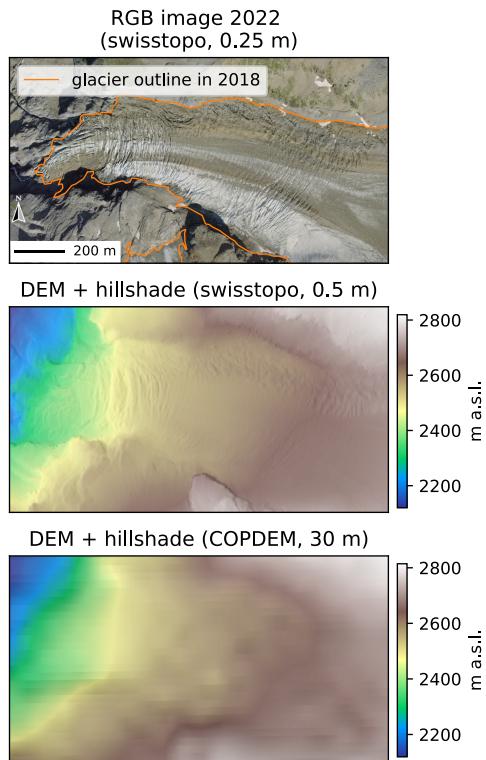


Figure 4: DEMs at two different spatial resolutions. The figure displays the tongue of the Rottalglletscher glacier in Switzerland (46.52° N, 7.94° E) (see also Figure 2). The DEMs are extracted from the swissALTI^{3D} DEM (swisstopo, 2024a) and the Copernicus GLO-30 DEM, and are here displayed with a superimposed shaded relief. While both DEMs capture the shape of the valley the glacier flows in, only the DEM with a sub-meter resolution captures smaller scale features such as crevasses.

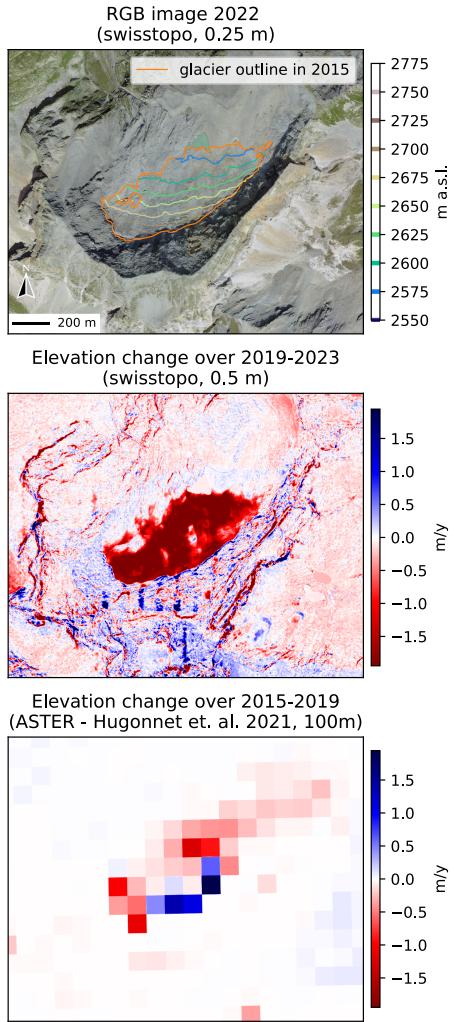


Figure 5: Differencing of DEMs. Here, we illustrate the role of resolution in DEM differencing for an entirely debris-covered glacier, Glatscher da Sut Fuina, in Switzerland (46.535° N, 9.473° E). While both DEM differences allow identifying the location of the glacier, the VHR version (central panel), based on two swissALTI^{3D} DEMs (swisstopo, [2024a]), is significantly more accurate. Note that for this VHR DEM differencing, the two DEMs were not co-registered before differentiation, which would allow for some artefacts to be removed. The DEM differencing map based on the lower-resolution DEMs (Hugonet et al., [2021]) (lower panel) is only able to roughly capture the location of the glacier, with some potential outliers (e.g., the significantly positive pixels). This figure illustrates that when using DEM differencing to detect glaciers, the role of the spatial resolution becomes important for relatively small glaciers.

4 Literature Overview

This section briefly reviews some of the key studies that employed DL models for automatic glacier delineation. Note that this is not an exhaustive review of all existing methods but rather a short overview of some of the most relevant and innovative works, emphasizing the particular challenges that have been addressed and shedding light on the obstacles that will require further research. Additionally, we highlight the various data sources used in each work to illustrate the importance of relying on multi-sensor approaches when mapping glaciers through DL. Given the significant differences in input data sources, labels, and/or considered regions, the evaluation scores mentioned throughout this section should not be overinterpreted or used to compare the performance of the various methods.

The section is structured as follows. First, we describe the works on glacier extent mapping in Section 4.1, with three sub-categories: i) standard methods, i.e. those that focus in general on glacier extent mapping with the primary goal of automatizing the process, ii) studies that perform glacier mapping on multiple acquisitions to quantify temporal glacier area changes and, iii) studies that map the extent of rock glaciers, which we treat as a separate category given their significant differences compared to typical glaciers, thereby usually requiring more specialized methodologies. The various studies and corresponding methods are then summarized in Tables 1 to 3 respectively. The second part covers the works on calving front detection, which are then summarized in Table 4. Note that most of the paragraphs cover a single study, with some exceptions where follow-ups are also included.

4.1 Glacier Extent Mapping

4.1.1 Glacier Extent Mapping - Standard Methods

The methods included in the following paragraphs treat the glacier extent mapping problem as a single-image segmentation task and propose various modifications to existing DL architectures, usually with the goal of improving the performance for the debris-covered parts of the glacier which remain hard to detect. One common characteristic of these methods is their data fusion capability: as highlighted in Table 1, most works use at least two data sources, with optical data being the main one.

GlacierNet (Xie et al., 2020) The first work that employs a DL architecture is GlacierNet (Xie et al., 2020). Whereas previous studies that used ML relied on classical approaches, e.g. support vector machine (SVM), random forest (RF) or shallow networks like multi-layer perceptrons (MLPs) (Khan et al., 2020; Zhang et al., 2019b), GlacierNet proposes a Convolutional Neural Network (CNN) architecture built upon the fully convolutional model SegNet (Badrinarayanan et al., 2017). The main purpose of the work was to address the challenge of detecting debris-covered glaciers. It was trained using all eleven Landsat 8 bands and the AW3D30 DEM, from which additional features were derived, i.e. slope angle, slope-azimuth divergence index (SADI), profile curvature, tangential curvature and unsphericity curvature. The study focuses on two sub-regions from HMA with different glacier conditions and properties: Nepal Himalaya and the central Karakoram in Pakistan. The glacier boundaries used for training were obtained from GLIMS (Raup et al., 2007) and further modified i) by improving the termini delineations, which can be at a different position due to the time mismatch between the imagery and the inventory, and ii) by removing the snow-covered accumulation zones, as it is indistinguishable from the surrounding snow-covered terrain. Once the model is trained and inferences are made, a post-processing step is applied to improve the predictions by i) connected-region size thresholding, ii) gap-filling, iii) improving the predictions over lake-contact termini based on the Normalized Difference Water Index (NDWI). Using only the imagery and the DEM as inputs, the method achieves 80.9% Intersection Over Union (IOU) and 89.5% F1 on the testing regions, which further increase to 84.1% and 91.4%, respectively, when including all the DEM-derived features.

In a different study focused on the central Karakoram region, Xie et al. (2021) compare GlacierNet (Xie et al., 2020) against five different CNN-based segmentation models. Instead of GLIMS, the authors used for the ablation areas the more-accurate contours from the Glacier Area Mapping for Discharge in Asian Mountains (GAMDAM) dataset (Nuimura et al., 2015). The chosen baseline methods include three versions of U-Net (Ronneberger et al., 2015): Mobile-UNet (Jing et al., 2022) i.e. U-Net with a MobileNetV2 backbone (Sandler et al., 2018), Res-UNet i.e. U-Net with a ResNet34 backbone (He et al., 2016), and R2U-Net (Alom et al., 2018) i.e. U-Net with recurrent convolutional layers. The last two are FCDenseNet (Jégou et al., 2017) and DeepLabv3+ (Chen et al., 2018) with an Xception backbone (Chollet, 2017). They found that DeepLabv3+ performs the best, with an 86.2% IOU. However, GlacierNet gives the second-highest score, only 0.2% lower, but at a smaller computational cost.

GlacierNet2 (Xie et al., 2022) In a follow-up study, Xie et al. (2022) proposed an improvement of GlacierNet by i) using multiple models to improve the predictions over the glacier termini, ii) including also the snow-covered accumulation zones, which were previously discarded. GlacierNet2 can be considered as a two-member ensemble model as it combines the previous GlacierNet with the predictions from a DeepLabv3+ model (Chen et al., 2018) with an Xception backbone (Chollet, 2017). Each sub-model is trained independently, then their weights are frozen, and lastly, a final 1x1 convolutional layer is trained to fuse their predictions. However, the predictions from GlacierNet alone are still kept and post-processed in parallel with those from the fused GlacierNet-DeepLabv3+ model. Adding to the post-processed steps proposed in the previous work (i.e. gap-filling and region-size thresholding), the authors implement an additional step where the final predictions of the two sub-components are compared at the termini and disagreements are addressed through a k-Nearest Neighbor (KNN) classifier. Also, a particular post-processing pipeline is applied to the accumulation areas for the snow-covered pixels detected using the NDSI. An ablation study shows that the final model reaches an 88.4% IOU and a 93.8% F1 score, improving by 1-2% the baselines, i.e. GlacierNet, DeepLabv3+ and their combination, when evaluated on ablation-zone mapping.

Tian et al. (2022) An improved U-Net (Ronneberger et al., 2015) architecture is used by Tian et al. (2022) to segment glaciers in the Pamir Plateau. The original architecture is improved by incorporating the channel attention module from Roy et al. (2018), the so-called channel squeeze and excitation block. On top of this, the authors also used a conditional-random field (CRF) method as post-processing to refine the results. For training, they used optical data from Landsat-8 and the SRTM DEM with the labels based on GLIMS (Raup et al., 2007). These were manually corrected to account for a temporal gap between the imagery and the labels during which glacier changes occurred. The model achieved an F1 score of 89.5% after adding the attention mechanism, further improved to 89.8% with the CRF-based refinement, performing better than the original U-Net, which obtained 88.9%. They also tested the GlacierNet model (Xie et al., 2020), which obtained an F1 score of only 84.9%.

Chu et al. (2022b) Similar to the previous study of Tian et al. (2022), in this work, Chu et al. (2022b) also incorporate an attention mechanism, i.e. convolutional block attention module (CBAM) (Woo et al., 2018). They build upon the DeepLabv3+ architecture (Chen et al., 2018) with a ResNet-34 backbone (He et al., 2016). Additional improvements were obtained using test-time augmentation and depth-wise separable convolution at the end of the segmentation (Chollet, 2017). They used high-resolution (8m) multi-spectral data (R, G, B, Near-InfraRed (NIR)) from the Gaofen-6 satellite, which also provides a panchromatic band with 2m resolution. For training, they manually delineated glaciers from the Tanggula, Kunlun and the Qilian Mountains. Finally, they compared their results based on the model's predictions with existing regional inventories, showing a more accurate extraction of the debris-free glaciers. Their model achieves an F1 score of 98.54%. They compared to multiple baselines, the best performing one being the original DeepLabv3+ model, which achieves already an F1 score of 98.3%, followed by U-Net (Ronneberger et al., 2015) with a ResNet-18 backbone (He et al., 2016) which obtains an F1 score of 97.8%.

Peng et al. (2023) The transformer-based architecture builds upon Swin-Unet (Cao et al., 2023), with a series of improvements from various other works. The decoder was coupled with locally-grouped self-attention and global sub-sampled attention modules from Twins-SVT (Chu et al., 2021), together with the conditional position encoding from Chu et al. (2022a). The decoder uses Local-Global CNN Blocks, ending with a Feature Refinement Head as the segmentation head, both from Unet-Former (Wang et al., 2022). The study focuses on the Qilian Mountains, using the inventory from Li et al. (2020). As input data, the following were used: five optical bands (Sentinel-2), i.e. R, G, B, NIR and SWIR (B11), from which three indices were obtained i.e. Normalized Difference Vegetation Index (NDVI), NDWI and NDSI; two SAR backscatter intensity images (VV and VH polarizations), using the Level-1 Ground Range Detected from Sentinel-1; and a DEM using the 8m High Mountain Asia Digital Elevation Model (HMADEM) (Shean, 2017) as the main source, complemented with the SRTM one for the regions not covered by the former. The final model achieves an F1 score of 84.3%, followed by the Swin transformer (Liu et al., 2021) with 82.9% and the model from the previous study of Chu et al. (2022b), based on DeepLabv3+ (Chen et al., 2018), with 82.2%. A detailed ablation study on various input features groups shows that optical bands are the most important but also highlights the benefit of combining multi-source datasets.

Thomas et al. (2023) This work focuses on mapping debris-covered glaciers from three HMA sub-regions, i.e. Hunza (Karakoram), Manaslu (Central Himalayas) and the Khumbu (Central Himalayas). The methodology is similar to the one from a previous study (discussed in Section 4.1.3), focused on mapping rock glaciers (Robson et al., 2020). It

consists of a five-layer CNN with a post-processing stage using OBIA. The model's output is finer compared to previous studies, as it distinguishes among seven classes, i.e. supraglacial debris, clean ice, snow cover, lakes, vegetation, shadows, and non-glacial material. As input, the model uses a wide range of features (21 in total):

- ten optical bands (Sentinel-2) from which three indices were obtained i.e. NDVI, NDWI and NDSI
- SAR coherence (Sentinel-1)
- a thermal band (Landsat-8)
- a DEM (AW3D30) with five derived features, i.e. slope angle, profile curvature, planform curvature, aspect, and shaded relief

The model is trained using the GAMDAM inventory (Nuimura et al., 2015), and the analysis of the results shows that the integration of OBIA as a post-processing step helps reduce the model's errors. Additionally, for one of the regions (Manaslu), the authors apply their methodology to declassified panchromatic imagery from the Corona KH-4B satellite. Although the results show the limitations caused by the absence of the multi-spectral dimension, the approach illustrates the potential towards building historical inventories for HMA.

GlaViTU (Maslov et al., 2023) Following the advances in Computer Vision research, an architecture based on a VisionTransformer (ViT) (Dosovitskiy et al., 2020) is used. The authors propose a hybrid CNN-transformer model, called Glacier-VisionTransformer-U-Net or GlaViTU, by combining SEmgentation TRansformer (SETR) (Zheng et al., 2021) with a ResNet-backbone U-Net (He et al., 2016; Ronneberger et al., 2015). The method is compared to three baselines. First, TransUNet (Chen et al., 2021), i.e. a different type of U-Net that uses a hybrid ResNet-50 (He et al., 2016) + ViT (Dosovitskiy et al., 2020) encoder. The second, ResU-Net, is the U-Net with a classical ResNet backbone, also used in the previously mentioned study of Xie et al. (2021). The third baseline is a SETR-B/16 model with progressive upsampling decoder (Zheng et al., 2021). For a fair comparison, all the baselines are modified by adding a data-fusion CNN-based head as in the proposed architecture. This block was proposed to better fuse the three input data modalities, i.e. multi-spectral optical data from Landsat 7, 8 and Sentinel-2, σ_0 -calibrated amplitude images from Envisat and Sentinel-1, and DEMs (Cop30DEM and AW3D30). The dataset is much larger compared to previous studies, and it covers six regions worldwide: the European Alps, High-Mountain Asia, Indonesia, New Zealand, the Southern Andes and Scandinavia. On average, over these six regions, GlaViTU achieves an 87.5% IOU, followed by TransU-Net and ResU-Net with 86.3% and 85.2%, respectively.

In a more recent study, Maslov et al. (2024) extend the previous dataset towards a global one. Compared to the previous studies, the dataset is much larger, covering most of the glacier regions outside the two ice sheets. With around 19,000 glaciers, it covers ca. 7% of the total glaciated area. The training labels are mainly based on GLIMS (Raup et al., 2007) and RGI (Pfeffer et al., 2014), but for 8 out of the 23 sub-regions, local inventories are used. In addition to the data sources used in their previous work, i.e. optical, SAR and DEM, the authors also investigate whether adding thermal data when available from Landsat-8 helps but did not find it effective. Moreover, in this work, InSAR coherence images are also considered (when available) instead of the amplitude images, improving performance. The original GlaViTU model is further improved by updating the data fusion block, which is now equipped with feature weighting and includes squeeze-and-excitation blocks. The DeepLabv3+ model (Chen et al., 2018), with a ResNeSt-10 backbone (Zhang et al., 2022), is used as a baseline, after adding the same data fusion block as for GlaViTU. The average IOU over all the regions is 89.4%, almost 2% higher than the baseline, which obtains 87.7%. Instead of training a single global model, Maslov et al. (2024) also investigate four additional training strategies, i.e. regional training, finetuning and location encoding (by using either the region or the coordinate), showing that on average, the regional and the fine-tuned models perform the best, with almost 1% better than the global model. Lastly, it was investigated for the first time whether uncertainty estimation techniques, namely Monte Carlo dropout (Gal et al., 2016) and temperature scaling (Guo et al., 2017), can be used to get further insights into the predictions. They concluded that temperature scaling alone could already help to extract confidence intervals for the predictions and qualitatively found that the model is more uncertain on the debris-covered segments or those under the shadow, thus illustrating the practicality of the uncertainty estimates. So far, this study represents the most comprehensive dataset for glacier mapping using DL which can facilitate further methodological developments by directly using the processed data and the same validation procedure.

Publication (model name)	model architecture ³	data modality ⁴ (source)	ROI ⁵	code/ data ⁶ / output ⁷
Xie et al. (2020) (GlacierNet)	SegNet (Badrinarayanan et al., 2017)	<ul style="list-style-type: none"> optical (Landsat 8) DEM (AW3D30) 	HMA: Nepal Himalaya, central Karakoram	-
Xie et al. (2022) (GlacierNet2)	SegNet & DeepLabv3+ (Badrinarayanan et al., 2017; Chen et al., 2018)		HMA: central Karakoram	-
Tian et al. (2022)	U-Net (Ronneberger et al., 2015) with attention	<ul style="list-style-type: none"> optical (Landsat 8) DEM (SRTM) 	HMA: Pamir	-
Chu et al. (2022b)	DeepLabv3+ (Chen et al., 2018) with attention	<ul style="list-style-type: none"> optical (Gaofen-6) 	HMA: Tanggula, Kunlun, Qilian	C/D/-
Peng et al. (2023)	based on Swin-Unet (Cao et al., 2023), with various improvements	<ul style="list-style-type: none"> optical (Sentinel-2) SAR (Sentinel-1) DEM (HMADEM, SRTM) 	HMA: Qilian	-
Thomas et al. (2023)	custom (5-layers CNN)	<ul style="list-style-type: none"> optical (Sentinel-2) InSAR coherence (Sentinel-1) DEM (AW3D30) thermal (Landsat-8) 	HMA: Khumbu, Manaslu, Hunza	-
Maslov et al. (2023) (GlaViTU)	SETR & U-Net (Ronneberger et al., 2015; Zheng et al., 2021)	<ul style="list-style-type: none"> optical (Landsat 7/8, Sentinel-2) SAR⁸ (Envisat, Sentinel-1) DEM (AW3D30, Cop30DEM, SRTM) thermal⁹ (Landsat-8) 	European Alps, HMA, Indonesia, New Zealand, S Andes and Scandinavia	C/D/-
Maslov et al. (2024) (GlaViTU ¹⁰)			global	C/D/-

Table 1: Summary of DL-based studies for standard glacier-extent mapping. For details, see Section 4.1.1 or the corresponding publications. The full links are also provided in Section 7

4.1.2 Area Change Estimation

Since the area of a glacier is a crucial indicator of its health, tracking its evolution over time is an important application of glacier mapping methods. Traditionally, this is done by manually re-creating glacier inventories after a certain period, usually decades, such that long-term impacts of climate change can be observed (see Section 2 where we provide examples). In principle, once a DL model is trained (as presented in the previous section), it can be applied to delineate glaciers at different points, with the predicted results being used to analyze the temporal changes. However, a few

³Note that sometimes changes to the original architecture are made, see Section 4.1.1.

⁴Often features are further derived e.g. NDVI from optical or slope from DEM.

⁵We only indicate the main Regions of Interest (ROIs) from where the training data was extracted. Still, the coverage can vary significantly, e.g. from a few image scenes to almost complete coverage.

⁶This refers to the processed training data, not the raw one. The latter is usually openly available from the specified source.

⁷In case the study has an output product (e.g. an inventory).

⁸In both studies σ_0 -calibrated amplitude images are used. In the second study, InSAR coherence images are also included when available.

⁹used only in the second study

¹⁰In this work, the original GlaViTU model from Maslov et al. (2023) is slightly modified, see Section 4.1.1

challenges arise in practice when detecting glacier changes from outlines derived from a DL framework, requiring additional assumptions. To list a few:

- To extract an area change rate for each individual glacier, the same ice divides have to be used, as usually the methods only classify a pixel as being part of the glacier or not, without any knowledge about what a glacier is as an independent object. This can be problematic when tributaries are becoming disconnected from the main glacier.
- In general, to increase the signal-to-noise ratio, there should be a relatively long period between the considered outlines. However, most models are trained on imagery from a single source, e.g., Landsat-8, usually constrained by the static inventory that provides the labels, e.g., the RGI. If the resulting model is then applied on data from different sensors to increase the temporal coverage, we potentially need to deal with generalization issues.
- Even if the data comes from the same sensor, temporal generalization issues can still occur. One such situation is when the imagery has different characteristics compared to the one used for training, e.g. different cloud coverage or seasonal snow conditions. See Figure 6 as an example.
- Increasing debris-coverage in a warming climate (Compagno et al., 2022; Tielidze et al., 2020) can also affect the temporal generalization. Since automatic methods are still significantly more affected by errors on debris-covered glacier parts than clean ice, the change in the debris coverage percentage can introduce biases in the results. An example is provided in Figure 7.
- Once a model is trained, we ideally want to employ it on the entire ROI to cover all the glaciers. However, this requires applying the model also on the data on which it was trained. Given the risk of memorization, especially when training with noisy labels (Arpit et al., 2017), we can expect that the model errors over time are not independent, thus breaking the temporal generalization assumption.

To summarize, additional challenges arise when relying on DL for glacier area change quantification. These challenges perhaps explain why there are only a relatively small number of DL works that explicitly focus on glacier changes. A few works that attempted to identify glacier changes are presented in the following three paragraphs and summarized in Table 2.

GlacierCoverNet (Roberts-Pierel et al., 2022) The availability of long time-series data from the Landsat program offers some opportunities for glacier studies. Roberts-Pierel et al. (2022) exploit the data starting from 1985 up to 2020 covering Alaska to study how glacier surface area evolved over time, an important indicator of the glaciers' health. After building a temporal mosaic to account for clouds and seasonal snow, they obtained 18 biannual image composites, each with full spatial coverage of the selected region. Based on this, they extract five features: NDSI, NDVI, Normalized Burn Ratio (NBR) and Tasseled Cap Brightness & Wetness (Kauth et al., 1976). Additionally, a DEM is used as input, together with three features derived from it: curvature and aspect intensity (North and South), i.e. a method of scaling the cosine of aspect by the sine of slope (Kirchner et al., 2014). The proposed model is based on the FSPNet architecture (Zhao et al., 2017) with a ResNeSt-101 backbone (Zhang et al., 2022). The glacier outlines from RGI 6.0 were used for training. Another significant difference compared to the previous studies is that the model is trained to predict separately whether a pixel is no glacier, supra-glacial debris or debris-free glacier. This was achieved by developing a method for identifying the debris pixels and then assuming that those within the RGI outlines are debris-covered glacier pixels. Once the model was trained using the image composites as close as possible to the RGI dates, it was applied on the entire time-series, thus producing 18 glacier inventories for Alaska. These show a significant area loss: the total glaciated area decreased by 8425 km² (i.e. -13%), with sizeable sub-regional variability. The relatively fine temporal resolution also allows the inspection of the glacier area changes through time, showing a much stronger shrinkage over the last 15 years (2005-2020) compared to the earlier period. Lastly, the distinction between debris and clean ice allows studying the debris evolution over time, revealing an increase of around 64% over 1985-2020, also with significant sub-regional variability.

Rajat et al. (2022) A similar investigation but on a much smaller scale shows that the Himachal glaciers retreated significantly, from an estimated total area of around 4,021 km² in 1994 to only 2199 km² in 2021, resulting in an

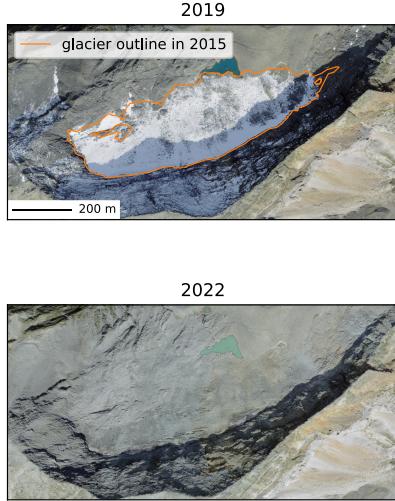


Figure 6: Debris-covered glacier with fresh snow. The effect of fresh snow when mapping a fully debris-covered glacier is here shown for Glatscher da Sut Fuina, in Switzerland (46.535° N, 9.473° E) (same glacier as in Figure 5). The two aerial images from swisstopo (2024b) capture an important issue that can affect automatic glacier extent mapping & area change analysis, i.e. predicting the presence of a glacier using, as a proxy, the superimposed fresh snow, hardly present outside the glacier surface (upper panel, where the snow only remains over the debris-covered glacier, which is colder than the non-glaciated surroundings). Since automatic methods, including DL-based ones, are usually challenged in the case of debris cover, in this example, one would probably underestimate the glacier surface in 2022 (without snow) and thus overestimate the shrinkage rate. This figure illustrates that choosing imagery with similar climatic conditions, ideally without any snow, should be a priority, which is especially important for glacier area change analyses. See also Figure 7.

annual retreat rate of $\sim 68 \text{ km}^2$ ($\sim 1.68\%$). To obtain these results, a U-Net was trained on manually annotated glaciers by visualizing Landsat-8-based NDSI and a DEM from USGS, with additional derived features. The model, a U-Net (Ronneberger et al., 2015), is trained using a subset of four bands and then applied on four different years, i.e. 1994, 2001, 2011 and 2021, using input data from Landsat 4/5/8.

Diaconu et al. (2023) This is another study that makes use of DL for investigating glacier area changes, focused on a different region, the European Alps (RGI-11). The input data consists of a subset of five bands (R, G, B, NIR, SWIR-B12) from Sentinel-2 and a DEM. A major advantage of this study is the relatively good quality of the labels: they are based on a new inventory (Paul et al., 2020) and estimated to be of better quality than the previous RGI one. It is based on Sentinel-2 data from (mainly) 2015, ensuring a perfect match between the satellite data used for training and the glacier outlines. The model used is a U-Net (Ronneberger et al., 2015) with a ResNet34 backbone (He et al., 2016). To avoid using inferences made on the training data, which can lead to biases, five models are trained using a regional cross-validation scheme. Once the models are trained, they are applied on the most recent data, i.e. 2023, which was a strong melt year according to Glacier Monitoring in Switzerland (GLAMOS) (GLAMOS, 2023). This offers, however, ideal conditions for glacier mapping as it reduces the chances of seasonal snow, which can cause many false positives. Based on the models' predictions, the areas are estimated both at the inventory time and in 2023, which are then compared to estimate change rates for each individual glacier. To increase the signal-to-noise ratio, the authors make two important assumptions: i) the glaciers do not grow in the given period, supported by geodetic mass-balance studies (Hugonet et al., 2021) and ii) the models make systematic errors. The second assumption, if true, implies that the segments missed by the models, e.g. debris-covered or under shadow, do not significantly affect the estimated change rates. The regional cross-validation scheme helps support this assumption since it decreases the chances that models would perform differently at the inventory time compared to 2023. To further reduce the impact of the models' errors on the area change rates, an outlier filtering scheme was used to drop the glaciers for which the model performs poorly. After this step, individual estimates for around 1,300 glaciers are provided, representing 87% of the glacierized area in the region. Regionally, the estimate is around -1.8% loss per year, which illustrates the high sensitivity of the glaciers in this region to climate change. A glacier-level analysis further shows significant inter-glacier variability.

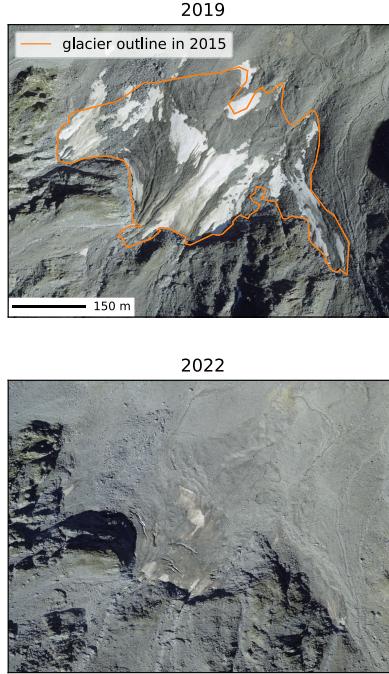


Figure 7: A glacier with increasing debris coverage. For many debris-covered glaciers, there is a tendency for debris cover to increase over time, as illustrated here for the Tambogletscher, a glacier in Switzerland (46.504° N, 9.291° E). The two aerial images from swisstopo (2024b) show that initially, in 2019, the glacier was only partially covered by debris (upper panel), while three years later (lower panel), it has become almost completely debris-covered. Such transitions will affect automatic glacier extent mapping & area change analysis since automatic methods, including DL-based ones, usually face important challenges related to detecting the presence of debris. In the example presented here, many detection methods will likely underestimate the glacier area in 2022 and thus overestimate the glacier shrinkage rate. This figure suggests that capturing the uncertainties in the methods becomes critical to avoid significant biases in the estimates. See also Figure 6.

Publication (model name)	model architecture ^[1]	data modality ^[2] (source)	ROI ^[3]	code/ data ^[4] / output ^[5]
Roberts-Pierel et al. (2022) (GlacierCoverNet)	FSPNet (Zhao et al., 2017)	<ul style="list-style-type: none"> optical (Landsat 4/5/7/8) DEM (USGS-3DEP) 	Alaska	-/-/O
Rajat et al. (2022)	U-Net (Ronneberger et al., 2015)	<ul style="list-style-type: none"> optical (Landsat 4/5/8) 	HMA: Himachal	-
Diaconu et al. (2023)	U-Net (Ronneberger et al., 2015)	<ul style="list-style-type: none"> optical (Sentinel-2) DEM (NASADEM) 	European Alps (RGI-11)	C/D/-

Table 2: Summary of DL-based studies focused on glacier area change analysis. For details, see Section 4.1.2 or the corresponding publications. The full links are also provided in Section 7.

^[1]Note that sometimes changes to the original architecture are made, see Section 4.1.1

^[2]Often features are further derived e.g. NDVI from optical or slope from DEM.

^[3]We only indicate the main ROIs from where the training data was extracted. Still, the coverage can vary significantly, e.g. from a few image scenes to almost complete coverage.

^[4]This refers to the processed training data, not the raw one. The latter is usually openly available from the specified source.

^[5]In case the study has an output product (e.g. an inventory).

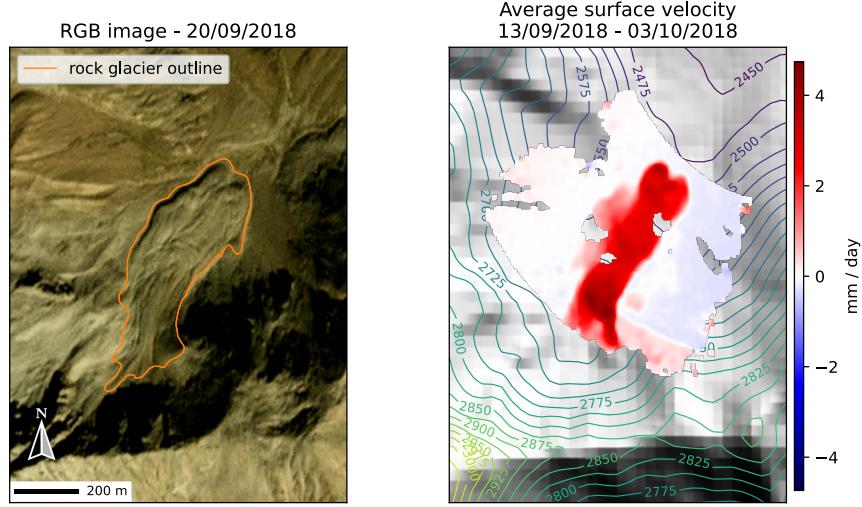


Figure 8: An active rock glacier. Some rock glaciers that have substantial amounts of ice are subject to a distinct downslope (gravitational) flow, as illustrated here for the Lazau glacier in northern Italy (46.746° N, 10.755° E). The satellite image (left panel) from PlanetScope (3 m GSD) shows the extent of the glacier, which is difficult to distinguish from the surrounding landforms. On the right panel, we show the average daily surface velocity from Bertone et al. (2023) measured using ground-based SAR (with Copernicus GLO-30 DEM in the background). This figure illustrates how similar a rock glacier is to the surrounding landscape and how utilizing surface velocity observations can contribute to a better distinction.

4.1.3 Rock Glaciers Mapping

A special class of glaciers is the so-called rock glaciers. These are essentially a mixture of frozen debris and ice^[16]. As opposed to typical glaciers, which, by definition, are flowing, rock glaciers can be both active and non-active. They are important for cryosphere studies as they can indicate the permafrost distribution in the region. Rock glaciers are also affected by climate change but are more resilient due to the insulated effect of the rocky material and the active layer (Robson et al., 2020). Detecting this type of glaciers from optical data is even more difficult as, by definition, they are covered by debris. Consequently, the studies on this area usually have a multi-modal approach, e.g., by including SAR coherence that can capture small deformations that could occur over time (at least for the active ones). An example of a rock glacier is provided in Figure 8. In the following paragraphs, we describe three major studies focused on rock glaciers and then summarized in Table 3.

Robson et al. (2020) Rock glaciers pose even more challenges to automated methods compared to glaciers that are only partially covered by debris due to the spectral similarity between them and the surrounding material. Robson et al. (2020) propose to use a CNN model combined with OBIA to further improve the predictions using additional morphological and spatial characteristics. Given the difficulty of the task, various data sources are used: optical data from Sentinel-2, InSAR coherence data from Sentinel-1 and a Pléiades DEM (processed by the authors). The study focuses on the La Laguna and Poiqu catchments, in the Andes, and the central Himalayas, respectively. The ground truth labels are obtained from a dataset by Schaffer and Macdonell for the La Laguna catchment and the one from Bolch et al. (2020) for Poiqu. The authors used a relatively small DL model, a five-layers CNN trained with eleven input features: five optical bands (R, G, B, NIR, SWIR) + three derived indices (NDVI, Modified Normalized Difference Water Index (MNDWI)(Xu, 2006), Soil Adjusted Vegetation Index (SAVI) (Alba et al., 2012)), InSAR coherence, DEM + derived curvature. With the full pipeline, i.e. CNN + OBIA, they automatically mapped 108 of the 120 glaciers considered in the validation set. The authors also investigated, for a small sub-region of the Poiqu catchment, whether using higher resolution optical data and the corresponding DEM from Pléiades (2m) boosts the performance and found a 9% increase in precision but only 1% increase in recall.

Hu et al. (2023) If there are already many inventories for typical glaciers, including RGI (Pfeffer et al., 2014), which has global coverage, rock glaciers are yet to be identified in some regions. The Western Kunlun Mountains represent

^[16]For a more technical definition, see RGK (2023).

one such case. Hu et al. (2023) built an inventory for the active rock glaciers in the region using InSAR from ALOS and VHR images from Google Earth. Using this data, a DeepLabv3+ model (Chen et al., 2018) with an Xception backbone (Chollet, 2017) was trained on Sentinel-2 (only RGB) and applied to the entire region to further identify glaciers that were previously missed, increasing the initial number of 290 glaciers to 413. The outlines produced by the model had however to be manually inspected and corrected, therefore further research is necessary towards a fully-automatic pipeline.

Sun et al. (2024) For the first time, a regional-scale inventory for rock glaciers was built by Sun et al. (2024). The final benchmark dataset contains 44,273 glaciers covering $\sim 6000 \text{ km}^2$ ($\mu = 0.14 \text{ km}^2$). This was achieved by initially compiling multiple existing inventories, not only from the Tibetan Plateau but also from other regions, into a large dataset containing 4085 rock glaciers. This was then used to train a DL model using Planet Basemaps. The model architecture is the same as in Hu et al. (2023), i.e. a DeepLabv3+ model (Chen et al., 2018) with an Xception backbone (Chollet, 2017). The initial predictions of the DL model were investigated and manually corrected following a strict guideline (RGIK, 2023), with the effort of seven mappers and two independent reviewers. At this stage, high-resolution Google Earth images and ESRI basemaps were also utilized. When comparing the initially DL-produced polygons with the final ones revised by experts, a 63% F1 score was obtained (precision = 55%, recall = 73%).

Publication (model name)	model architecture ¹⁷	data modality ¹⁸ (source)	ROI ¹⁹	code/ data ²⁰ / output ²¹
Robson et al. (2020)	custom (5-layers CNN)	<ul style="list-style-type: none"> optical (Sentinel-2, Pléiades) InSAR coherence (Sentinel-1) DEM (generated from Pléiades) 	La Laguna catchment (Andes) Poiqu catchment (Himalaya)	-
Hu et al. (2023)	DeepLabv3+ (Chen et al., 2018)	<ul style="list-style-type: none"> optical (Sentinel-2) 	HMA: Western Kunlun	C/-/O
Sun et al. (2024)	DeepLabv3+ (Chen et al., 2018)	<ul style="list-style-type: none"> optical (Planet Basemaps) 	HMA: Tibetan Plateau	-/-/O

Table 3: Summary of DL-based studies focused on rock glaciers mapping. For details, see Section 4.1.3 or the corresponding publications. The full links are also provided in Section 7

4.2 Calving Front Detection

For marine-terminating glaciers, changes in the calving front are an essential indicator of the underlying glacial dynamics, with shifts in the calving front hinting at melt processes or surge events. As the mass loss of the major ice sheets in Antarctica and Greenland is projected to be a major contributor to global sea level rise (Calvin et al., 2023), understanding and monitoring these developments is paramount.

Various DL approaches have been suggested in recent years to automatically monitor calving fronts for the Earth's ice sheets. Generally, these methods use either optical or SAR imagery as their primary source of imagery, with some methods taking auxiliary data as additional inputs, such as elevation data. The following is an overview of relevant existing works on DL for calving front detection:

Baumhoer et al. (2019) Seeing the potential of DL methods for automated detection of ice sheet calving fronts in Antarctica, Baumhoer et al. (2019) adapted the U-Net model (Ronneberger et al., 2015) for this task. As a main

¹⁷Note that sometimes changes to the original architecture are made, see Section 4.1.1

¹⁸Often features are further derived e.g. NDVI from optical or slope from DEM.

¹⁹We only indicate the main ROIs from where the training data was extracted. Still, the coverage can vary significantly, e.g. from a few image scenes to almost complete coverage.

²⁰This refers to the processed training data, not the raw one. The latter is usually openly available from the specified source.

²¹In case the study has an output product (e.g. an inventory).

input feature, dual-polarized Sentinel-1 SAR data was chosen for its year-round availability and robustness against weather conditions. For training, the calving front was manually annotated in SAR imagery for the Sulzberger, Victoria Land, Wilkes Land and Shackleton regions. Four additional regions were selected for testing purposes: Ekstromisen, Wordie, Marie Byrd Land and Oats Land. As an additional input feature for the model, the HH/HV ratio was derived. Finally, TanDEM-X elevation data was merged into the data stack as a fourth input to guide the model in regions with confounding backscatter behaviour.

Mohajerani et al. (2019) In a first case study for Greenland glaciers, Mohajerani et al. (2019) trained a U-Net model (Ronneberger et al., 2015) to detect calving fronts for three glaciers in Landsat imagery. For Landsat 5, the green band was extracted, while the panchromatic band was used for Landsat 7 and 8. Calving fronts were annotated manually for each image. The studied regions cover the Helheim, Sverdrup, Kangerlussuaq, and Jakobshavn glaciers. The imagery is segmented into two classes, namely a “background” class and a “calving front” class, following an edge-detection approach rather than a zonal segmentation approach. The final calving front prediction is then derived via a path-finding algorithm and re-projection to the original geographical coordinates.

Zhang et al. (2019a) Seeing the value of SAR data for calving front detection, Zhang et al. (2019a) manually delineated coastline positions for the Jakobshavn Isbrae glacier in Greenland using imagery from the TerraSAR-X mission. As a segmentation model, they train a U-Net model (Ronneberger et al., 2015) to segment the imagery into a “glacier” and an “ocean” class. The calving front is then extracted by tracing the boundary between these two classes in a post-processing step.

CALFIN (Cheng et al., 2021) Cheng et al. (2021) built a large-scale dataset of calving front observations in Greenland by manually annotating Landsat data from 1972 to 2019, covering 66 glaciers. Similar to Mohajerani et al. (2019), single-channel data is used (in this case, from the near-infrared band). However, the imagery is pre-processed into three channels by applying contrast normalization algorithms and stacking the results. Using this dataset, a DeepLabv3+ model (Chen et al., 2018) with an Xception backbone (Chollet, 2017) is then trained to both segment the imagery into ocean and land classes, as well as directly mark the calving front. In a post-processing step, the calving front is then extracted by constructing a minimum-spanning tree and finding the longest path within this tree.

Zhang et al. (2021) Following up on their previous study (Zhang et al., 2019a) extend their dataset to span three glaciers, namely Jakobshavn, Kangerlussuaq, and Helheim. Further, multiple data modalities are included, namely optical imagery from Landsat-8 and Sentinel-2, as well as SAR imagery from Envisat, ALOS-1, TerraSAR-X, Sentinel-1 and ALOS-2. Multiple models are trained to perform a binary segmentation into land/glacier and ocean, with the best performance observed for a DeepLabv3+ model (Chen et al., 2018) with a DRN backbone (Yu et al., 2017).

HED-UNet (Heidler et al., 2022) Seeing the advantages of both segmentation and edge detection approaches for the task of calving front detection, Heidler et al. (2022) combined these two tasks in a single model. This dual training enhances model predictions near the calving front, which tend to be imprecise and blurry for segmentation-based models. The developed model takes inspiration from both the U-Net (Ronneberger et al., 2015) and the HED edge detector (Xie et al., 2015). The dataset from Baumhoer et al. (2019) is used as training and evaluation data. However, the DEM input channel is identified as a potential confounder in this study, which can cause the model to overfit to this static data product and ignore the actual SAR imagery. The authors, therefore, advocate against using elevation data as a direct input feature for the DL model and instead suggest only incorporating this data in a separate post-processing step. IceLines (Baumhoer et al., 2023) is a continuously updated data product for the entire Arctic derived using this model.

Loebel et al. (2022) Seeing that previous studies for calving front extraction in Greenland mostly relied on single-channel imagery, Loebel et al. (2022) studied the benefits of incorporating various additional data modalities into the training process. Using Landsat-8 imagery as the main imagery source, they manually delineate calving fronts for 23 glaciers in Greenland glaciers and two glaciers on the Antarctic Peninsula. The used DL model is a deepened version of the U-Net (Ronneberger et al., 2015) model, including two additional down- and up-sampling stages to allow for a larger spatial context. Starting with panchromatic imagery as a baseline input modality, the authors test various configurations by adding multi-spectral channels, statistical texture information, and topography data from the BedMachine Greenland v3 dataset (Morlighem et al., 2017). The largest gains in model accuracy are observed when multi-spectral information

is included. Meanwhile, the inclusion of textural and topographic information seems to introduce a trade-off, where the model performance becomes more robust on challenging scenes, but, in turn, accuracy decreases for most other images.

Gourmelon et al. (2022) In an effort to make the task of calving front detection more approachable for researchers from the field of computer vision, Gourmelon et al. (2022) build CaFFE, a “machine-learning ready” dataset of calving front positions in Greenland, Antarctica, and Alaska. The imagery used comes from various SAR sensors, namely ERS-1/2, RADARSAT 1, Envisat, ALOS, TerraSAR-X, TanDEM-X, and Sentinel-1. As ground truth annotations, the authors provide both calving front masks for edge-oriented approaches and zone labels that segment the imagery into the classes “ocean”, “rock”, “glacier”, and “NA”. As a baseline model for future comparisons, the authors train an adapted U-Net model (Ronneberger et al., 2015), which is augmented by an atrous spatial pyramid pooling layer (Chen et al., 2018) in the bottleneck of the network. Notably, this study quantifies the effect of the seasons and image resolution on the prediction accuracy, suggesting that summer images are to be preferred over winter images and that higher resolution can help with prediction accuracy in some cases.

Periyasamy et al. (2022) Seeing the wide-spread use of the U-Net model (Ronneberger et al., 2015) for calving front detection, Periyasamy et al. (2022) conducted a systematic study to better understand the influence of specific hyperparameters for this task and give recommendations on how to tune models for calving front detection. Using an earlier version of the CaFFE dataset (Gourmelon et al., 2022), they optimize various components of the training process, such as data preprocessing, data augmentation, loss function, bottleneck, normalization layers and dropout layers. They observe optimal performance when using adaptive histogram equalization.

COBRA (Heidler et al., 2023) Nearly all existing studies employing DL for calving front detection are trained to provide dense predictions, either in the form of semantic segmentation or edge detection. Heidler et al. (2023) take another approach: by adopting the idea of deep active contours (Peng et al., 2020), they introduce a model that directly predicts a contour line, parameterized by a sequence of vertices in the image space. In this way, the model is encouraged to focus on the actual calving front during training, and the predictions can be used without any post-processing steps. This model is trained on the CALFIN dataset (Cheng et al., 2021), and has been applied for a large-scale study of calving front dynamics in Svalbard (Li et al., 2024).

Zhang et al. (2023) To leverage the increasing amount of openly available remote sensing data, Zhang et al. (2023) developed an automated pipeline for calving front extraction using both SAR (from Sentinel-1) and optical data (Landsat 5,7,8 and Sentinel-2). For training labels, the authors make use of TermPicks (Goliber et al., 2022), a large-scale dataset with manually digitized calving fronts. To quantify the uncertainties in the model (a DeepLabv3+ (Chen et al., 2018)), Monte Carlo dropout (Gal et al., 2016) is employed, combined with a temporal ensemble (i.e. combining multiple predictions for the same date). The high-temporal resolution of their results allows them to capture the seasonal variability, with the final product, AutoTerm, covering 295 outlet glaciers in Greenland and 278,239 calving fronts.

AMD-HookNet (Wu et al., 2023) A common observation in calving front detection research is the need for large spatial context windows. Naive solutions to addressing this requirement, such as training on larger image patches and increasing the size of convolutional filters, are computationally inefficient. Setting out to address this issue in a more elegant manner, Wu et al. (2023) introduce AMD-HookNet, a deep neural network designed to operate on two versions of a satellite scene at different spatial resolutions. By interlocking two U-Net (Ronneberger et al., 2015) branches with attention layers, more general information with a wide spatial context can be applied to improve the predictions of the high-resolution branch. This model is trained on the CaFFE dataset introduced by Gourmelon et al. (2022). Evaluations show that this dual-resolution approach can indeed improve prediction accuracy considerably.

HookFormer (Wu et al., 2024) Recently, vision transformers (Dosovitskiy et al., 2020) have become the tool of choice for many computer vision tasks. Following the ideas introduced in the previous study, Wu et al. (2024) combined a Swin Transformer model (Liu et al., 2021) with the HookNet approach, where a high-resolution and a low-resolution branch are interleaved. The resulting model is again trained on the CaFFE dataset Gourmelon et al. (2022). Compared to the previous evaluations, this transformer-based model outperforms the previous CNN-based models.

Publication (model name)	model architecture ²²	[dataset ²³] data modality ²⁴ (source)	ROI ²⁵	code/ data ²⁶ / output ²⁷
Baumhoer et al. (2019)	U-Net (Ronneberger et al., 2015)	• SAR (Sentinel-1) • DEM (TanDEM-X)	Antarctica (8 sites)	-
Heidler et al. (2022) (HED-UNet)	custom; based on HED (Xie et al., 2015) & U-Net (Ronneberger et al., 2015)			C/-O
Mohajerani et al. (2019)	U-Net (Ronneberger et al., 2015)	• optical (Landsat 5,7,8)	Greenland (4 glaciers)	C/D/-
Zhang et al. (2019a)	U-Net (Ronneberger et al., 2015)	• SAR (TerraSAR-X)	Jakobshavn Isbrae (Greenland)	-/-O
Zhang et al. (2021)	DeepLabv3+ (Chen et al., 2018)	• optical (Landsat-8, Sentinel-2) • SAR (TerraSAR-X, Envisat, ALOS-1/2, Sentinel-1)	Greenland (3 glaciers)	C/D/O
Cheng et al. (2021) (CALFIN)	DeepLabv3+ (Chen et al., 2018)	CALFIN dataset (Cheng et al., 2021) • optical (Landsat 5,7,8)	Greenland (66 glaciers)	C/D/O
Heidler et al. (2023) (COBRA)	custom; based on Deep Snake (Peng et al., 2020)			C/↑O
Loebel et al. (2022)	U-Net (Ronneberger et al., 2015)	• optical (Landsat-8) • bed topography (Bed- Machine Greenland v3 (Morlighem et al., 2017))	Greenland (23 glaciers) & Antarctica (2 glaciers)	C/D/-
Periyasamy et al. (2022)	U-Net (Ronneberger et al., 2015)	earlier version of CaFFE (see next line)		-
Gourmelon et al. (2022)	U-Net (Ronneberger et al., 2015)	CaFFE dataset (Gourmelon et al., 2022): • SAR (ENVISAT, ESR 1&2, Sentinel-1, TerraSAR-X, TanDEM-X, ALOS, RADARSAT-1)	6 glaciers from Antarctica, Greenland, and Alaska	C/D/O
Wu et al. (2023) (AMD-HookNet)	custom; two branch U-Net (Ronneberger et al., 2015)			-/↑/-
Wu et al. (2024) (HookFormer)	custom; two branch Swin Transformer (Liu et al., 2021)			-/↑/-
Zhang et al. (2023)	DeepLabv3+ (Chen et al., 2018)	• SAR (Sentinel-1) • optical (Sentinel-2, Landsat-5,7,8)	Greenland (295 glaciers)	C/-O

Table 4: Summary of DL-based studies focused on calving-front detection. For details, see Section 4.2 or the corresponding publications. The full links are also provided in Section 7

²²Note that sometimes changes to the original architecture are made, see Section 4.2.

²³Indicates whether a ‘benchmark’ dataset is used.

²⁴Often features are further derived e.g. NDVI from optical or slope from DEM.

²⁵We only indicate the main ROIs from where the training data was extracted. Still, the coverage can vary significantly, e.g. from a few image scenes to almost complete coverage. We provide the number of glaciers, where possible, but this should only be taken as a rough estimate as it depends on how glacier systems are separated into individual glaciers.

²⁶This refers to the processed training data, not to the raw one. The latter is usually openly available from the specified source. ↑ means see the data link from the previous publication.

²⁷In case the study has an output product (e.g. an inventory).

5 Discussion

In this work, we provided an overview of glacier mapping with DL, with a first part focusing on mapping (delineating) the full glacier extent (Section 4.1) and a second part detailing the automatic extraction of calving fronts (Section 4.2). Although the two fields evolved relatively independently, the utilized methodologies are generally similar, with many works relying on fully convolutional segmentation models that are well established in the DL community.

Based on the data sources that these studies use, we can conclude that glacier mapping heavily relies on data fusion, as most of the methods make use of at least two different data modalities, e.g. for glacier extent mapping, usually, at least an optical and a DEM are combined. This multimodal characteristic indicates the advantage of using DL for glacier mapping tasks, as it can automatically learn to extract useful information from each modality. We note, however, that most of the glacier mapping studies simply concatenate the input bands coming from each source, which may be sub-optimal (see, e.g. Li et al. (2022) for a review on data fusion with DL, focused on remote sensing applications). Some of the limitations we identified in the considered studies (Section 4) include:

- Some studies do not clarify how the training-test data split was performed and whether an additional validation set was used for hyper-parameter tuning. Determining this information is particularly challenging in cases where the source code is missing.
- Most of the studies that propose new methodologies (e.g. improvements for DL architectures or pre/post-processing pipelines) usually rely on different datasets, thus making it difficult to compare the added value of the proposed improvements with respect to other studies and existing methods. Additionally, a detailed ablation study (i.e. a systematic analysis where components of a model are successively removed) is often missing to empirically justify some of the methodological contributions.
- Most studies do not investigate how sensitive the proposed methods are to random initialization or the training-validation split.
- There is a risk of inconsistent/biased results in the studies that apply the same single model on the entire ROI since this implies including the results based on the data on which the model was trained.

Based on these limitations and other more general aspects, we provide a few recommendations in the next section that could be adopted in future studies. We then briefly discuss in Section 5.2 another promising area of glacier-related research where DL also starts playing a significant role, i.e. glacier mass balance and evolution modelling. We close with an outlook in Section 5.3.

5.1 Recommendations

Addressing the limitations we identified in the studies we discussed in our review could potentially facilitate model inter-comparison and accelerate the progress in the field of DL-based glacier mapping. Therefore, we here list a few recommendations, most of them being standard best practices in ML-based scientific research²⁸:

- **Open Data & Code.** Publishing both the source code and the (processed) data is crucial for advancing scientific research and promoting transparency and reproducibility. If publishing the data is not possible (e.g. due to proprietary constraints), we strongly encourage to at least publish the output of the models, e.g. as an inventory in the case of glacier extent mapping. Additionally, even if the raw data is often available from the original sources, publishing the ready-to-train data can stimulate new methodological developments and facilitate model inter-comparison. We also encourage re-using existing datasets instead of building new ones from scratch when the main goal is to propose a new methodology and quantify its performance.
- **Benchmark Datasets.** As an extension to the previous point, we encourage efforts towards building (large-scale) benchmark datasets similar to the prominent ones in ML, e.g. ImageNet (Deng et al., 2009). Such datasets enable researchers to evaluate, compare, and improve the methods, accelerating progress and innovation. See, e.g. Long et al. (2021) for guidance on how to build a benchmark dataset in Remote Sensing

²⁸For a more detailed perspective on this, we recommend, e.g. (Heil et al., 2021) and guidelines like REFORM (Reporting Standards For Machine Learning Based Science) (Kapoor et al., 2023).

applications. For calving-front detection, such examples exist, e.g. CALFIN (Cheng et al., 2021) and CaFFE (Gourmelon et al., 2022). For glacier extent mapping, the work of Maslov et al. (2024) can be considered a first attempt towards building a large-scale benchmark dataset.

- **Random Initialization/Data Split Sensitivity Analysis.** DL models are sensitive to the random weights initialization, with a non-trivial relationship between initialization and final performance (Arpit et al., 2019). To account for this effect, one should ideally also evaluate the impact of the training-validation split on the model performance, especially for cases where the weight initialization is fixed (e.g. when using pre-trained models).
- **Detailed Ablation Studies.** When proposing various methodological improvements to existing architectures, one should always perform ablation studies to evaluate the impact of each improvement, thereby providing empirical evidence for the proposed benefits. Where possible, the same applies to post-processing pipelines. Furthermore, ablation studies can also be performed to study the contribution/importance of the individual data modalities in case multiple are used.
- **Uncertainty Quantification.** Although being an active area of DL research (Abdar et al., 2021; Gawlikowski et al., 2023), some effort should be invested into quantifying the uncertainty in the predictions, e.g. by training an ensemble, and investigating the quality of the uncertainties, which can for instance be realized by checking if the uncertainties agree with the actual errors the model is making.
- **Cross-Validation with Regional Split.** When one has to use a model for an entire dataset, e.g. to build a regional inventory, we recommend training multiple models with a regional cross-validation scheme to avoid running the model on the training data. In particular, for studies focused on glacier area change analysis, regional cross-validation can prevent the risk of biased results that are a consequence of memorization (see section 4.1.2).

5.2 Deep Learning for Modelling Glacier Mass Balance and their Evolution

DL for glacier mass balance (typically relying on regression-based types of approaches) is generally at an earlier stage of development in comparison to the glacier mapping efforts (classification-based types of approaches) on which we focused in this review. The few studies that have so far used ML for glacier mass balance problems have mostly relied on more classical (i.e. non-DL) models or shallow Multi-Layer Perceptron (MLP) networks. A major issue that these studies face relates to the generally limited data availability for training, while mass balance models would, ideally, need high spatial and temporal resolution observations to be able to capture the complex interaction between climate and topography. For instance, to be able to capture the seasonal mass variability as a response to the local climatic conditions, we would need both accurate meteorological records at the glacier level, especially for precipitation, and also mass balance measurements with a (sub)seasonal frequency. This is rarely the case in practice since only a few hundred glaciers are being monitored with a (sub)annual frequency, and, second, weather stations are very sparse in mountain regions. As a result, we have to rely on geodetic MBs, which are usually available only at a multi-annual scale, thus averaging out the intraannual variability driven strongly by the local weather. Additionally, since weather measurements are limited, many studies use reanalysis data, which usually comes at a very coarse resolution, e.g. ERA5 at 0.25° resolution (around 27–28 km at the equator) (Hersbach et al., 2020).

Some of the first works on MB using ML are those of Bolibar et al. (2020a**b**), who introduced the ALpine Parameterized Glacier Model (ALPGM²⁹), a glacier evolution model that uses an MLP network with four hidden layers for MB estimation. The MLP model takes as input 34 topo-climatic predictors, e.g. monthly average temperature and snowfall, mean and max altitude, the slope of the glacier tongue, and it is trained with glacier-wide annual MBs for 32 French Alpine glaciers. By combining this ML-based MB component with a glacier evolution module (i.e. a simple geometry parametrization that models the ice dynamics), the transient evolution of glaciers can be modelled. The model was also applied to infer the evolution of the French Alpine glaciers until the end of the century, using climate models' projections under various scenarios, predicting a volume loss between 75 and 88% (Bolibar et al., 2022).

Other ML studies have also attempted to model the mass balance component of glaciers, for instance focusing on the winter mass balance (Guidicelli et al., 2023) using gradient boosting regressor (GBR) or Anilkumar et al. (2023) who

²⁹<https://github.com/JordiBolibar/ALPGM>

used point mass balances (vs. glacier-wide mass balance in works of Bolibar et al. (2020b, 2022)) through various techniques, i.e. random forest (RF), GBR, support vector machine, and MLPs. Diaconu et al. (2024) made use of the Open Global Glacier Model (Maussion et al., 2019) to reconstruct annual MBs and use the resulting dataset to systematically study various uncertainty estimation methods for ML models (e.g. the ensemble method), analysing their impact on the quality of the predictions.

Recently, DL efforts have also been focused on modelling the ice dynamical processes within glaciers. Jouvet et al. (2022) created the Instructed Glacier Model (IGM), which employs a convolutional neural network architecture to emulate the behaviour of computationally expensive ice flow models that are based on solving Navier-Stokes (NS) equations (typically referred to as 'Full-Stokes' in glaciology). Harnessing its extreme gain in computational costs (about three orders of magnitude compared to original NS calculations), IGM was then used to invert for various key components in glacier evolution modelling, such as ice thickness (Cook et al., 2023; Jouvet, 2023). In a recent update, IGM was retrained to not only reproduce 'expensive' simulations performed with NS but also to include more physical constraints (Jouvet et al., 2023). Other noteworthy recent advances for ice flow modelling through DL include the use of universal differential equations (UDEs) to model glacier flow (Bolibar et al., 2023). In the latter work, by combining an ice flow model with differential equations with an embedded neural network, a parameter (i.e. the creep component of the ice flow) can be automatically learnt from data as a nonlinear function. This approach combines the advantages of mathematical models (e.g. interpretable, incorporates domain knowledge) with those of ML (e.g. flexible, data-driven), which can also serve as a framework for discovering new empirical laws for glacier processes.

To conclude, DL applied for glacier mass balance and evolution modelling is a young but rapidly evolving field that is likely to substantially change future models for glacier projections. Like glacier mapping through DL, this field will largely benefit from new data becoming available.

5.3 Outlook

We see many promising directions in the field of automatic glacier mapping based on DL. First, we expect (and encourage) the development of large-scale benchmark datasets that will accelerate the methodological developments. Second, we recommend using time-series input data, as this can potentially mitigate the errors caused by (partial) occlusions in single images, e.g. by clouds. Alternatively, single-image model approaches can be used on multiple acquisitions, from which a (temporal) prediction ensemble can be built. Third, we suggest exploring more sophisticated data fusion techniques instead of simply concatenating the input bands from all the sources, as in most current studies.

We also expect a rapid development of DL-based methods for glacier evolution modelling, briefly discussed here in Section 5.2. These developments will largely profit from the increasing number of (global) products in the field, e.g. Hugonnet et al. (2021), with many other promising datasets currently being finalized (Dussaillant et al., 2024; Zemp et al., 2024). Additionally, given the complexity of glacier-climate interactions, we advocate for developing methods that can make use of expert knowledge, e.g. by including physical constraints in the models, with the recent work by Bolibar et al. (2023) as an example.

6 Summary

In this work, we provided an overview of the glacier mapping literature based on DL, highlighting the methodological contributions, the regions on which each study is focused and the data modalities used for training. We divided the studies into two major sub-fields in glacier mapping, which evolved in parallel: i) glacier extent mapping and ii) calving front detection. Both types of classification problems are usually tackled using multi-source datasets, showing the benefit of using DL for automatically extracting the relevant features from each modality. We then provide a compact summary of the available resources (data and source codes) to facilitate further experimentation.

7 Resources

In Table 5, we compiled a list of general resources useful for glacier-related studies, most of them having been used by the studies discussed in this work. Additionally, in Table 6 we provide the full links to the code, data and generated outputs from the studies that make them publicly available (note that for the Web version of this article, the same information is also provided in Tables 1 to 4 as hyperlinks).