Image classification of marine-terminating outlet glaciers in Greenland using deep learning methods

Melanie Marochov, Chris R. Stokes, Patrice E. Carbonneau

Department of Geography, Durham University, Durham, DH1 3LE, UK

Correspondence: Melanie Marochov (melanie.marochov@durham.ac.uk) and Patrice E. Carbonneau (patrice.carbonneau@durham.ac.uk)

Abstract. A wealth of research has focused on elucidating the key controls on mass loss from the Greenland and Antarctic ice sheets in response to climate forcing, specifically in relation to the drivers of marine-terminating outlet glacier change. The manual methods traditionally used to monitor change in satellite imagery of marine-terminating outlet glaciers are time-consuming and can be subjective, especially where mélange exists at the terminus. Recent advances in deep learning applied to image processing have created a new frontier in the field of automated delineation of glacier calving fronts. However, there remains a paucity of research on the use of deep learning for pixel-level semantic image classification of outlet glacier environments. Here, we apply and test a two-phase deep learning approach based on a well-established convolutional neural network (CNN) for automated classification of Sentinel-2 satellite images. The novel workflow, termed CNN-Supervised Classification (CSC) is adapted to produce multi-class outputs for unseen test imagery of glacial environments containing marine-terminating outlet glaciers in Greenland. Different CNN input parameters and training techniques are tested, with overall F1 scores for resulting classifications reaching up to 94% for in-sample test data (Helheim Glacier) and 96% for out-of-sample test data (Jakobshavn Isbrae and Store Glacier), establishing a state-of-the-art in classification of marine-terminating glaciers in Greenland. Predicted calving fronts derived using optimal CSC input parameters have a mean deviation of 56.17 m (5.6 pixels) and median deviation of 24.7 m (2.5 pixels) from manually digitised fronts. This demonstrates the transferability and robustness of the deep learning workflow despite complex and seasonally variable imagery. Future research could focus on the integration of deep learning classification workflows with free cloud-based platforms, to efficiently classify imagery and produce datasets for a range of glacial applications without the need for substantial prior experience in coding or deep learning.

1 Introduction

Quantifying glacier change (e.g., volume, area, geometry, surface hydrology, and terminus position) from remote sensing data is essential to improve our understanding of the impacts of climate change on glaciers (Vaughan et al., 2013; Hill et al., 2017). In many glaciated areas, well-established semi-automated techniques such as image band ratio methods are used to extract glacier outlines for this purpose and to create glacier inventories (Paul et al., 2016). These methods accurately classify areas of debris-free ice in contrast to surrounding topography and are widely used in studies of mountain glaciers and ice caps (e.g., Bolch et al., 2010; Frey et al., 2012; Rastner et al., 2012; Guo et al., 2015; Stokes et al., 2018). However, these approaches are less effective for accurately mapping more complex glaciers and glaciated landscapes such as marine-terminating outlet glaciers which often have seasonally variable areas of a spectrally similar mélange (a mixture of sea-ice and icebergs) near their calving fronts (e.g., Amundson et al., 2020). As a result, manual digitisation, and more recently, bespoke (semi-) automated techniques have been relied upon for delineation of glaciers in more complex settings (e.g., Robson et al., 2015; Baumhoer et al., 2019; Mohajerani et al., 2019; Zhang et al., 2019; Cheng et al., 2021).

The importance of processes occurring at marine-terminating outlet glaciers on a range of timescales (Amundson et al., 2010; Juan et al., 2010; Chauché et al., 2014; Carroll et al., 2016; Bunce et al., 2018; Catania et al., 2018, 2020; King et al., 2018; Bevan et al., 2019; Sutherland et al., 2019; Tuckett et al., 2019) highlights the growing need for a method to efficiently quantify outlet glacier change in an era of increasingly available satellite data. Since manual digitisation remains the most common technique used to delineate marine-terminating glaciers (e.g., Miles et al., 2016, 2018; Carr et al., 2017; Wood et al., 2018; Brough et al., 2019; Cook et al., 2019; King et al., 2020), studies which analyse seasonal glacier dynamics are often limited to individual glaciers or small spatial areas due to the labour-intense and time-consuming nature of the method (Seale et al., 2011). In contrast, where studies encapsulate larger numbers of glaciers over increased spatial areas, monitoring is often constrained to inter-annual to decadal scales (e.g., Moon and Joughin, 2008), removing the opportunity to understand seasonal changes and drivers.

To confront this challenge, some specialised automated techniques for extracting ice fronts have been developed, exemplified in a small number of studies which delineate the boundaries of marine-terminating glaciers and ice shelves at the margins of the Greenland (Sohn and Jezek, 1999; Seale et al., 2011; Krieger and Floricioiu, 2017) and Antarctic ice sheets (Liu and Jezek, 2004; Yu et al., 2019). These methods generally rely on tools from the fields of image processing and computer vision, namely semantic segmentation, and edge detection. Semantic segmentation is a term used interchangeably with pixel-level semantic classification and refers to the process of dividing an image into its constituent parts based on groups of pixels of a given class, assigning each pixel a semantic label (Liu et al., 2019). Throughout the remainder of this study, we refer to this generally as classification. The technique was used by Liu and Jezek (2004) to partition Synthetic Aperture Radar (SAR) imagery into two major semantic classes (ice/land and water). Following substantial post-processing, they applied an edge detection algorithm

to the classified image to extract the boundary between ice/land and water around Antarctica. Edge detection identifies areas in an image with abrupt changes in pixel brightness, therefore providing a useful tool to detect boundaries from satellite imagery (Chen and Hong Yang, 1995). In further work, Seale et al. (2011) applied an edge detection algorithm to Moderate Resolution Imaging Spectroradiometer (MODIS) satellite imagery of Greenland to detect glacier calving fronts with a similar level of accuracy to manual digitisation. Despite adequate levels of accuracy, edge detection techniques require substantial pre- and post-processing, and have since only been used for calving front delineation in a few studies (e.g., Joughin et al., 2008a; Christoffersen et al., 2012). Meanwhile, traditional statistical classification techniques (e.g., maximum likelihood) are not considered robust when it comes to extracting ice fronts where there is little contrast between glacier ice/ice shelves and spectrally similar areas of mélange or even water containing icebergs (Baumhoer et al., 2019). Moreover, classification techniques which rely solely on individual pixel values often miss contextual, class representative shapes and textures, meaning that for land cover classification of medium resolution satellite imagery, pixel-based approaches rarely produce satisfactory levels of accuracy (Blaschke et al., 2000) and commonly result in noisy classifications (Li et al., 2014). Therefore, edge detection and traditional pixel-based classification methods have not yet overcome the widespread use of manual digitisation for monitoring marine-terminating outlet glaciers.

More recently, deep learning methods have been developed to extract ice front outlines and overcome these drawbacks (Baumhoer et al., 2019; Mohajerani et al., 2019; Zhang et al., 2019, Cheng et al., 2021). Deep learning is a type of machine learning in which a computer learns complex patterns from raw data by building a hierarchy of simpler patterns (Goodfellow et al., 2016). Convolutional Neural Networks (CNNs) are deep learning models specifically designed to process multiple 2D arrays of data such as multiple image bands (LeCun et al., 2015). They differ from conventional classification algorithms based solely on the spectral properties of individual pixels by detecting the contextual information of images such as texture, in the same way a human operator would. Previous studies which apply deep learning to detect the calving fronts of marine-terminating glaciers used a type of CNN called a Fully Convolutional Neural Network (FCN) (Ronneberger et al., 2015), and various post-processing techniques to extract the boundaries between 1) ice and ocean in Antarctica (Baumhoer et al., 2019), and 2) marine-terminating outlet glaciers and mélange/water in Greenland (Mohajerani et al., 2019; Zhang et al., 2019, Cheng et al., 2021). Calving fronts detected using these methods have mean errors from manual delineations ranging from 38 to 108 m (<2 to 6 pixels), providing an accurate automated alternative to manual digitisation.

These deep learning methods have so far relied on a binary classification of input images. For example, Baumhoer et al. (2019) used only two classes (land ice and ocean), as did Zhang et al. (2019) who classified the input image into ice mélange regions and non-ice mélange regions (the latter including both glacier ice and bedrock). While these methods are incredibly useful for extracting glacier and ice shelf fronts to quantify fluctuations over time, they perhaps overlook the ability of deep learning methods to create highly accurate image classification outputs which contain more than two classes (i.e., not just ice and no-ice areas). Moreover, deep learning has been used successfully in other disciplines to classify entire landscapes or image scenes

to a high level of accuracy (Sharma et al., 2017; Carbonneau et al., 2020a). Despite this, multi-class image classification of entire marine-terminating outlet glacier environments has not yet been tested using deep learning.

CNNs have achieved success in mapping debris-covered land-terminating glaciers (Xie et al., 2020), rock glaciers (Robson et al., 2020), supraglacial lakes (Yuan et al., 2020) and snow cover (Nijhawan et al., 2019), but the use of deep learning in glaciology is still in its infancy. Given the abundance of available satellite imagery, deep learning methods could be a significant aid in the automation of image processing for marine-terminating glacial settings. Image classification using deep learning techniques has the clear potential to not only reduce the labour-intensive nature of manual methods but facilitate automated analysis in numerous research areas. Aside from calving front delineation, a method which quickly produces accurate multi-class image classifications of complex and seasonally variable outlet glacier environments could provide an efficient way to further elucidate processes and interactions controlling outlet glacier behaviour at high temporal resolution (e.g., calving events, the buttressing effects of mélange, subglacial plumes, and supra-glacial lakes). The compatibility of deep learning image classification methods with platforms such as Google Earth Engine (Gorelick et al., 2017) and its integration with Geographic Information Systems (GIS) software could also improve the efficiency of such analysis and remove the need for prior expertise in deep learning and coding. This in turn could allow the incorporation of a more detailed understanding of marine-terminating outlet glacier dynamics and interactions in models used to project future sea-level changes (Csatho et al., 2014).

The aim of this paper is to adapt a two-phase deep learning method which was originally developed to classify airborne imagery in fluvial settings (Carbonneau et al., 2020a) and test it on satellite imagery of marine-terminating outlet glaciers in Greenland. We first modify and train a well-established CNN called VGG16 (Simonyan and Zisserman, 2015) using labelled image tiles from 13 seasonally variable images of Helheim Glacier, south east Greenland. In the first phase of the workflow, this transferable, pre-trained model is applied to an unseen image from an outlet glacier environment. The resulting class predictions are then used as training data for a phase two model which is specific to the unseen input image. This phase two model produces a final pixel-level classification, from which calving front outlines are detected and error is estimated from manually delineated validation labels. We assess the sensitivity of the classification workflow to different image band combinations, training techniques, and model parameters for fine-tuning and transferability. Our objective is to establish and evaluate a workflow for multi-class image classification for glacial landscapes in Greenland which can be accessed and used rapidly without having specialised knowledge of deep learning or the need for time-consuming generation of substantial new training data. Furthermore, we aspire to exceed the current state-of-the-art and advance accuracy levels (F1 scores >90%) for pixel-level image classification of marine-terminating outlet glacier landscapes in Greenland. The methods developed here are trained and tested on outlet glaciers in Greenland with a pre-defined set of image classes. However, in future work the workflow may be applicable to mapping outlet glaciers elsewhere in the world, dependant on suitable adaptations to training data inputs and further fine-tuning.

2 Methods

2.1 Overview of CNN-Supervised Classification

The classification workflow used here is termed CNN-Supervised Classification (CSC), and was originally developed and tested on airborne imagery (<10 cm resolution) to produce pixel-level landcover classifications of fluvial scenes (Carbonneau et al., 2020a). CSC is a two-phase workflow based on convolutional architectures which concatenates a CNN to a multilayer perceptron (MLP) or compact CNN (cCNN) to produce pixel-level classifications. The two-phase approach was designed to simulate traditional supervised classification techniques (Carbonneau et al., 2020a). In effect, a pre-trained CNN is used in the first phase of CSC to automatically detect training areas for each individual input image instead of manual collection of training data, which is typically required for traditional supervised machine learning classifiers (Carbonneau et al., 2020a). The phase one CNN removes the need for intensive manual digitisation for every new image and its critical role is to produce training data which is locally specific to each image. In other words, the CNN accounts for image heterogeneity, and incorporates the specific illumination/weather conditions, acquisition angles and seasonal characteristics of each unseen image by detecting local predictive features like brightness, texture, and geometric features (e.g., crevasses) in relation to class. The predictions of the phase one CNN therefore provide bespoke training data for pixel-level image classification in phase two.

The pre-trained CNN applied in phase one of CSC falls into the category of supervised learning (Goodfellow et al., 2016) and is trained with a sample of image tiles which have been manually labelled according to class (training dataset). Each tile used to train the phase one CNN represents a sample of pure class (i.e., one class covers over 95% of the tile area) allowing the CNN to learn predictive features associated with class, and subsequently make class predictions for a tiled input image not previously seen in training (test dataset). During phase one of CSC, unseen test images are tiled and encoded in the form of 4D tensors which contain several separate tiles (Dimensions: Tiles, X, Y, Image bands). The pre-trained phase one CNN predicts a class for each input tile and the tiles are subsequently re-assembled in the shape of the original input image (Fig. 1). As shown in Fig. 1, this produces a one band class raster made up of tiles, each of which is denoted by a single integer number representing its predicted class. In the second phase of CSC, this class raster and input image features are used to train a robust second model specific to the unseen input image. The predictions of this phase two model result in a final, pixel-level image classification (Fig. 1).

Since the phase one CNN predictions take the form of a tiled class raster, it is expected that individual tiles may straddle more than one class and result in inaccurate class boundaries. As a result, this will generate error in the phase two training data. However, the phase two models are robust to noise and have been shown to overcome these errors (Carbonneau et al., 2020a) with resulting pixel-level classifications following class boundaries much more accurately.

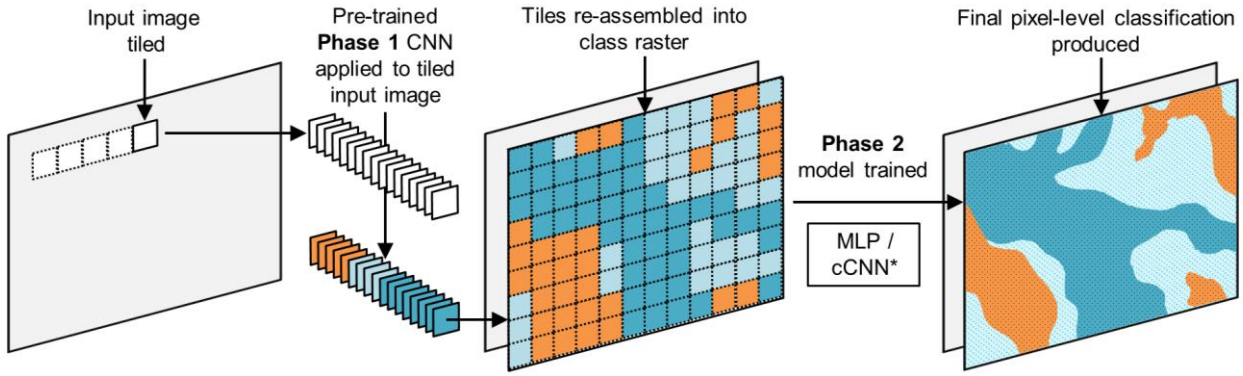


Figure 1: Conceptual diagram of the CNN-Supervised Classification workflow showing the production of a tiled class raster in phase one which is used as image-specific training data for the phase two model. The predictions of the phase two model produce a final pixel-level classification.

2.2. Study areas

2.2.1 Training area: Helheim Glacier, SE Greenland

An area spanning 68.8 x 37.2 km (6875 x 3721 pixels) (Fig. 2a) which includes Helheim Glacier (66.4° N, 38.8° W), a major outlet of the south-eastern Greenland Ice Sheet (GrIS), was chosen to adapt CSC for classification of marine-terminating outlet glacier landscapes and train the phase one CNN. Helheim is one of the five largest outlet glaciers of the GrIS by ice discharge (Howat et al., 2011; Enderlin et al., 2014) and has flow speeds of 5-11 km a^{-1} (Bevan et al., 2012). The glacier has a 48,140 km^2 drainage basin (Rignot and Kanagaratnam, 2006) equivalent to ~4% of the ice sheet's total area (Straneo et al., 2016), from which several tributaries converge into a ~6 km wide terminus. There is an extensive area of ice mélange (a mixture of sea-ice and icebergs) adjacent to the terminus where it enters Sermilik Fjord and is influenced by ocean currents (Straneo et al., 2016) (Fig. 2a). Inspection of available satellite imagery from 2019 revealed that the area of mélange varied seasonally with monthly variations in extension and composition as previously observed (Andresen et al., 2012, 2013). Additionally, a gap in the mélange at the glacier terminus appeared at the beginning of July and persisted until mid-August, suggesting the presence of an active meltwater-fed glacial plume (Straneo et al., 2011).

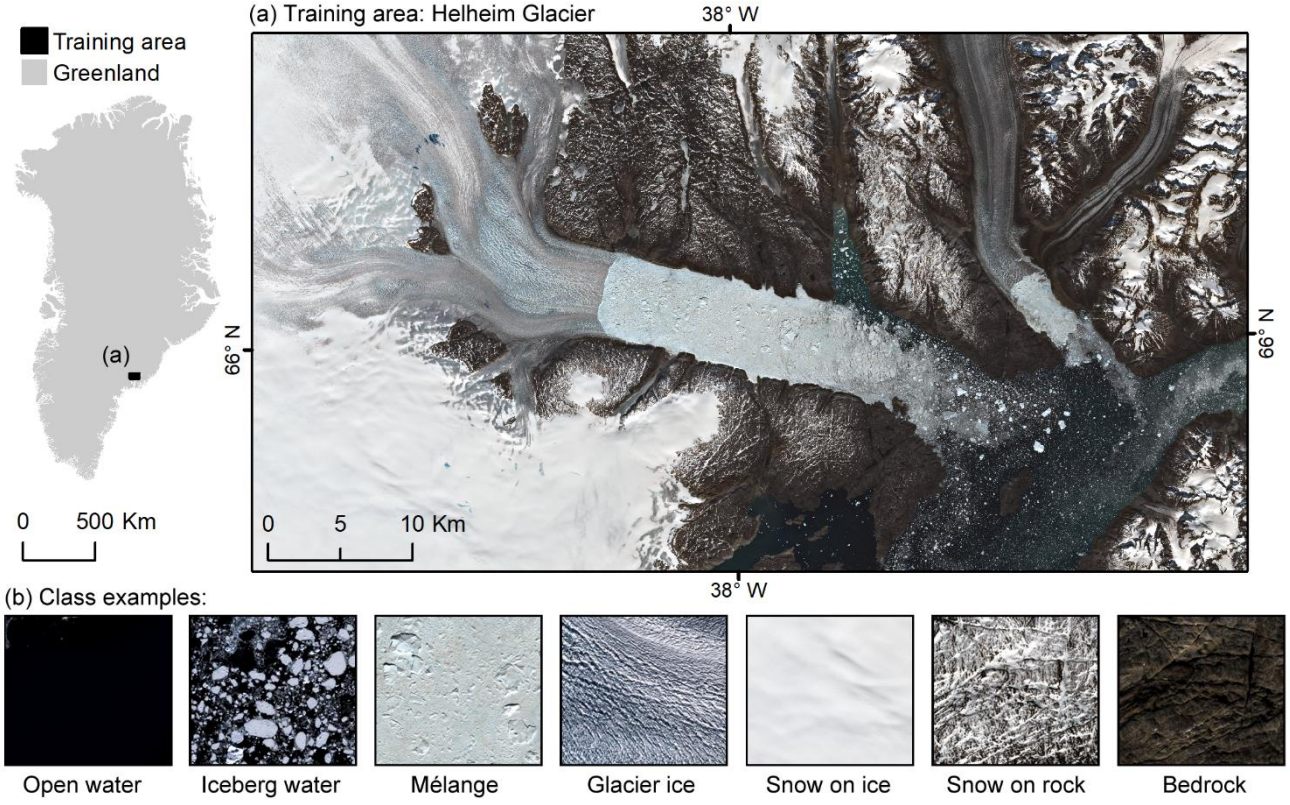


Figure 2: (a) Location of the area from which phase one CNN training data was extracted, showing Helheim glacier and the surrounding landscape. Sentinel-2 image acquired on 15 June 2019. (b) Shows example image samples for each of the seven semantic classes used to train the phase one CNN. The outline of Greenland is from Gerrish (2020).

The glacier, fjord, and surrounding landscape provide an ideal training area for the deep learning workflow because it contains a number of diverse elements that vary over short spatial and temporal scales and are typical of other complex outlet glacier settings in Greenland. These characteristics include 1) seasonal variations in the degree of surface meltwater on the glacier and ice mélange; 2) weekly to monthly changes in the extent and composition of mélange; 3) short-lived, meltwater-fed glacial plumes which result in polynyas adjacent to the terminus; 4) sea-ice in varying stages of formation; 5) varying volumes and sizes of icebergs in fjord waters and 6) seasonal variations in snow cover on both bedrock and ice. The resulting spectral variations over multiple satellite images in addition to potential variations resulting from changes in illumination and weather, pose a considerable challenge to image classification. However, capturing these characteristics at the scale of an entire outlet glacier image scene is important for a more efficient and integrated understanding of how numerous glacial processes interact. Examination of imagery showing the seasonal change of the glacial landscape throughout the year resulted in the establishment of seven semantic classes, including: 1) open water, 2) iceberg water, 3) mélange, 4) glacier ice, 5) snow on ice, 6) snow on rock, and 7) bare bedrock (see class examples in Fig. 2 and detailed criteria for each in Table 1). Training and validation data

for the phase one CNN applied in CSC was collected from the Helheim study area shown in Fig. 2 and labelled according to these seven classes.

Table 1: Descriptions of each of the seven semantic classes used to train the phase one CNN in the deep learning workflow. Example image samples of each class can be found in Figure 2.

Class	Class description
1. Open water	Open water with no icebergs
2. Iceberg water	Water with varying amounts of icebergs or disintegrated mélange/sea-ice
3. Mélange	Mixture of sea-ice and icebergs of varying sizes
4. Glacier ice	Glacier ice, with seasonally variable surface meltwater
5. Snow on ice	Snow/ice with a smooth appearance
6. Snow on rock	Bedrock with varying amounts of snow cover
7. Bedrock	Bedrock with no snow cover

2.2.2 Test areas: Helheim, Jakobshavn, and Store Glaciers

The ability of a model to accurately predict the class of pixels in an unseen test image is called generalisation (Goodfellow et al., 2016) and determines the transferability of the model. To test the transferability of the CSC workflow adapted for marine-terminating glacial landscapes in Greenland, we apply CSC to a test dataset composed of seasonally variable imagery from in-sample and out-of-sample study sites (Fig. 3). CSC is never tested on any image that was used in training. Rather the in-sample test dataset is compiled of images from the same glacier used in training, but acquired on different dates to the training data. The in-sample test site includes Helheim Glacier (Helheim) and has a slightly smaller area (47.1 x 39.9 km, or 4711 x 3986 pixels) compared to the training site (Fig. 3a).

The out-of-sample test areas contain Jakobshavn Isbrae (Jakobshavn) and Store Glacier (Store) in central west (CW) Greenland, and they represent outlet glacier landscapes never seen during training (Fig. 3b and c). The Jakobshavn site spans 35.7 x 22.7 km (3566 x 2265 pixels) while the Store site spans 28 x 20.9 km (or 2797 x 2089 pixels). Both out-of-sample test sites have notably different characteristics compared to the Helheim site, specifically in terms of glacier, terminus, and fjord shape, providing an adequate test of spatial transferability. Jakobshavn is the largest (by discharge) and fastest flowing outlet of the GrIS (Mouginot et al., 2019). The glacier discharges 45% of the CW GrIS (Mouginot et al., 2019) and has been undergoing terminus retreat, thinning, and acceleration over the past few decades (Howat et al., 2007; Joughin et al., 2008b). As a result, the terminus of Jakobshavn is composed of two distinct branches which are no longer laterally constrained by fjord walls in the same manner as Helheim. Store Glacier is responsible for 32% of discharge from the CW GrIS (Mouginot et al.,

2019), but has remained relatively stable over the last few decades (Catania et al., 2018). The calving front of Store is laterally constrained by the walls of Ikerasak Fjord (Fig. 3c) and both Jakobshavn and Store glaciers have different flow directions in comparison to Helheim. The seven classes identified from the training area are also present in the out-of-sample test sites, including mélange which continuously occupied the fjord at Jakobshavn, and was sporadically present in front of Store Glacier throughout the range of test imagery acquired in 2020 (Fig. 3).

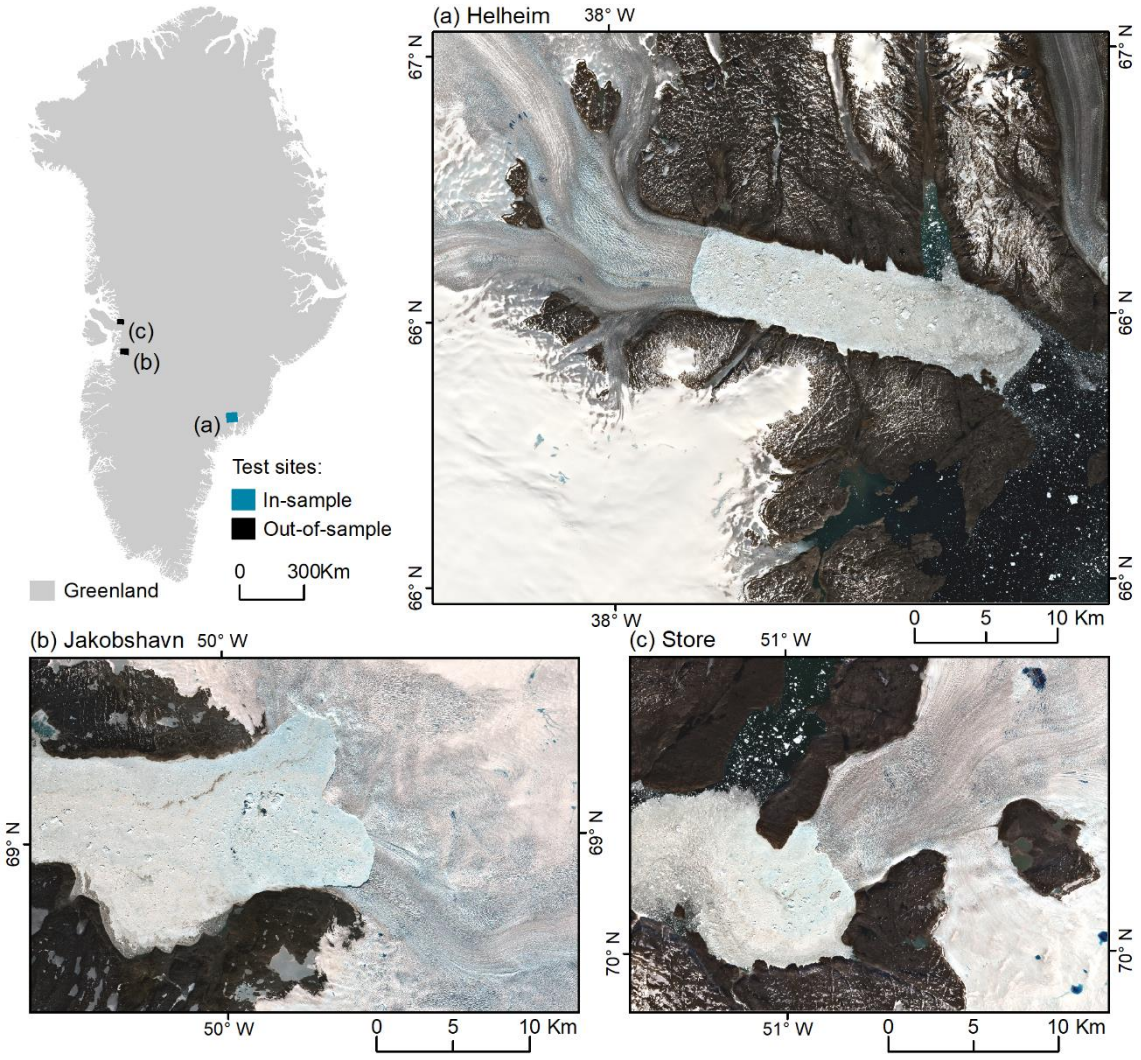


Figure 3: Test areas used to quantify the transferability of the CSC workflow. (a) The in-sample test area including Helheim Glacier. Example image acquired on 18 June 2019. (b) The out-of-sample test areas of Jakobshavn Isbrae (example image acquired on 21 May 2020) and (c) Store Glacier (example image acquired on 28 June 2020). The outline of Greenland is from Gerrish (2020).

2.3 Imagery

To train and test the CSC workflow adapted for marine-terminating glacial landscapes, Sentinel-2 image bands 2 (blue), 3 (green), 4 (red), and 8 (Near Infrared (NIR)) were used at 10 m spatial resolution. The red, green, and blue (RGB) bands were chosen because they are commonly used in image classification with deep learning architectures, making existing, pre-trained, models easily transferable for the purpose of this study. Additionally, snow and ice have high reflectance in the NIR band which is often used in remote sensing of glacial environments, for example in band ratios to automatically identify glacier outlines (e.g., Alifu et al., 2015). As a result, the CSC workflow was tested using both RGB and RGB+NIR band combinations.

Cloud cover and insufficient solar illumination present challenges when using optical satellite imagery such as Sentinel-2 data and resulting data availability is limited to cloud-free imagery that spans from February to October for our study areas. Despite these limitations, optical image availability still provides sufficient data to train and test CSC on seasonal timescales. Therefore, to best encompass the seasonally variable landscape characteristics and collect sufficient training data to represent intra-class variation in all seven classes, 13 cloud-free Sentinel-2 images of the Helheim training area, taken between February and October 2019, were acquired for phase one CNN training (Table S1 in the Supplement). Similarly, for the test dataset we aimed to acquire seasonally variable imagery, resulting in a dataset of nine in-sample test images acquired on different dates to training data in 2019, and 18 out-of-sample test images acquired from February to October 2020 (Table S1 in Supplement). Level-2A images were downloaded from Copernicus Open Access Hub (available at: https://scihub.copernicus.eu/dhus/#/home, last accessed: 20/07/20) and a simple set of pre-processing steps were applied. First, the RGB and NIR bands were combined into composite four band images and saved in GeoTIFF format. Second, the images were cropped to the training and test areas. Additionally, to test CSC over a larger spatial scale (i.e., more than a single glacier), we also created a four-band composite image for a whole unseen Sentinel-2 tile which included the entire landscape surrounding Helheim Glacier (10980 x 10980 pixels) acquired on 13 September 2019.

2.4 CSC model architectures and training

2.4.1 Phase 1: model architecture

For the base architecture of the pre-trained CNN used in phase one of CSC we adapt a well-established CNN called VGG16 (Simonyan and Zisserman, 2015) which outperformed the state-of-the-art performance of AlexNet in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014. The VGG model we use consists of five stacks of 13 2D convolutional layers which have filters with a 3x3 pixel kernel size (Fig. 4). The filter spatially convolves over the input image to create a feature map, using the filter weights. The dimensions of the output filters increase from 64 in the first stack of convolutional layers to 512 in the last (Fig. 4). All the convolutional layers use rectified linear unit (ReLU) activation and are interspersed with five max-pooling layers. The convolutional and pooling stacks are followed by three fully connected (dense) layers (i.e.,

a normal fine-tuned neural network) without shared weights, typical of CNN architectures. This section allows the features learned by the CNN to be allocated to a class by a final Softmax layer with the same number of units as classes. The dense layers use L^2 regularisation to reduce over-training (Goodfellow et al., 2016; Carbonneau et al., 2020a). The input image tile size for the first convolutional layer in the original VGG16 model architecture was fixed as a 224x224x3 RGB image. However, here we test the impact of tile size by using three datasets with different tile sizes of 50x50, 75x75, and 100x100 pixels. Thus, we adjust the input image size, so it matches our three tile sizes (Fig. 4 shows an example of an input tile size of 100) and adjust the number of input channels depending on the number of image bands used for training (i.e., three: RGB or four: RGB+NIR).

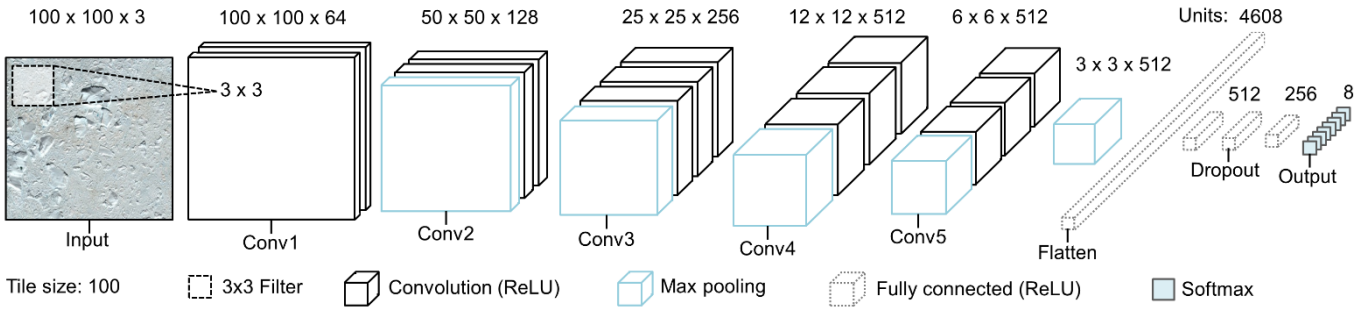


Figure 4: Architecture of phase one convolutional neural network, adapted for three tile size datasets from the original VGG16 model architecture (Simonyan and Zisserman, 2015). Diagram shows an example of an RGB tile of 100x100 pixels. There are five stacks of 2D convolutional layers (Labelled 'Conv#') which extract features from input tiles using a 3x3 filter. The convolutional stacks are followed by a fully connected neural network and Softmax activation for final class predictions which are subsequently used as localised training data for phase two models.

2.4.2 Phase 1: model training

We tested several different training inputs for the phase one CNN in order to find the best performing input parameters. Firstly, we tested training the CNN with tiles composed of 1) RGB bands, and 2) RGB+NIR bands. Secondly, we tested three different input tile sizes of 50x50 pixels, 75x75 pixels and 100x100 pixels to find the best tile size for identifying landscape features at the scale of the 10 m resolution imagery. Overall, this produced six pre-trained phase one CNNs to test.

To train each of the six CNNs, we employed early stopping to control hyperparameters and inhibit overfitting which occurs when a model is unable to generalize between training and validation data (Goodfellow et al., 2016). To do this, we designed a custom callback that trains the network until the validation data (20% set aside with a train-validate-split) reaches a desired target accuracy threshold. These targets ranged from 92.5 to 99% and determined the number of epochs each CNN was trained for. We used categorical cross entropy as the loss function and Adam gradient-based optimisation (Kingma and Ba, 2017) with

a learning rate of 10×10^{-4} and batch sizes of 30. Training the phase one CNNs took under an hour using an I7 processor at 5.1Ghz, and an Nvidia RTX 2060 GPU.

When applying CSC to multiple sites, we came to a similar conclusion to Carbonneau et al. (2020a) which found that model transferability was improved when the phase one CNN was trained with data from more than one site. We therefore deployed a joint fine-tuning training procedure where a CNN initially trained only on data from Helheim was trained further with a small set of extra tiles (5,000 samples per class) using only two images (one from winter and one from summer) for all 3 glaciers. This fine tuning was done at a low learning rate of 10×10^{-5} and smaller batch sizes of 10 in comparison to initial CNN training (which used a learning rate of 10×10^{-4} and batch size of 30). The rational for this is that if a glacier is identified for monitoring, the addition of two available scenes to produce data used to fine-tune an existing CNN is not an onerous task and can deliver significant improvements to the final results. For clarity, we will refer to CNN training without this extra level of fine-tuning as 'Single' training and CNN training with this added fine-tuning as 'Joint' training. We test the Joint training by applying it with tile sizes of 50x50 pixels and RGB+NIR bands due to the good general performance of these parameters during Single training. Alongside the six phase one CNNs with Single training, this resulted in an additional glacier-specific CNN with Joint training for each of the three test areas.

2.4.3 Phase 1: model training data production

A dataset of 210,000 training samples with 30,000 image tiles per class was used to train and validate each phase one CNN. To create the training tiles, the cropped four-band (RGB+NIR) images extracted from 13 Sentinel-2 acquisitions were manually labelled according to the seven semantic classes using QGIS 3.4 digitising tools. Vector polygons labelled by class number were rasterised to produce a per-pixel class raster the same size as the training area. Both the input image and class raster were then tiled using a script which extracted tiles of 100x100 pixels with high overlap using a stride of 20 (number of pixels the window moves before extracting another tile) (Fig. 5). Each tile was extracted, assigned a class label based on the class raster and any tiles occupied by less than 95% pure class were rejected, removing tiles containing mixed classes. Once extracted, each tile was augmented by three successive rotations of 90 degrees (Fig. 5). Data augmentation is a common step for bolstering training datasets in deep learning, and usually entails slightly altering existing data to increase the number of training samples (Chollet, 2017). Tile rotation also allows the model to learn classes which may appear at different orientations in unseen images, for example accounting for different glacier flow directions, providing the potential for increased transferability. Following augmentation, tiles were normalised by a constant value of 8192 to convert raw Sentinel-2 data to 16-bit floating point data and saved to disk in TIF format. This was because a GPU with a Turing architecture was used in CNN training, enabling the use of the TensorFlow mixed precision training method for which the input is 16-bit floating point data. The tiles were randomly allocated to training and validation folders with an 80/20% training-validation split for phase one CNN training.

Overall, this resulted in a dataset upwards of 1 million tiles with a large imbalance that ranged from 50,000 tiles in class one to 900,000 tiles in class four. However, class imbalance can have negative impacts on model performance (Johnson and Khoshgoftaar, 2019), so we then drastically cut the tile population and randomly subsampled 30,000 tiles from each class, thus resulting in a balanced training dataset. The final number of 30,000 tiles per class was chosen after trial and error revealed that we could run all needed CNN models with all the tiles loaded in an available RAM space of 64GB with a 32GB paging file. These populations of 100x100 pixel tiles composed of RGB+NIR bands were then sliced as needed to create the tiles of 50x50 or 75x75 pixels in either RGB or RGB+NIR format. For the Joint fine tuning of phase one CNNs, a small dataset of 5,000 samples per class was extracted from a single winter image and a single summer image for each of the three glaciers.

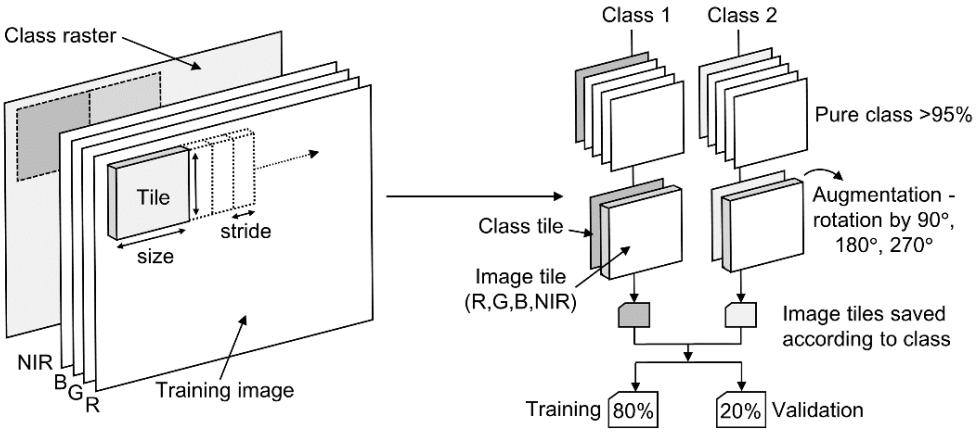


Figure 5: Conceptual diagram of tiling process used to create training and validation data. A specified tile size (100x100 pixels) and stride (20 pixels) were used to extract tiles from the class raster and image bands. These tiles were filtered and augmented and saved to individual class folders using an 80/20% split for training and validation data.

2.4.4 Phase 2: model architectures and training

To classify airborne imagery of fluvial scenes at pixel-level using the CSC workflow, Carbonneau et al. (2020a) applied a pixel-based approach using an MLP in the second phase of the workflow, achieving high levels of accuracy (90-99%). We propose that applying pixel-based techniques to coarser resolution imagery such as Sentinel-2 data may be less effective compared to applying the workflow to high resolution imagery. Furthermore, particularly in landscapes containing marine-terminating glaciers, many distinct classes may be covered in snow or ice and therefore be very spectrally similar (i.e., all classes are white), and where this is the case a pixel-based MLP would predictably struggle to differentiate between classes. An example of this is shown in Fig. 11a in the results where the majority of the image is snow-covered and different classes are very spectrally similar. We therefore adopt a patch-based approach which uses a small window of pixels to determine the class of a central pixel, as in Sharma et al. (2017). This approach is based on the idea that a pixel in remotely sensed imagery is spatially dependent and likely to be similar to those around it (Berberoglu et al., 2000). Sharma et al. (2017) use a patch size

of 5x5 pixels for patch-based classification of medium resolution Landsat 8 imagery. This use of a region instead of a single pixel allows for the construction of a small CNN (dubbed 'compact CNN' or cCNN: Samarth et al., 2019) with fewer convolutional layers that assigns a class to the central pixel according to the properties of the region (Carbonneau et al., 2020b). It therefore combines spatial and spectral information. Here we test both pixel- and patch-based approaches using an MLP and cCNN in the second phase of the workflow (the architectures and application of which are detailed in the following sections 2.4.4.1 and 2.4.4.2). Specifically, five patch sizes of 1x1 (pixel-based), 3x3, 5x5, 7x7, and 15x15 pixels were tested.

2.4.4.1 Multilayer Perceptron

For the pixel-based classification in phase two we use an MLP (Fig. 6). An MLP is a typical deep learning model (also commonly known as an artificial neural network (ANN)) which consists of three (or more) interconnected layers (Rumelhart et al., 1986; Berberoglu et al., 2000). The MLP used here has five layers consisting of four fully connected (dense) layers and one batch normalisation layer (Fig. 6). The first dense layer has the same number of input dimensions as image bands and 64 output filters. This is followed by a batch normalization layer which helps to reduce overfitting in a similar way to dropout layers, by adjusting the activations in the network to add noise. This is followed by two more dense layers with 32 and 16 filters, respectively. Each dense layer uses L^2 regularisation and ReLU activation except the output layer. The final output layer in the network is a dense layer with Softmax activation and eight output filters, to match the number of output classes. We used categorical cross entropy as the loss function and Adam gradient-based optimisation (Kingma and Ba, 2017) with a learning rate of 10×10^{-3}.

We used conventional early stopping with a patience parameter and a minimum improvement threshold to train the MLP. The minimum improvement was set as 0.5%. For the MLP, we found that training did not stabilise for at least 20 epochs and we set the patience to 20. This means that if training does not improve the validation accuracy by 0.5% after a period of 20 epochs, the training will stop. Since the MLP is pixel-based, the number of parameters is smaller compared to the patch-based model, with 3,128 and 3,192 trainable parameters for RGB and NIR+RGB imagery, respectively.

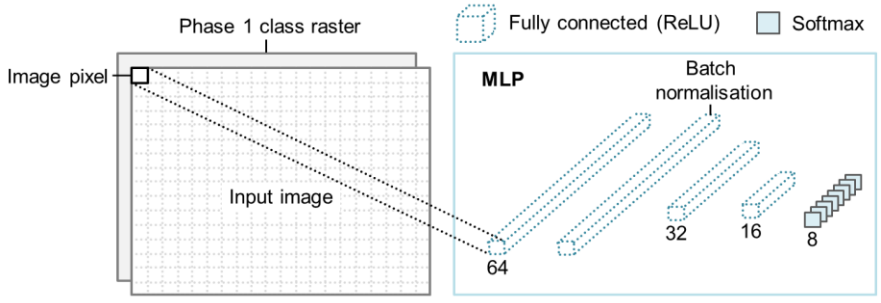


Figure 6: Architecture of phase two multilayer perceptron used for pixel-based classification.

2.4.4.2 Compact Convolutional Neural Network

For the patch-based classification in phase two we use a cCNN (Fig. 7). We refer to this model architecture as a compact CNN (cf. Samarth et al., 2019) because the cCNN architecture contains fewer convolutional layers in comparison to conventional CNNs. The cCNN trains to learn the class of a central pixel in a patch as a function of its neighbourhood. So, for each pixel in the input image, a small image tile is extracted with square dimensions of the patch size (e.g., 3x3, 5x5, 7x7, or 15x15 pixels). The central pixel from the phase one predicted class raster is used as the associated class label. The number of image bands in each patch is determined by which band combination is being tested (i.e., three for RGB or four for RGB+NIR bands). The patches are fed into the cCNN in the form of 4D tensors.

The architecture of the cCNN is composed of a deepening series of convolution layers which change depending on the patch size. In effect, we use as many 3x3 filters as can be accommodated by the patch size without the recourse to padding. Therefore, for 3x3 image patches, we use a single 2D convolution layer since the convolution of a 3x3 image with a 3x3 kernel returns a single scalar value. An example of the cCNN architecture for a 3x3 pixel patch is shown in Fig. 7. For the 5x5 image patch, we use two 2D convolution layers. The first convolution of the 5x5 image with a 3x3 kernel leaves a 3x3 image which is rendered to a scalar after a second 3x3 convolution. For the 7x7 image patch size, we use three 2D convolution layers. Finally, for the 15x15 patch size we use seven 2D convolution layers. In all cases, each convolution layer uses 32 filters and therefore passes 32 equivalent channels to the following layer, with the exception of the final layer which passes a set of 32 scalar predictors. These scalars are flattened and fed into a dense top which terminates in the usual Softmax layer for class prediction (Fig. 7).

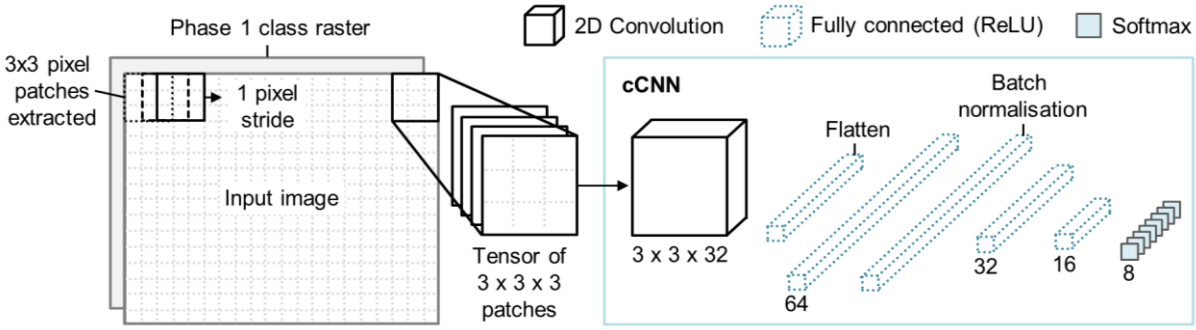


Figure 7: Architecture of the cCNN used in phase two for patch-based pixel level classification. Patches are extracted from the input image with a stride of one pixel, assigned a class label according to the class raster produced in phase one, and compiled into 4D tensors which are then fed into the cCNN. An example of a 3x3 RGB patch is shown in this diagram which uses an architecture with a single 2D convolutional layer with 32 3x3 filters. The convolutional layer feeds into a fully connected network for class prediction.

As with the MLP, when the cCNN was applied we used conventional early stopping with a patience parameter and a minimum improvement threshold. The minimum improvement was set as 0.5%. Using patch sizes of 3x3, we used a patience of 15 and

for patches of 7x7 and 15x15, a patience of 10. The number of trainable of parameters ranged from 5,880 in the case of RGB imagery with a patch size of 3x3 pixels to 231,582 for NIR+RGB imagery with a patch size of 15x15 pixels.

2.5 CNN-Supervised Classification performance

The performance of CSC was tested in two ways to allow comparison to previous deep learning methods. Firstly, classification accuracy was measured using manually collected validation labels. Secondly, a calving front detection method was implemented, and error was quantified using manually digitised calving front data for all test images.

Model performance is often measured by classification accuracy (the number of correct predictions divided by the total number of predictions). However, some models require more robust measures of accuracy which also take into account confusion between predicted classes (Goodfellow et al., 2016; Carbonneau et al., 2020a). We use an F1 score as the primary performance metric for the models used in both phases of the classification workflow. The F1 score is defined as the harmonic mean between precision (p) and recall (r):

$$F1 = \frac{2pr}{p + r} \quad (1)$$

where precision finds the proportion of positive predictions that are actually correct by dividing the number of true positives by the sum of both true (correct) positives and false (incorrect) positives. Recall finds the proportion of positive predictions that were identified correctly by dividing the number of true positives by the sum of true positives and false negatives (misidentified positives). Thus, the inclusion of recall provides a metric which represents confusion between class predictions and takes into account class imbalance (Carbonneau et al., 2020a). F1 scores range from 0 to 1 with 1 being equivalent to 100% accuracy. Carbonneau et al. (2020a) used classification results from 862 images to compare F1 and accuracy. They found that they are closely correlated (accuracy = 1.03F1 +4.1% with an R^2 of 0.96), with F1 and accuracy converging at 100%.

The validation labels used to calculate F1s were digitised manually using QGIS 3.4 digitising tools. Due to the manual nature of the data collection, this resulted in some unlabelled areas where classes were particularly difficult to define. This often occurred at class boundaries or where very small areas of different classes were mixed (at the scale of a few pixels). For example, in areas where the snow on rock class transitioned to bare bedrock, the structure of the underlying rock would often result in snow-covered areas spanning just a few pixels. In cases like this, digitising small patches of snow at pixel-level would become very time-consuming and as a result some areas of the image remained unlabelled. Despite this, we aimed to cover as much of each test image with validation labels as possible.

The F1 scores were calculated based on the concatenation of all the predictions for all available test images within the given parameters of tile size, patch size, number of bands, CSC phase, type of training (Single or Joint), glacier, and type of validation (in-sample or out-of-sample). Given that the calculation of F1 scores for gigapixel samples can be very computationally intensive, each F1 score presented here was estimated from a sample of 10 million pixels of the available data.

In addition to classification performance, we implemented a calving front detection method based on morphological geodesic active contours (see Fig. S1 in the Supplement). The method is based on the definition of a calving front as the contact between 'ocean' pixels (open water, iceberg water, or mélange) and glacier ice pixels. Since the final classification output from CSC is at pixel-level, this allowed for calving front detection at the native spatial resolution of Sentinel-2 imagery (10 m). Error was quantified for each predicted calving front by measuring the Euclidean distance between each predicted calving front pixel and the closest pixel in manually digitised calving fronts. From this, the mean, median, and mode error was quantified for each predicted calving front. Calculating the median and mode values allows the elimination of outliers in calving front predictions (Baumhoer et al., 2019). As with the classification validation labels, calving fronts were digitised in QGIS 3.4 and rasterised to form a single pixel-wide line.

3 Results

3.1 Classification performance

3.1.1 Phase 1: sensitivity to tile size and image bands

F1 scores for the tiled phase one classifications of unseen test imagery using a CNN with Single training are shown in Fig. 8. F1 scores range between 83.2% for 50x50 RGB tiles from out-of-sample data to 92.2% for 50x50 RGB+NIR tiles for in-sample data (Fig. 8). In general, these F1 scores suggest that the phase one CNN can classify unseen images to a sufficient level of accuracy to produce training data for phase two pixel-level classification. As shown in Fig. 8, the performance of models trained with RGB bands tend to improve with larger tile sizes, suggesting that the greater proportion of information stored in larger tiles is beneficial when using only three bands. In contrast, with RGB+NIR bands, classification performance declines with larger tile sizes, suggesting that spectral information is more important than spatial information. Phase one CNN predictions are more accurate for in-sample test data than out-of-sample test data, which is to be expected since the CNN was trained with data from Helheim. Overall, F1 scores from all test data combined suggest that the optimum tile size and band combination in phase one is 50x50 pixel tiles with RGB+NIR bands.

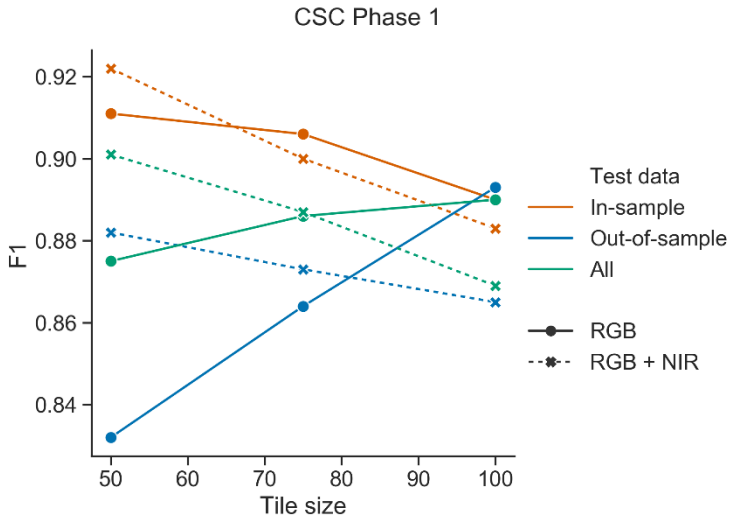


Figure 8: Performance of the phase one CNN with Single training showing the impact of tile size and image band combinations. For all test data combined we see that RGB+NIR tiles of 50x50 pixels produce optimum F1 scores. Note also that the RGB band combination tends to reach better performance with larger 100x100 pixel tiles whereas for RGB+NIR images, performance declines for larger tile sizes.

3.1.2 Phase 2: pixel- vs patch-based methods

For phase two pixel-level classification, the performance of both pixel- and patch-based techniques were tested using an MLP and cCNN, respectively. The F1 scores for pixel-level phase two classifications using the optimum band combination (RGB+NIR) and Single training are shown in Fig. 9, combining both in-sample and out-of-sample test data. The pixel-based technique is clearly outperformed by the patch-based technique, with F1 scores for the pixel-based method ranging from 73.2% (with 100x100 RGB tiles) to 81.4% (with 75x75 RGB+NIR tiles) for all test data combined. Comparatively, F1 scores from all test data combined using the patch-based method reach up to 89.7% for 3x3 patches, 91.6% for 5x5 patches, 91.8% for 7x7 patched and 92.2% for 15x15 patches (all with 50x50 RGB+NIR tiles).

The F1 scores for patch-based results from combined in-sample and out-of-sample test data using Single training are shown in Fig. 10. In general, larger patches of 5x5, 7x7 and 15x15 pixels outperform the 3x3 patch size. However, the range in F1s between different patch sizes is small (i.e., a 3.3% range), with the worst performing patch producing an F1 of 88.9% (3x3 patch using 100x100 RGB+NIR tiles) to the best performing patch size producing an F1 of 92.2% (15x15 patch with 50x50 RGB+NIR tiles). In addition, the trend shown in Fig. 8 which exemplifies a trade-off between spatial and spectral information in phase one is carried through to phase two. As in phase one, the phase two results show that smaller tile sizes of 50x50 pixels perform best with RGB+NIR bands rather than RGB alone. Using the RGB band combination requires larger tile sizes to produce better results. Given this, the optimum input parameters for the CSC workflow with Single training for GrIS marine-terminating glaciers are 50x50 RGB+NIR tiles with larger patch sizes from 5x5 to 15x15 pixels.

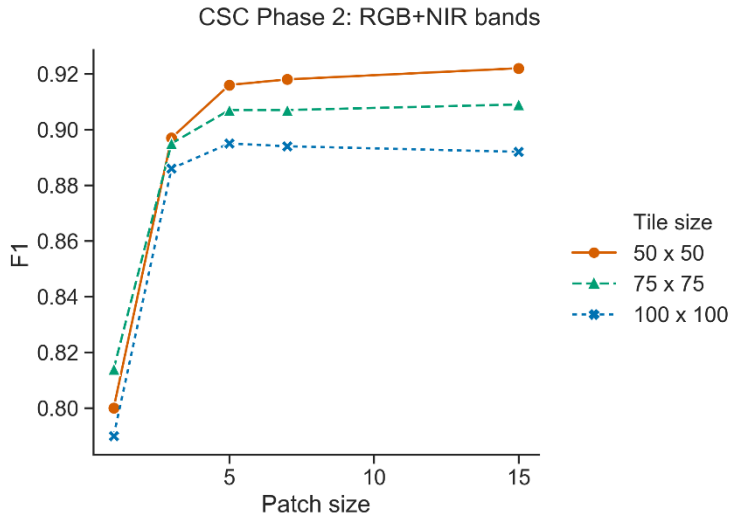


Figure 9: F1 scores for final phase two classifications of all test data combined, produced using the RGB+NIR band combination. Shows that the patch-based method significantly outperforms the pixel-based method for phase 2 classifications.

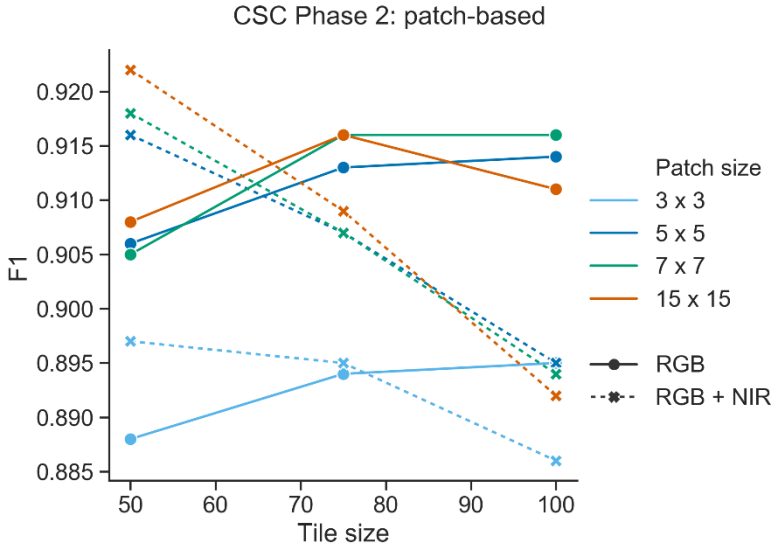


Figure 10: F1s of final patch-based phase 2 classifications for all test data combined (both in-sample and out-of-sample). Note the similar trend to phase one whereby RGB bands perform better with larger tile sizes, but with RGB+NIR bands, smaller tiles improve performance. Additionally, larger patch sizes deliver optimal performance, with 15x15 pixel patches producing best overall F1.

Confusion matrices for output classifications of each test glacier using the optimal patch size of 15x15 pixels with Single training can be found in the Supplement (Fig. S2). In summary, Fig. S1 shows good agreement between predicted and actual classes for all glaciers, with the exception of the open water class for Helheim and Jakobshavn where confusion occurs between the bergy water and bedrock classes. Open water is the smallest class for both sites, with open water often covering only small areas in each individual image. Moreover, at Helheim and Jakobshavn, land-based lakes (which are labelled as open water in

the manually digitised validation labels) often cover such small areas that they are missed during the application of the phase one CNNs. This is because most lakes occur at smaller scales than individual image tiles.

3.1.3 Seasonal performance

The test data for both in-sample and out-of-sample sites was seasonally variable, with images acquired between February and October 2019/2020 which contained different seasonal elements and varied illumination conditions. The generally good F1 scores for both in-sample and out-of-sample sites using 50x50 RGB+NIR tiles, and the patch-based technique, suggests that CSC has good seasonal transferability. For example, Fig. 11 shows three example test images acquired on different dates throughout 2019 for the Helheim test site, and alongside each image are the associated pixel-level CSC outputs (using Single training). Figure 11a shows an input image from 5 March 2019 where the landscape is covered in snow and the illumination angle has resulted in some areas of deep shadow. The corresponding classification in Fig. 11b has an F1 of 93.3% and calving front error of 10 m (equivalent to 1 pixel). The image acquired on 5 June 2019 shown in Fig. 11c is notably different in terms of seasonal characteristics and illumination. Snow covers fewer areas of the landscape and meltwater is also apparent on Helheim Glacier. The corresponding CSC classification has an F1 of 94.5% and calving front error of 36.06 m (equivalent to 3.6 pixels). The image acquired on 1 October 2019 shown in Fig. 11e also has different characteristics, with a larger ice sheet ablation area, little to no snow cover resulting in bare bedrock, and some shadows along the fjord edge similar to Fig. 11a. The corresponding classification in Fig. 11f has an F1 score of 95.5% and calving front error of 40 m (equivalent to 4 pixels) with one small area of glacier ice near a nunatak which has been misclassified as mélange, possibly due to very little textural difference (i.e., no crevassing).

Similarly, Fig. 12 shows examples of CSC with Single training applied to seasonally variable imagery from the out-of-sample study site which includes Store Glacier. Figure 12a shows an image from 28 June 2020 where mélange was present at the terminus and the glacier ice has surficial meltwater. The corresponding classification shown in Fig. 12 b has an F1 of 86.3% with several areas of glacier ice which have been misclassified as mélange. The calving front error is 14.14 m, equivalent to 1.4 pixels. Figure 12c shows an image from 14 September 2020, with little to no melange and less meltwater visible on the glacier surface. There are also areas of bedrock which are deeply shadowed, resulting in some misclassifications of bedrock areas as open water in the corresponding CSC output (Fig. 12d). The resulting classification has an F1 score of 89.9% and calving front error of 10 m (1 pixel). The image from 8 October 2020 shown in Fig. 12e has more snow cover, no mélange, and a lower angle of illumination with the resulting classification (Fig. 12f) producing an F1 of 91.1% and calving front error of 14.14 m (1.4 pixels). The CSC outputs for the out-of-sample site using Single training are more prone to misclassification compared to the in-sample site, which prompted testing of the Joint fine-tuning method. Additionally, it was noted that CSC did not produce very accurate classifications for images with extremely low illumination angles. This is most likely because images with very low illumination angles occurred most frequently at the beginning or ends of the image availability season

and made up a smaller proportion of phase one training data. To improve the ability of CSC to classify imagery with deep shadow and extremely low illumination angles, the proportion of imagery containing these qualities could be increased in training data for phase one CNNs.

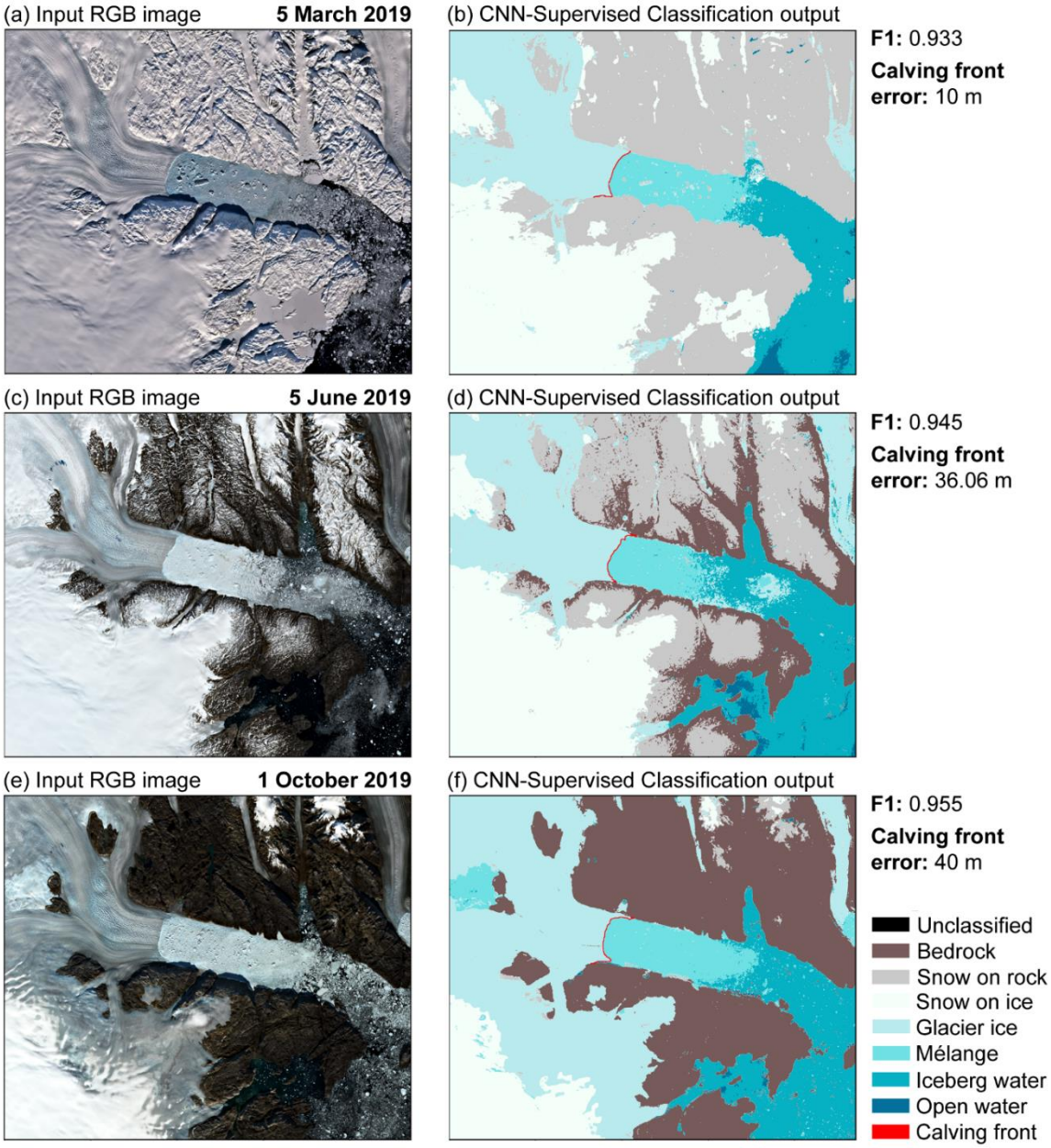
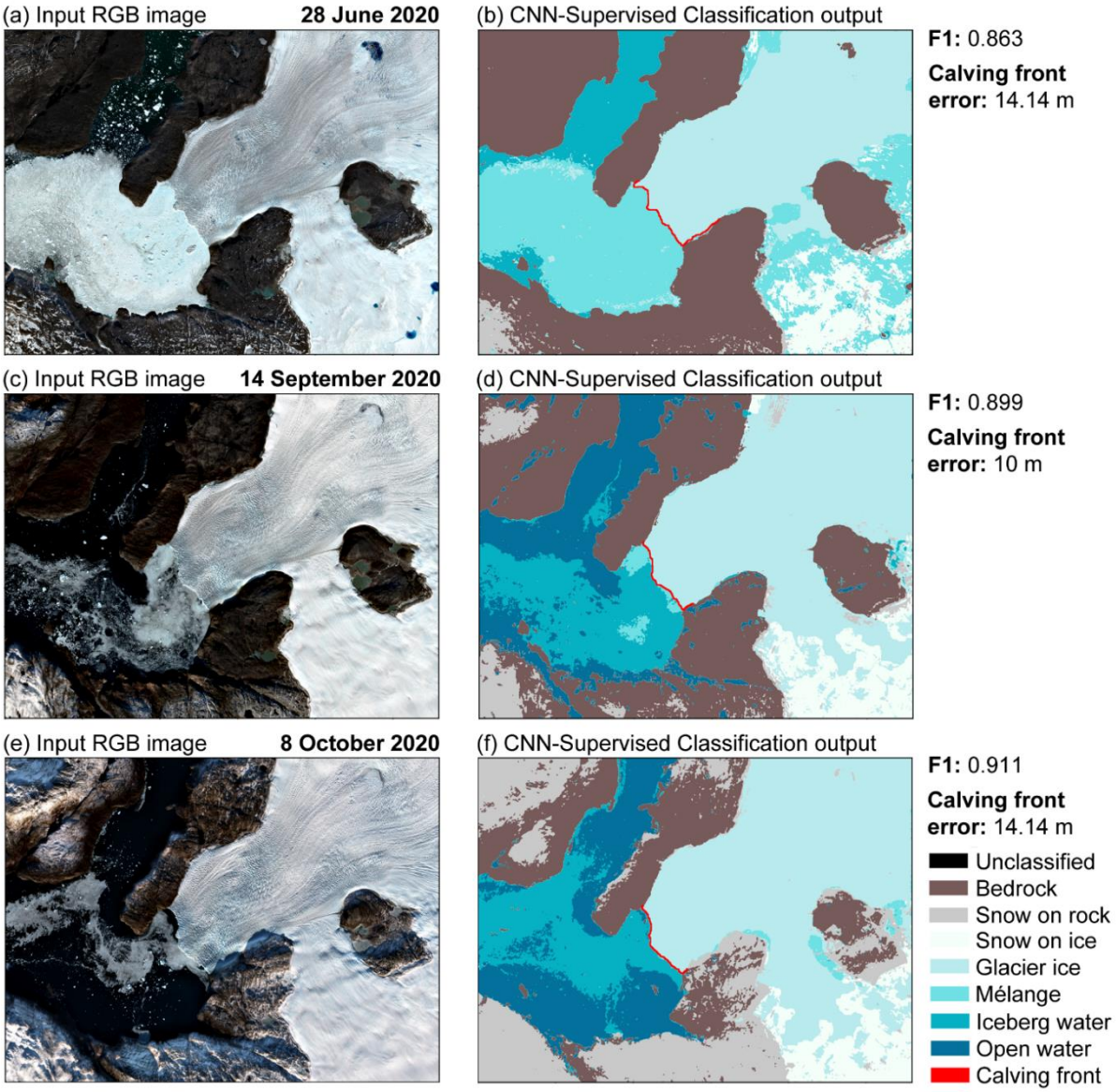


Figure 11: Examples of pixel-level classification outputs for seasonally variable imagery from the in-sample test site showing input RGB images of Helheim in the first column, which were acquired on (a) 5 March 2019, (c) 5 June 2019, and (e) 1 October 2019 with the associated CSC outputs shown in (b), (d), and (f). The F1 scores and calving front error are shown next to each classification. These classifications were produced using Single training with RGB+NIR 50x50 tiles and a phase two patch size of 7.



525 **Figure 12: Examples of pixel-level classification outputs for seasonally variable imagery from the out-of-sample test site showing input RGB images of Store in the first column, which were acquired on (a) 23 June 2020, (c) 14 September 2020, and (e) 8 October 2020 with the associated CSC outputs shown in (b), (d), and (f). The F1 scores and calving front error is shown next to each classification. These classifications were produced using Single training with RGB+NIR 50x50 tiles and a phase two patch size of 5.**

530 Overall, these examples show the ability of CSC to classify in- and out-of-sample imagery of marine-terminating glacial landscapes in Greenland with different seasonal characteristics. Using the optimum CSC input parameters produces classifications with good F1 scores and subsequent calving front predictions that vary by only a few pixels from manual delineations.

3.1.4 Performance of CSC on an entire Sentinel-2 tile

The size of input imagery to the CSC workflow is not limited to a specified set of dimensions. Since collection of validation labels for each test image required manual digitisation, the test sites were restricted to ~20 to 50 km to allow collection of seasonal data for individual glacial landscapes. Despite this, CSC can also be applied to entire Sentinel-2 tiles, so validation labels were also collected for a whole Sentinel-2 tile of the landscape surrounding Helheim Glacier. The outputs of the CSC workflow applied to the entire Sentinel-2 image are shown in Fig. 13. Figure 13a shows the input image collected on 13 September 2019. The tiled phase one predictions are shown in Fig. 13c and the final pixel-level classification is shown in Fig. 13d. The overall F1 score of this classification was 92%.

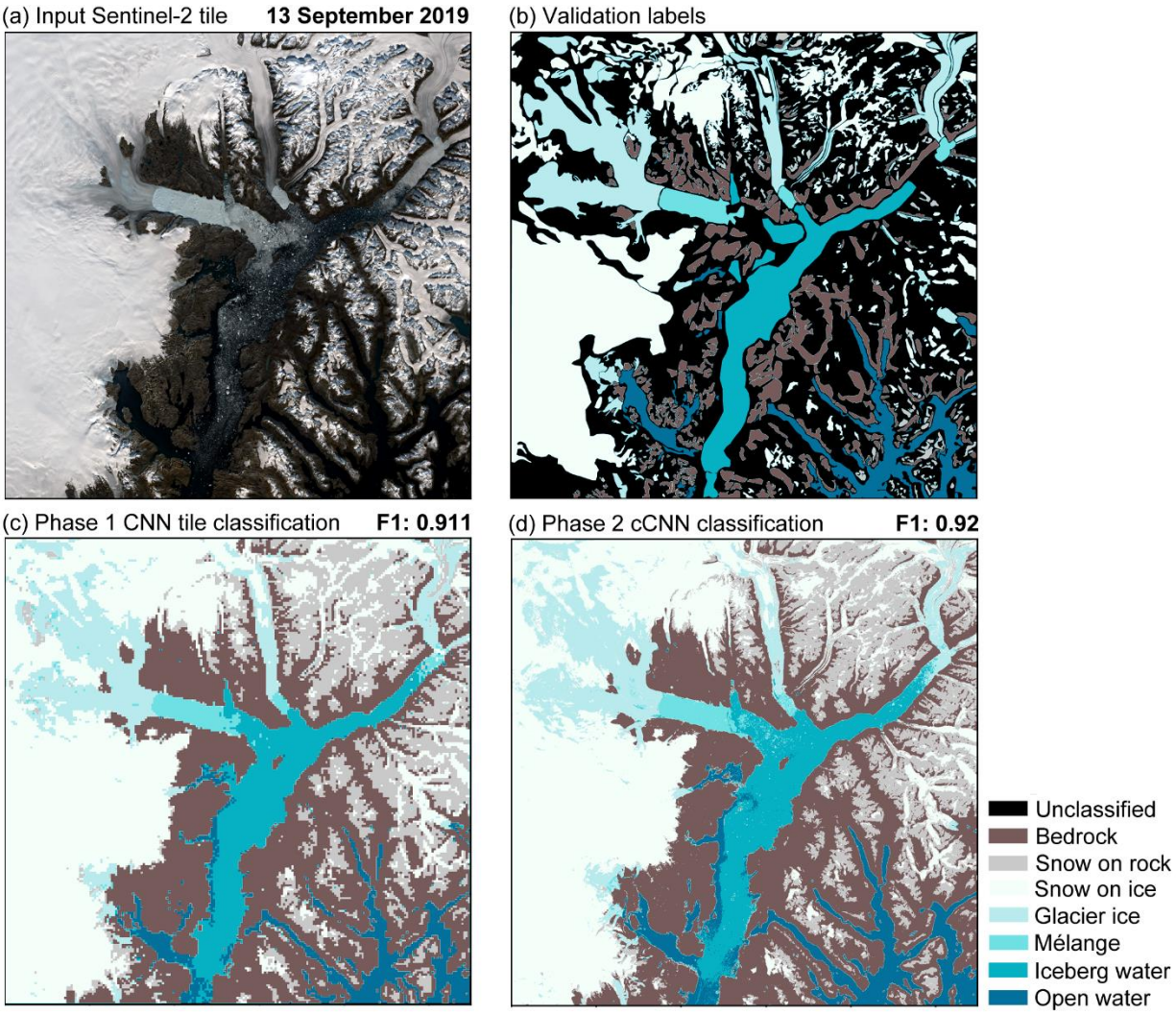


Figure 13: CSC performance on (a) an entire Sentinel-2 tile. (b) Shows validation labels. (c) Shows the tiled classification output of phase 1 which was used as training data for phase 2, producing a final pixel-level classification shown in (d). The final classification was produced using RGB+NIR tiles with a size of 50x50 pixels and a cCNN patch size of 7 using Single training.

3.1.5 Performance of CSC using Joint fine-tuning

The addition of a small number of extra image tiles (5,000 per class) from two glacier-specific images (one from winter and one from summer) used for the Joint fine-tuning of phase one models significantly improved classification accuracy for both out-of-sample study sites (Fig. 14b and c). Results were only marginally improved for the in-sample study site (Fig. 14a), which was to be expected since phase one models were already trained on data from Helheim. An example comparing classification outputs using Single and Joint training is shown in Fig. 15 which shows an image of the out-of-sample Store site acquired on 22 August 2020. The phase one CNN with Single training misclassified a large area of the glacier as mélange (Fig. 15c) and as a result the subsequent phase two classification also had large areas of misclassified glacier ice (Fig. 15d). The addition of Joint fine-tuning rectified this and led to fewer glacier ice tiles being misclassified as mélange in the phase one predictions (Fig. 15e). The robustness of the phase two model resulted in a final classification with correct classification for the majority of the glacier, with an overall F1 score of 97.5% as opposed to 84.7% with Single training.

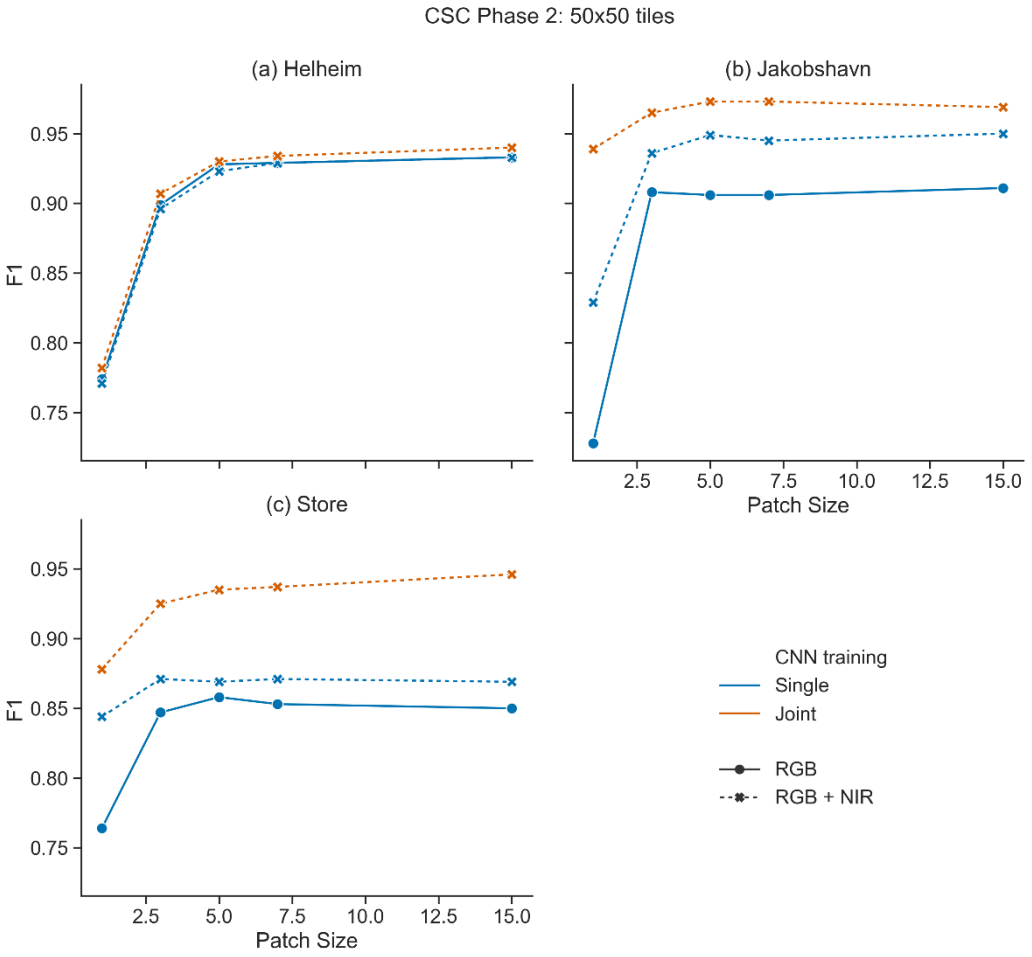


Figure 14: Comparison of CSC performance using Single vs Joint training approaches for (a) Helheim, (b) Jakobshavn, and (c) Store. F1s shown as a function of image bands and patch size using a tile size of 50x50 pixels.

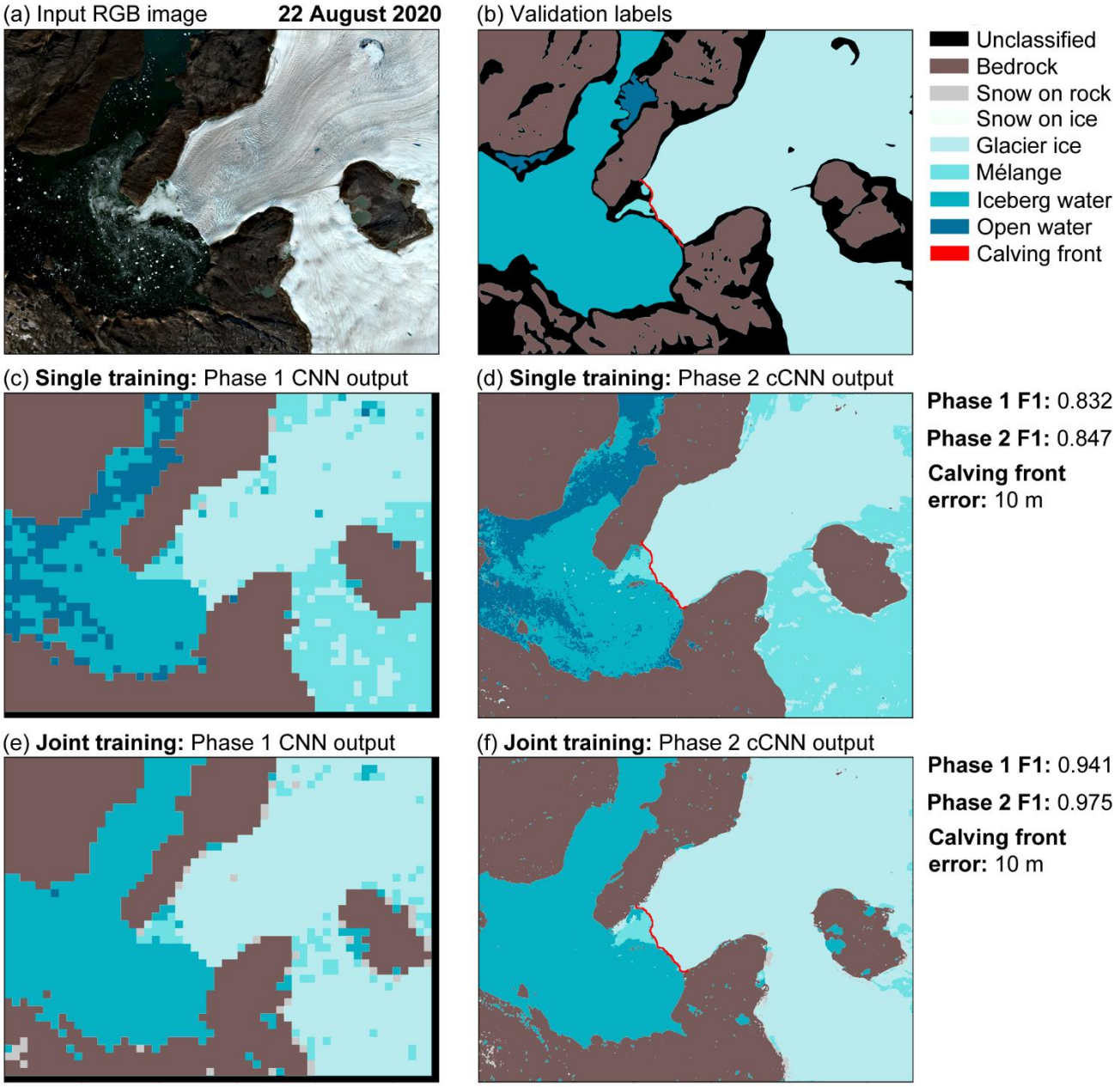


Figure 15: Comparison of Single and Joint training methods for (a) an image of Store glacier acquired on 22 August 2020. (b) shows the manually collected validation labels. (c) Shows the phase 1 tiled output using Single training and (d) shows the resulting CSC output. Note the area of glacier ice which has been misclassified using Single training. (e) Shows the phase 1 output using Joint training with the associated pixel-level phase 2 output shown in (f). A tile size of 50, patch size of 5 and RGB+NIR bands were used in the examples shown here. The Joint training method rectifies the misclassified area of glacier ice.

In terms of per-glacier improvements, the best overall F1 score for Helheim went from 93.3% with Single training to 94% with Joint training (improvement of +0.7%), both using 50x50 RGB+NIR tiles with 15x15 patches. An example classification using Joint training is shown in Fig. 16 with a resulting classification F1 score of 94.9% and calving front error of 50m.

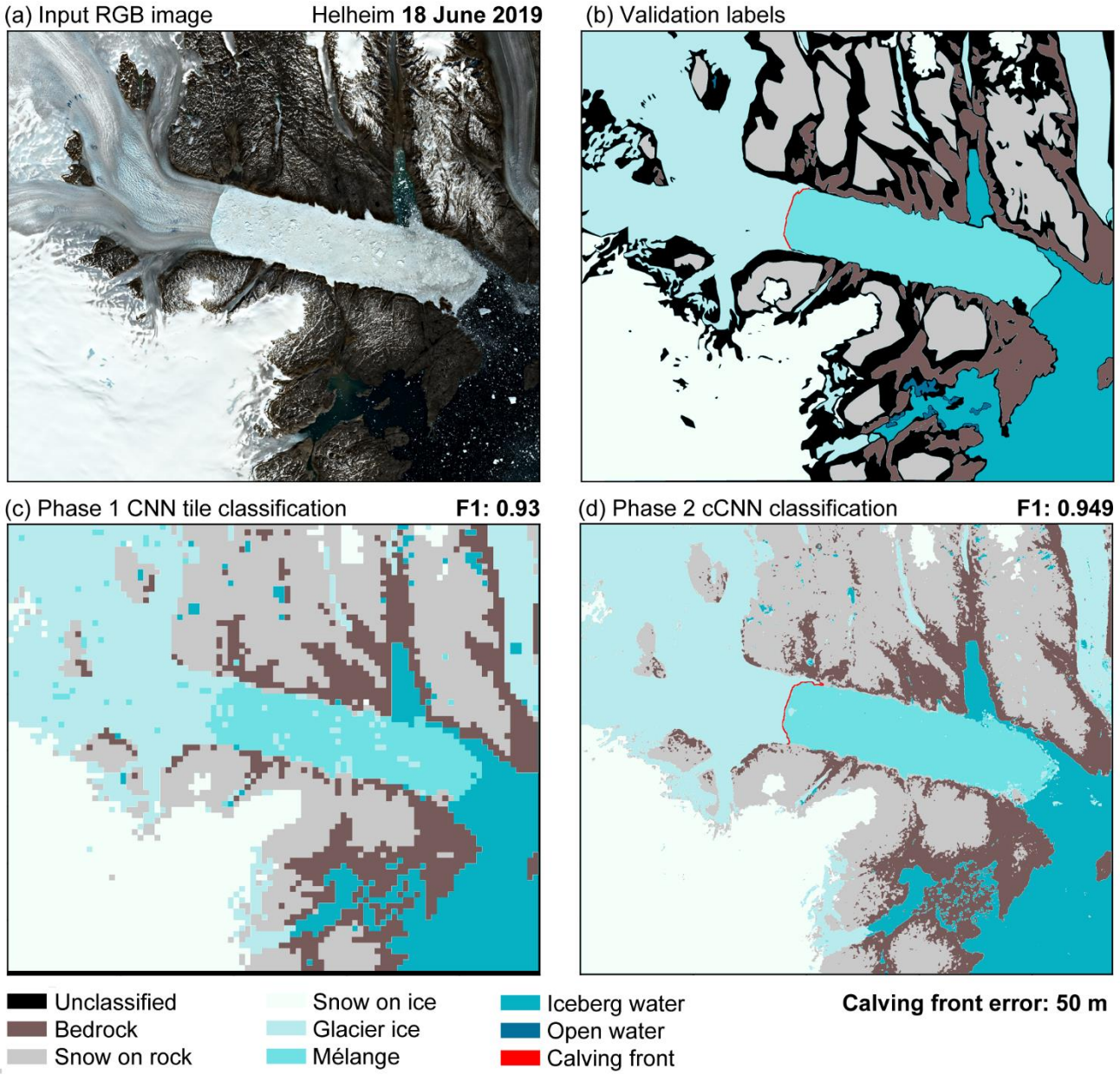


Figure 16: Example of CSC using Joint training for (a) an unseen image of Helheim acquired on 18 June 2019. (b) Shows the manually collected validation labels. (c) Shows the tiled output of the phase 1 CNN and (d) shows the final pixel-level classification with an associated calving front detection. The optimum classification parameters with a tile size of 50, patch size of 15 and RGB + NIR bands were used in this example.

For Jakobshavn, the best F1 score went from 95% with Single training (50x50 RGB+NIR tiles and 15x15 patches) to 97.3% with Joint training (50x50 RGB+NIR tiles and 5x5 patches). An example of one of the Jakobshavn test images and the corresponding output classification using Joint training can be seen in Fig. 17. The final classification (Fig. 17d) had an F1 score of 98.1% and an average calving front error of 50 m.

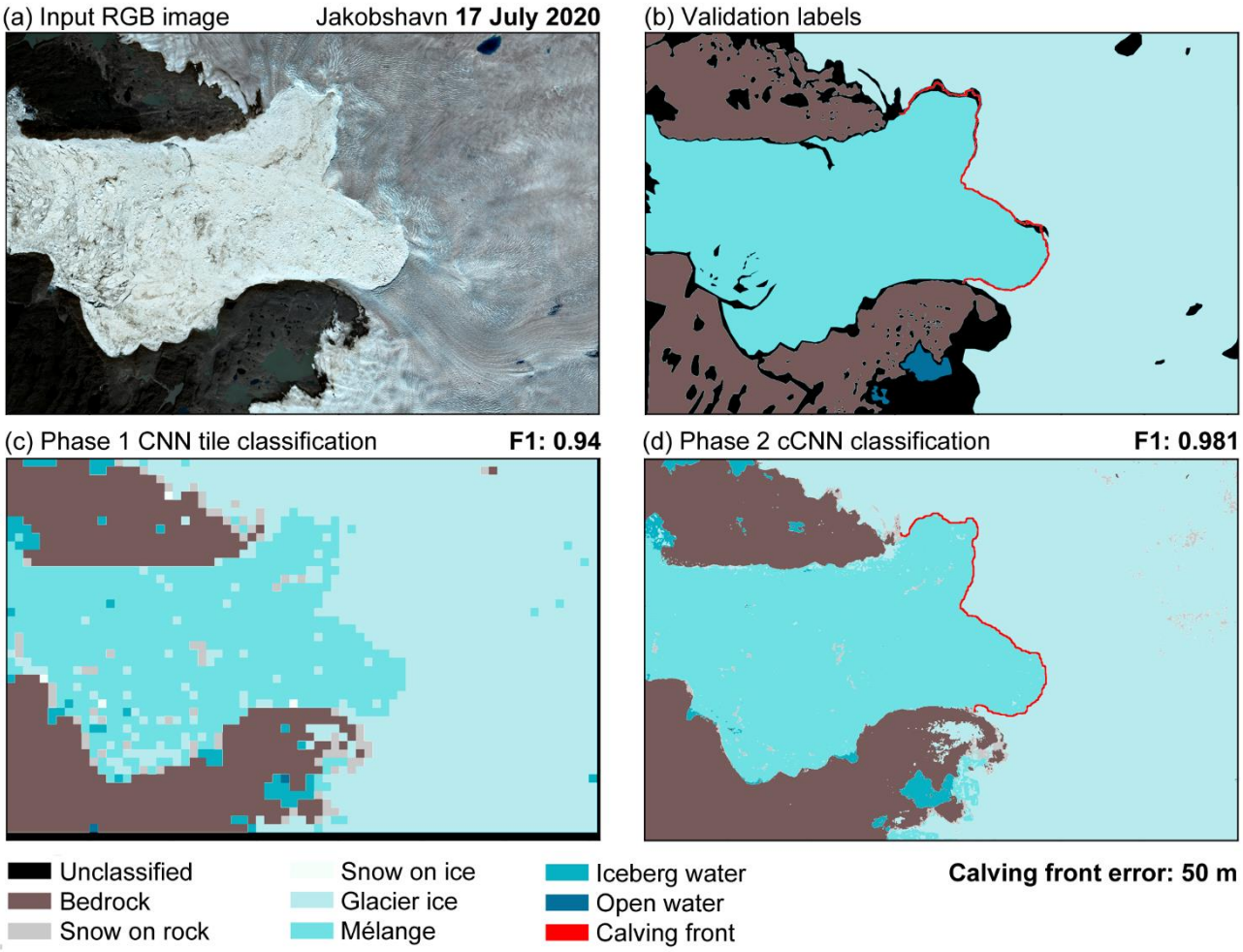


Figure 17: Example of CSC using Joint training for (a) an unseen image of Jakobshavn acquired on 21 May 2020. (b) Shows the manually collected validation labels. (c) Shows the tiled output of the phase 1 CNN and (d) shows the final pixel-level classification with an associated calving front detection. The optimum classification parameters with a tile size of 50, patch size of 5 and RGB + NIR bands were used in this example.

For the Store study site, the best overall F1 increased from 91.4% with Single training (using 100x100 RGB+NIR tiles and 3x3 patches) to 94.6% with Joint training (using 50x50 RGB+NIR tiles and a patch size of 15x15 pixels). An example of a CSC output with Joint training is shown in Fig. 15. In summary, this suggests that digitising an additional 2 images for the purposes of fine-tuning an existing pre-trained CNN for glacier-specific classification is worth the improvements in classification accuracy.

3.2 Calving front error estimation

Since the predictions of CSC phase two are pixel-level, this allows the implementation of a calving front detection algorithm which uses morphologic active contours and other binary morphology operators to establish a calving front. The error of calving front predictions was calculated based on the distance from manually delineated fronts. Overall, the optimal CSC input parameters which produced the lowest mean error were 50x50 tiles, RGB+NIR bands, and 5x5 patches with Single training. Using these parameters resulted in a mean calving front error of 56.17 m (equivalent to 5.6 pixels) for the test dataset as a whole (with individual mean errors of 58.81 m for Helheim, 70.6 m for Jakobshavn, and 39.1 m for Store). For the same parameter set, median error was 24.7 m (30 m for Helheim and Jakobshavn, and 14.1 m for Store), suggesting that mean values are increased by extreme values. Figure 18 shows the full error distribution for every predicted calving front pixel detected in classifications produced with RGB+NIR bands and 50x50 tiles. The data shown is for all glaciers combined.

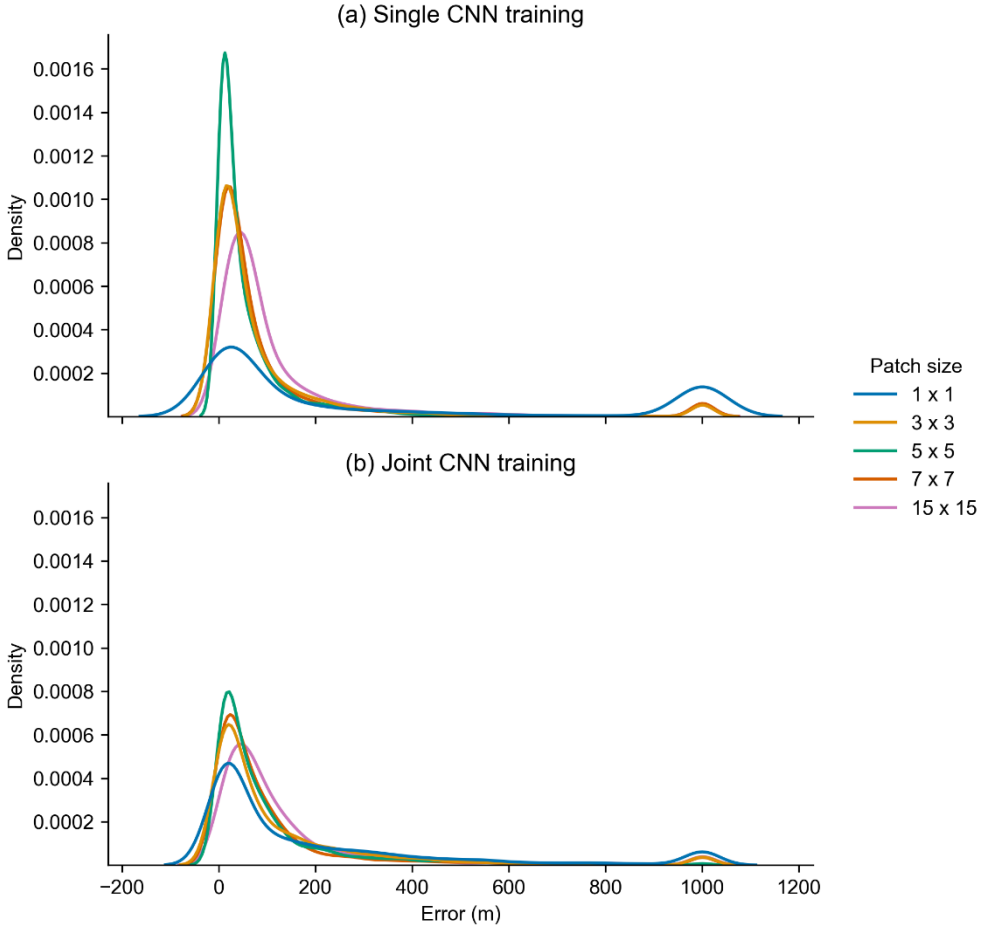


Figure 18: A kernel density estimate (KDE) plot of the full error distribution for all calving front predictions derived from all test sites using classifications produced with RGB+NIR bands and 50x50 pixel tiles using (a) Single CNN training or (b) Joint CNN training. Error values above 1000 m are grouped into a single bin to prevent long tails in the plots and show a second peak which represents catastrophic errors in calving front prediction. Note that low calving front errors occur most often with 5x5 patches, followed by 7x7 and 3x3 patches, with highest error occurring for 15x15 patches and 1x1 patches (pixel-based).

Figure 18 shows that minimal error is achieved using 5x5 patches, followed by 7x7, 3x3, 15x15 patches with pixel-based results producing the worst calving front errors. Firstly, we note that small classification errors of a few pixels (often caused by shadows at the front) can lead to errors in the range of 5 to 10 pixels. On this front, the smaller scale information provided in a 5x5 pixel patch is clearly optimal in comparison to overall classification accuracy which achieves good results with patch sizes from 5x5 to 15x15 pixels. Secondly, we note a small tail of data where large errors can occur. In Fig. 18, a secondary peak which represents calving front errors of 1000 m and above shows where calving front predictions were catastrophically erroneous. In all of our test data, one of the 27 test images severely failed to detect the calving front (despite a high F1). Nonetheless, a mean calving front error of 56.17 m derived from classifications using Single CNN training with RGB+NIR bands, 50x50 tiles and 5x5 patches suggests that CSC has the ability to detect calving fronts with reasonable accuracy.

Furthermore, Fig. 19 shows calving front errors as a function of tile size and patch size for each glacier using RGB+NIR bands with Single CNN training while Fig. 20 shows the calving front errors as a function of patch size for Joint training. The modal error for both Single and Joint training (Fig. 19a and 20a) ranges from 10 to 50 m. Median errors fall below 100 m for the majority of calving front detections, especially those produced using optimal parameters (Fig. 19 b). Crucially, these figures do not show a systematic increase of error with the patch size which suggests that calving front errors can be attributed to classification errors at glacier fronts.

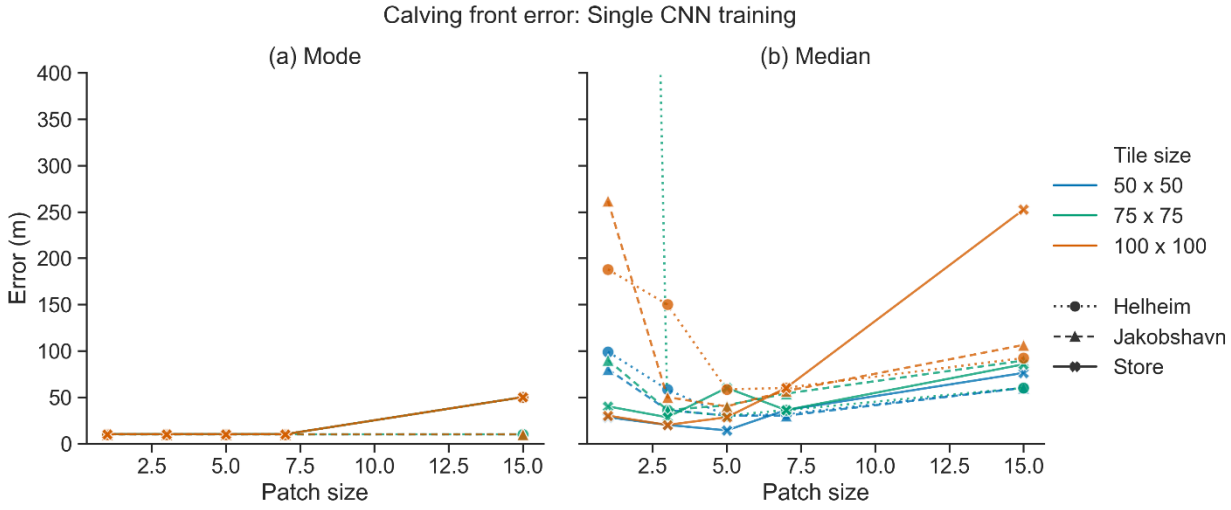


Figure 19: (a) modal and (b) median calving front errors as a function of patch and tile size for each glacier using Single training.

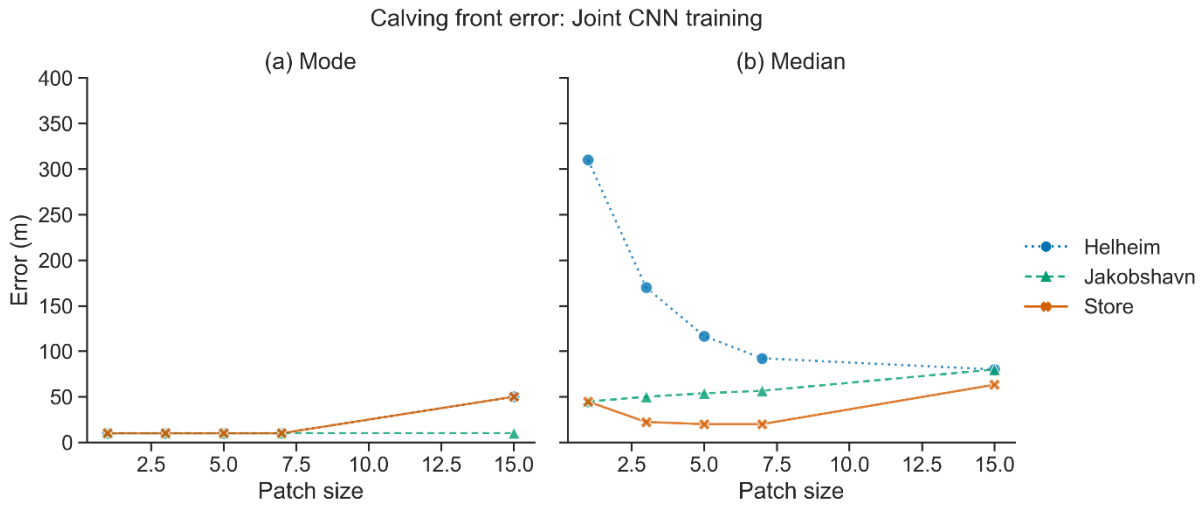


Figure 20: (a) modal and (b) median calving front error as a function of patch size for each glacier using Joint training.

4 Discussion

4.1 Performance of CNN-Supervised Classification

The results reported here demonstrate that the CSC workflow adapted for landscapes containing marine-terminating outlet glaciers in Greenland produces state-of-the-art pixel-level classifications for seasonally variable imagery. By testing the performance of different band combinations, tile sizes, and patch sizes on seasonally variable test imagery, we find that classifications reach F1 scores of up to 93.3% for in-sample test imagery, and 91% for out-of-sample test imagery when using a phase one CNN trained only with data from Helheim Glacier and the overall optimal input parameters (of 50x50 RGB+NIR tiles with 15x15 patches). With the addition of Joint fine-tuning, F1 scores increased to 94% for in-sample test data and 96% for out-of-sample test data. In terms of calving front accuracy, a mean error of 56.17 m (5.6 pixels) and median error of 24.7 m (2.5 pixels) was achieved from classifications produced with Single training, 50x50 tiles, RGB+NIR bands and a patch size of 5x5 pixels. In comparison, manually digitised calving fronts usually have error of up to 2-3 pixels. For example, Carr et al. (2017) calculated a mean frontal position error of 27.1 m using repeat digitisations. Overall, this suggests that the accurate multi-class outputs of CSC are capable of producing datasets with sufficient levels of accuracy, for example to monitor calving front change. Given that CSC can identify seven different semantic classes, this also provides scope for analysis in other research areas, beyond calving front monitoring. For example, changes in other class boundaries could be monitored, for instance to detect changes in snowline position and quantify ablation area change (Noël et al., 2019). Similarly, the multi-class outputs could be used to quantify seasonal changes in the area of a specific class, for example to monitor changes in the area and extent of mélange (Foga et al., 2014; Cassotto et al., 2015). Moreover, while CSC operates at the scale of overall landcover classes, outputs could potentially be used to isolate a specific target class for detection of smaller scale features such as

supraglacial lakes (Hochreuther et al., 2021) and subglacial plumes (How et al., 2017; Everett et al., 2018). In addition, the outputs of the CSC script retain the geospatial information of the input data, meaning classification and calving front outputs can be manipulated in GIS software, for example to produce time series data of seasonal change with ease.

4.2 Sensitivity of CSC performance to tile size, image bands, and patch size

From both phase one and phase two results showing the sensitivity of classification performance to image bands and tile size, we see a trade-off between the use of spatial and spectral data. In summary, the increased spectral data provided by the RGB+NIR band combination is optimal with the smaller 50x50 pixel tiles, but performance declines with larger tile sizes (Fig. 10). In contrast, the performance of CSC using RGB bands is low for 50x50 tiles but increases with tile sizes of 75x75 and 100x100, suggesting that the increased spatial data in larger tiles is beneficial when only using three image bands. Overall, 50x50 tiles and RGB+NIR bands produce optimal results for the whole test dataset, suggesting that the combination of spectral and spatial information contained within a four-band 50x50 tile is ideal for detecting class-specific features at the scale of 10 m resolution imagery for marine-terminating glacial landscapes in Greenland. Testing the use of additional image bands to increase spectral data may be advantageous in future work. For example, Xie et al. (2020) used a CNN trained with 17 input bands derived from Landsat 8 imagery and DEM data and found that using more bands produced higher accuracy than using fewer bands for mapping debris-covered mountain glaciers. However, this may not necessarily be the case with marine-terminating outlet glaciers and using additional input channels is likely to increase processing time which should also be taken into account when considering that accurate results can be achieved using only RGB+NIR bands.

In phase 2 pixel-level classification, the patch-based method significantly outperformed the pixel-based method. The reason for testing pixel- and patch-based techniques was due to our use of medium resolution satellite imagery which tends to have spectral variations across images, making it difficult to distinguish class from the spectral characteristics of a pixel alone (Maggiori et al., 2016). We proposed that adopting a patch-based technique which includes contextual information surrounding a pixel would aid classification of complex and seasonally variable outlet glacier landscapes, as it has in other applications (Sharma et al., 2017) and found that this was true. This also validates similar findings that patch-based CNNs outperform standard pixel-based neural networks and CNNs (Sharma et al., 2017). Moreover, while a patch size of 15x15 pixels produced classifications with the highest F1s for all data combined, there was only a 3.3% range in F1s for all patch sizes (for classifications produced with 50x50 RGB+NIR tiles and Single training), suggesting that the patch-based method in general produces good results. This demonstrates that aside from the benefit of using patches instead of individual pixels for pixel-level classification, classification performance overall is not particularly sensitive to which patch size is used. However, for calving front detection, a patch size of 5x5 pixels was optimal, suggesting that the smaller scale contextual information contained within a 5x5 pixel patch in comparison to a 15x15 pixel patch is beneficial for classification at the glacier front where small areas of shadow can impact front prediction at the scale of a few pixels.

4.3 Impact of Joint fine-tuning for out-of-sample sites

Since phase one CNNs were initially only trained on data from one site, when the CSC workflow was applied to out-of-sample test images with Single training, outputs were more prone to misclassification in comparison to in-sample test imagery. As a result, Joint fine-tuning was tested, and we found that the method significantly improved classifications with the addition of training data from only two glacier-specific images. Considering the improvements to classification performance for out-of-sample sites made by fine-tuning the phase one CNNs, we suggest that the manual labour required to collect 5,000 additional samples per class derived from only two images is not substantial and may be worthwhile if a glacier is identified for monitoring. Despite this, the application of CSC using Single CNN training still produced an F1 score of up to 91% for out-of-sample test data, providing sufficient classification quality to detect calving fronts with a mean error of 54.86 m (5.5 pixels).

4.4. Comparison to previous work

Our results build on the work of deep learning-based classification methods for ice front delineation (Baumhoer et al., 2019; Mohajerani et al., 2019; Zhang et al., 2019; Cheng et al., 2021), with several key innovations and variations of note. Firstly, it is important to note that the CSC workflow produces multi-class outputs using seven semantic classes rather than binary outputs of previous methods. This fulfils the aim to provide meaningful information which could be used for a variety of applications at the scale of entire outlet glacier landscapes in Greenland. In terms of classification accuracy, CSC produces marginally better F1s in comparison to previous methods applied to marine-terminating glacial environments. Previous studies which focus on outlet glaciers of the GrIS do not provide F1 scores for their classification outputs. However, Baumhoer et al. (2019) apply their method to Antarctic marine-terminating environments and produce overall F1s of 89.5% for training areas (in-sample) and 90.5% for test areas (out-of-sample). In comparison, CSC produces F1 scores of up to 93.3% for in-sample test imagery, and 91% for out-of-sample test imagery when using a phase one CNN trained only with data from Helheim Glacier. By applying Joint CNN training to fine-tune the phase one CNN to each test glacier, F1 scores increased to 94% for in-sample test data and 96% for out-of-sample test data. It is worth noting that the characteristics of Antarctic outlet glacier environments can vary substantially from Greenlandic outlet glacier environments, potentially presenting different classification challenges. As such, this is a tentative comparison, especially given that CSC outputs contain seven classes at the scale of the whole landscape, rather than just two classes focused at the ice front.

Furthermore, since previous deep learning studies which produce binary classifications for Greenlandic outlet glaciers do not provide F1 scores, for further comparison we also integrated a calving front detection method into the CSC workflow which produced a mean error of 56.17 m (5.6 pixels) for all test images using Single training and optimal input parameters (RGB+NIR bands, 50x50 tiles and 5x5 patches). Table 2 shows the mean calving front errors produced in each of the previous deep learning studies designed to detect calving fronts using binary classifications. Mean calving front errors for test imagery from

both training sites (in-sample) and test sites (out-of-sample) are provided, however not all studies specified these values. In terms of the number of metres that predicted fronts deviate from manual digitisations, the predictions of CSC are comparable to those of previous studies. However, in terms of the equivalent number of pixels, CSC predictions deviate from manual digitisations by a few more pixels compared to previous studies (apart from Zhang et al., 2019), indicating that if a given application solely requires accurate calving front localisation of a known glacier, the method presented here is not necessarily the optimal choice.

Table 2: Mean calving front errors from previous deep learning methods designed specifically to detect ice fronts in comparison to the mean calving front errors produced by CSC in this study.

Study	Ice sheet	No. of test images	Mean calving front error (and equivalent in pixels)		
			Training site(s)	**Test site(s)** (sites not used in training)	**Both training and test sites combined**
Baumhoer et al. (2019)	Antarctic	11	78.25 m (< 2 pix.)	107.75 m (2.69 pix.)	93 m (2.33 pix.)
Mohajerani et al. (2019)	Greenland	10	-	96.31 m (1.97 pix.)	-
Zhang et al. (2019)	Greenland	84	38 m (6 pix.)	-	-
Cheng et al. (2021)	Greenland	162	-	-	86.76 m (2.25 pix.)
This study	Greenland	27	58.81 m (5.9 pix.)	54.86 m (5.5 pix.)	56.17 m (5.6 pix.)

The second major difference between CSC and previous methods is the deep learning architecture. All previous deep learning classification methods for delineating ice fronts in marine-terminating glacial environments (Baumhoer et al., 2019; Mohajerani et al., 2019; Zhang et al., 2019; Cheng et al., 2021) use FCN/U-Net architectures (Ronneberger et al., 2015). Hoeser et al. (2020) reviewed image segmentation and object detection in remote sensing and whilst they do conclude that FCN/U-Net architectures are dominant, they still find about 30% of published work uses patch-based approaches which are akin to the second phase of the CSC method presented here. This suggests that FCN architectures need not be considered the *de facto* algorithm for glacial landscape classification. Moreover, the advantage of CSC over one-stage patch-based methods using FCNs is that the initial phase one CNN in CSC provides transferability and delivers a bespoke training set for the pixel-level patch-based operator (as described in Section 2.1). We discuss the other major implications of the architectural differences between our work and FCNs in the following sections.

4.4.1 Data pre-processing and computational loads

CSC has certain practical advantages over FCNs in terms of data processing and computational loads. Firstly, the CSC method has low pre-processing requirements. In effect, test images were cropped to areas of ~2000-5000 by 2000-5000 pixels in order to produce large images containing whole marine-terminating glacier landscapes, yet still within a workable size for detailed

digitisation of validation labels. Then, for CSC the only pre-processing step required is normalisation by a constant factor of 8192 to convert raw Sentinel-2 data to 16-bit floating point data. Once this is done, CSC has a low computational load. Training the initial VGG16 model can be done in under one hour using an I7 processor at 5.1Ghz, and an Nvidia RTX 2060 GPU. When CSC is subsequently applied to a sample image of 3000x3000 pixels using optimal parameters of RGB+NIR bands and tiles of 50x50 pixels for the phase one CNN, and patch size of 7x7 pixels for the phase two cCNN, classification requires 4 minutes. We also coded a low-memory usage pathway in the main script that classifies a large image row-by-row with a threshold to define 'large' set by the user. Using this, we can classify a stack consisting of full bands 4, 3, 2, and 8 (RGB+NIR) for Sentinel-2 at native resolution (10980x10980 pixels each) in 12 minutes with a peak RAM consumption of 11GB. This makes CSC suitable for use in free cloud-based solutions such as Google Colaboratory, providing the potential to build on existing cloud-based tools for glacial mapping (e.g. Lea, 2018). Moreover, given the simplicity of data pre-processing steps required for CSC, the workflow has good accessibility and can be implemented easily by new users.

In contrast, for several of the previous studies which implement FCN architectures, a larger number of pre-processing steps are required, including but not limited to rotation for consistent glacier flow direction, edge enhancement, and pseudo-HDR toning (Mohajerani et al., 2019; Zhang et al., 2019; Cheng et al., 2021). Similarly, FCN architectures can be very demanding in terms of computer RAM and GPU RAM, especially when large images are used as inputs. When we tested this by implementing the popular FCN8 based on VGG16 which has ca. 130 million trainable parameters, we found that the largest dyadic image size that could be processed was 512x512. This general problem has been resolved in different ways in the Earth observation (EO)-facing literature. Baumhoer et al. 2019 used 40 m Sentinel-1 SAR data and a DEM at 90 m resolution as their base. Using a smaller FCN with ca 7.8 million parameters, they used image tiles of 780x780 with 4 channels (HH, HV, DEM, HH/HV polarisations) on a GTX 1080 GPU (8Gb vs 6Gb for the RTX2060). However, it is important to note that with 40 m data, 780 pixels still covers 31.2 km. If this were Sentinel-2 optical data, with a resolution of 10 m, the sample tiles would only cover 7.8 km. In contrast, the calving front of Jakobshavn in test imagery used in this work has a width of ~11 km. In order to get around this sort of issue using FCNs, downsampling is used. For example, Mohajerani et al. 2019 used an advanced pre-processing routine that involved a re-orientation and then a resampling of the scene to 200x300 pixels. This resampling resulted in imagery with varied resolutions across glaciers used in training and test data. In the end, the FCN they used only had 240x152 pixels in a single post-processed channel which was tested at a single site (Helheim Glacier) with a resampled spatial resolution of 49 m (from Landsat data with 15/30 m resolution). In contrast, the spatial resolution of the input images and resulting classification outputs using CSC always remains native to raw Sentinel-2 data (i.e., 10 m).

4.4.2 Training data volume

In terms of the number of training samples used for deep learning models, Goodfellow et al. (2016) note that, as a general rule, each class should contain at least 5,000 samples to reach satisfactory performance, but models can reach and exceed human-

level performance when trained on at least 10 million samples. With this in mind, the number of labelled samples produced by manually labelled training images and data augmentation in the datasets used here (210,000 tiles) makes them relatively small. However, in comparison to pre-trained models such as VGG16 which were trained on the ImageNet database using over 1000 classes, our adapted VGG16 architecture only uses seven classes, and therefore can be trained sufficiently with 'only' a few 100 thousand samples. This suggests that relatively few images are needed to produce highly accurate image classifications using our workflow, reducing the time required for initial creation of manually labelled training data. Furthermore, the number of satellite acquisitions used to produce the training data for the phase one CNN in CSC is smaller than that used to train models in previous FCN-based studies. Given that our basic phase one CNN training sample is no larger than 100x100 pixels, a very large number of samples can be extracted from a full Sentinel-2 tile of 10980x10980 pixels. In our initial training of the phase one CNN, we used sub-images of 6875 x 3721 pixels extracted from 13 Sentinel-2 acquisitions. In the joint-fine tuning step, we added data from six Sentinel-2 acquisitions (one winter and one summer for each of the three glaciers). So, in total, this work used data from 13 to 19 Sentinel-2 acquisitions. Comparatively, Baumhoer et al. (2019) used 38 Sentinel-1 satellite acquisitions, Zhang et al. (2019) used 75 TerraSAR-X acquisitions, Mohajerani et al. (2019) used 123 Landsat 5-8 acquisitions, and Cheng et al. (2021) used 1,872 images (1,541 from Landsat and 232 from Sentinel-1). So, overall, we argue that our results were obtained with less training data than those from comparator FCN-facing works.

4.4.3 Size of input imagery

The size of input imagery also represents an area where CSC has advantages over FCNs. In FCN architectures, the instance that must be classified must be well framed in the input image. Often in the case of higher resolution images where such framing would lead to image sizes in excess of 1000x1000 pixels, downsampling must be used unless extremely powerfully GPU are available. Another important point is that the pre-processing methods used in FCN papers start with a user actually knowing where the feature of interest is and performing a suitable clip of the data. For example, Mohajerani et al. (2019) crops imagery to within a 300 m buffer area of a pre-defined calving front and further crops training images to 150x240 pixels for FCN training inputs. In the resulting images, the calving front must be kept within the frame. This type of pre-processing is not required in CSC. Instead, CSC can process entire tiles of Sentinel-2 data at native resolutions without the need for downsampling, selection and clipping of a known target area, or extensive pre-processing (see Fig. 13). In order to produce digitised validation labels for a test dataset spanning seasonally variable imagery, our test areas were cropped to 2000 to 3000 pixels (digitisation of entire Sentinel-2 tiles to near pixel-levels of detail for seasonally variable test imagery would be a more onerous task), but the CSC method is not sensitive to where the data clip boundaries fall, and it performs well even when an image boundary cuts a glacier in half. It also works well when the user does not have previous knowledge of the location of a feature of interest. Admittedly, in the case of glaciers, this is arguably less important because we already have high quality glacier inventories. However, in terms of the wider scope of image classification in EO, there are many cases where a human user cannot be expected to know *a priori* the location of all features/class instances of interest in order to carry out the level of

pre-processing required by FCN architectures. In these cases, the lower levels of pre-processing required by CSC are advantageous and has allowed us to produce classifications for full Sentinel-2 tiles (Fig. 13) that are absent from other works based on FCNs and U-Nets.

4.4.4 Local textures vs object shapes

Finally, from a theoretical perspective, FCN architectures can be strongly dependent on object shapes and less dependent on inner textures. In the final stages of the encoder part of an FCN architecture, the simplified shape of the object will contribute to the weights learned in training (as will inter class relations). This means that an FCN must be trained to recognise specific shapes. As a result, an FCN trained only on data from Helheim could not be expected to perform well at the task of classifying Jakobshavn. There are no published examples where an FCN has been trained on a single glacier and displays transferability to very different glaciers. For example, Mohajerani et al. (2019) train their FCN on three glaciers (Jakobshavn, Sverdrup, and Kangerlussuaq) and only test it on Helheim Glacier. Similarly, the FCN used by Zhang et al. (2019) is only trained and tested on Jakobshavn, providing no test of spatial transferability. Instead, multiple sites must be included in FCN training in order to reach good transferability (e.g., Cheng et al., 2021). Contrastingly, in this study, even before the application of Joint fine-tuning, the phase one VGG16 CNN solely trained on data from Helheim successfully classified large areas of Jakobshavn leading to very high performance with final, phase two results with F1s in excess of 95%. This is because CSC is driven by spectral and textural properties within the object, whilst the downsampling often required in an FCN pipeline can remove local textures. FCNs compensate for this by making use of inter-class relations, which CSC does not consider. However, on the terrestrial surface, there is a strong correlation between the ontology of a semantic class and both colour and textural properties. This explains why a statistical learning algorithm such as maximum likelihood has been used with reasonable success by the EO community for nearly half a century (Lillesand and Kiefer, 1994). Furthermore, the learning of shapes, a strong point of FCN, is not so relevant in EO since many semantic classes have either variable shapes or no shapes at all. Good examples are forest and vegetated patches, water body shapes (including supraglacial lakes), rocky outcrop shapes, and sediment patches in rivers.

Overall, the empirical results presented here show that CSC has delivered a state-of-the-art performance for novel multi-class pixel-level classification of marine-terminating glacial landscapes in Greenland. In summary, when compared to FCN architectures, CSC has lower training data volume requirements and simpler pre-processing steps. Also, the workflow produces marginally better F1 scores but marginally poorer calving front detections (in terms of pixel dimensions). On balance, we argue that this shows that there is still a place in EO for patch-based classification methods such as CSC.

5 Conclusion

We develop and evaluate a workflow for novel multi-class image classification of seasonally variable marine-terminating outlet glacier scenes using deep learning. The development of deep learning methods for automated classification of outlet glaciers is an important step towards monitoring processes at high temporal and spatial resolution (e.g., changes in frontal position, mélange extent, and calving events). While still in its infancy in glacial settings, image classification using deep learning provides clear potential to reduce the labour-intensive nature of manual methods and facilitate automated analysis in an era of the burgeoning availability of satellite imagery. Our two-phase workflow, termed CNN-Supervised Classification, is adapted for classification of medium resolution Sentinel-2 imagery of outlet glaciers in Greenland. In phase one, the application of a well-established, pre-trained CNN called VGG16 replicates the way a human operator would interpret an image, rapidly producing tiled training data for a pixel-level phase two model. Application of the phase two model produces pixel-level classifications according to seven semantic classes characteristic of complex outlet glacier settings in Greenland.

Alongside an evaluation of various input parameters and training methods on model performance, we apply and test the workflow on 27 seasonally variable unseen images. The test dataset is composed of nine images from the training area of Helheim Glacier (in-sample), and 18 images from Jakobshavn and Store glaciers which represent landscapes not previously seen by the phase one CNN during training (out-of-sample). Resulting pixel-level classifications from the test dataset as a whole produce F1 scores up to 94% for in-sample test data and 96% for out-of-sample data with the implementation of a joint fine-tuning technique. The calving front detection method built into the CSC workflow predicts fronts with a mean error of 56.17 m (5.6 pixels) and median error of 24.7 m (2.5 pixels) when optimal CSC input parameters are used. Overall, this demonstrates that the CSC workflow has good spatial and temporal transferability to unseen marine-terminating glaciers in Greenland and can be used to classify entire landscapes and subsequently produce accurate secondary datasets (such as calving front data). The simplicity of data pre-processing and the low computational costs of CSC make it a useful tool which can be accessed and used without having specialised knowledge of deep learning or the need for time-consuming generation of substantial new training data. From a wider perspective, the results of this study strengthen the foothold of deep learning in the realm of automated processing of freely available medium resolution satellite imagery, especially building on the growing body of research using deep learning in glaciology (Baumhoer et al., 2019; Mohajerani et al., 2019; Zhang et al., 2019; Xie et al., 2020; Cheng et al., 2021).

Code and data availability: Sentinel-2 imagery is available from the Copernicus Open Access Hub (available at: https://scihub.copernicus.eu/dhus/#/home, last accessed: 20/07/20). The Python scripts for the full deep learning workflow and instructions on how to apply them are available at: http://doi.org/10.5281/zenodo.4081095 and can be cited as Carbonneau and Marochov (2020). The pre-trained CNN for phase one of CSC are available for download from this institutional repository:

http://doi.org/10.5281/zenodo.4081095 and can be cited as Marochov and Carbonneau (2020). The original code for the CSC workflow for classification of fluvial scenes is available at: https://github.com/geojames/CNN-Supervised-Classification.

Supplement: The supplement includes the full list of Sentinel-2 imagery used for training and testing the classification workflow (Table S1), a flow chart of the methodology used to produce calving fronts from pixel-level classifications (Fig. S1) and the confusion matrices for all three glaciers using the optimal classification parameter set with Single CNN training (Fig. S2).

Author contributions: PC developed the code with contributions and editing by MM. MM created training and validation data, implemented the code to perform image classifications and wrote the manuscript. CRS and PC supervised, discussed results and edited the manuscript.

Competing interests: The authors declare no conflict of interest.

Acknowledgements: We acknowledge the European Union Copernicus program for providing Sentinel-2 data.

References

Alifu, H., Tateishi, R. and Johnson, B.: A new band ratio technique for mapping debris-covered glaciers using Landsat imagery and a digital elevation model, International Journal of Remote Sensing, 36(8), 2063–2075, doi:10.1080/2150704X.2015.1034886, 2015.

Amundson, J. M., Fahnestock, M., Truffer, M., Brown, J., Lüthi, M. P. and Motyka, R. J.: Ice mélange dynamics and implications for terminus stability, Jakobshavn Isbræ, Greenland, Journal of Geophysical Research: Earth Surface, 115(F1), doi:10.1029/2009JF001405, 2010.

Amundson, J. M., Kienholz, C., Hager, A. O., Jackson, R. H., Motyka, R. J., Nash, J. D. and Sutherland, D. A.: Formation, flow and break-up of ephemeral ice mélange at LeConte Glacier and Bay, Alaska., Journal of Glaciology [online] Available from: https://scholarworks.alaska.edu/handle/11122/11343 (Accessed 1 December 2020), 2020.

Andresen, C. S., Straneo, F., Ribergaard, M. H., Bjørk, A. A., Andersen, T. J., Kuijpers, A., Nørgaard-Pedersen, N., Kjær, K. H., Schjøth, F., Weckström, K. and Ahlstrøm, A. P.: Rapid response of Helheim Glacier in Greenland to climate variability over the past century, Nature Geosci, 5(1), 37–41, doi:10.1038/ngeo1349, 2012.

Andresen, C. S., Sicre, M.-A., Straneo, F., Sutherland, D. A., Schmith, T., Hvid Ribergaard, M., Kuijpers, A. and Lloyd, J. M.: A 100-year long record of alkenone-derived SST changes by Southeast Greenland, Continental Shelf Research, 71, 45–51, doi:10.1016/j.csr.2013.10.003, 2013.

Baumhoer, C. A., Dietz, A. J., Kneisel, C. and Kuenzer, C.: Automated extraction of antarctic glacier and ice shelf fronts from Sentinel-1 imagery using deep learning, Remote Sensing, 11(21), 2529, doi:10.3390/rs11212529, 2019.

Berberoglu, S., Lloyd, C. D., Atkinson, P. M. and Curran, P. J.: The integration of spectral and textural information using neural networks for land cover mapping in the Mediterranean, Computers & Geosciences, 26(4), 385–396, doi:10.1016/S0098-3004(99)00119-3, 2000.

Bevan, S. L., Luckman, A. J. and Murray, T.: Glacier dynamics over the last quarter of a century at Helheim, Kangerdlugssuaq and 14 other major Greenland outlet glaciers, The Cryosphere, 6(5), 923–937, doi:10.5194/tc-6-923-2012, 2012.

Bevan, S. L., Luckman, A. J., Benn, D. I., Cowton, T. and Todd, J.: Impact of warming shelf waters on ice mélange and terminus retreat at a large SE Greenland glacier, The Cryosphere, 13(9), 2303–2315, doi:10.5194/tc-13-2303-2019, 2019.

Blaschke, T., Lang, S., Lorup, E., Strobl, J. and Zeil, P.: Object-Oriented Image Processing in an Integrated GIS/Remote Sensing Environment and Perspectives for Environmental Applications, 16, 2000.

Bolch, T., Menounos, B. and Wheate, R.: Landsat-based inventory of glaciers in western Canada, 1985–2005, Remote Sensing of Environment, 114(1), 127–137, doi:10.1016/j.rse.2009.08.015, 2010.

Brough, S., Carr, J. R., Ross, N. and Lea, J. M.: Exceptional retreat of Kangerlussuaq Glacier, East Greenland, between 2016 and 2018, Front. Earth Sci., 7, doi:10.3389/feart.2019.00123, 2019.

Bunce, C., Carr, J. R., Nienow, P. W., Ross, N. and Killick, R.: Ice front change of marine-terminating outlet glaciers in northwest and southeast Greenland during the 21st century, J. Glaciol., 64(246), 523–535, doi:10.1017/jog.2018.44, 2018.

Carbonneau, P. E. and Marochov, M.: SEE_ICE: Glacial Landscape Classification with Deep Learning, Zenodo., doi: 10.5281/zenodo.4081095, 2020.

Carbonneau, P. E., Dugdale, S. J., Breckon, T. P., Dietrich, J. T., Fonstad, M. A., Miyamoto, H. and Woodget, A. S.: Adopting deep learning methods for airborne RGB fluvial scene classification, Remote Sensing of Environment, 251, 112107, doi:10.1016/j.rse.2020.112107, 2020a.

Carbonneau, P. E., Belletti, B., Micotti, M., Lastoria, B., Casaioli, M., Mariani, S., Marchetti, G. and Bizzi, S.: UAV-based training for fully fuzzy classification of Sentinel-2 fluvial scenes, Earth Surface Processes and Landforms, doi:10.1002/esp.4955, 2020b.

Carr, J. R., Stokes, C. R. and Vieli, A.: Threefold increase in marine-terminating outlet glacier retreat rates across the Atlantic Arctic: 1992–2010, Annals of Glaciology, 58(74), 72–91, doi:10.1017/aog.2017.3, 2017.

Carroll, D., Sutherland, D. A., Hudson, B., Moon, T., Catania, G. A., Shroyer, E. L., Nash, J. D., Bartholomaus, T. C., Felikson, D., Stearns, L. A., Noël, B. P. Y. and Broeke, M. R. van den: The impact of glacier geometry on meltwater plume structure and submarine melt in Greenland fjords, Geophysical Research Letters, 43(18), 9739–9748, doi:10.1002/2016GL070170, 2016.

Cassotto, R., Fahnestock, M., Amundson, J. M., Truffer, M. and Joughin, I.: Seasonal and interannual variations in ice melange and its impact on terminus stability, Jakobshavn Isbræ, Greenland, Journal of Glaciology, 61(225), 76–88, doi:10.3189/2015JoG13J235, 2015.

Catania, G. A., Stearns, L. A., Sutherland, D. A., Fried, M. J., Bartholomaus, T. C., Morlighem, M., Shroyer, E. and Nash, J.: Geometric controls on tidewater glacier retreat in Central Western Greenland, Journal of Geophysical Research: Earth Surface, 123(8), 2024–2038, doi:10.1029/2017JF004499, 2018.

Catania, G. A., Stearns, L. A., Moon, T. A., Enderlin, E. M. and Jackson, R. H.: Future evolution of Greenland's marine-terminating outlet glaciers, Journal of Geophysical Research: Earth Surface, 125(2), e2018JF004873, doi:10.1029/2018JF004873, 2020.

Chauché, N., Hubbard, A., Gascard, J.-C., Box, J. E., Bates, R., Koppes, M., Sole, A., Christoffersen, P. and Patton, H.: Ice–ocean interaction and calving front morphology at two west Greenland tidewater outlet glaciers, The Cryosphere, 8(4), 1457–1468, doi:10.5194/tc-8-1457-2014, 2014.

Chen, G. and Hong Yang, Y. H.: Edge detection by regularized cubic B-spline fitting, IEEE Transactions on Systems, Man, and Cybernetics, 25(4), 636–643, doi:10.1109/21.370194, 1995.

Cheng, D., Hayes, W., Larour, E., Mohajerani, Y., Wood, M., Velicogna, I. and Rignot, E.: Calving Front Machine (CALFIN): glacial termini dataset and automated deep learning extraction method for Greenland, 1972–2019, The Cryosphere, 15(3), 1663–1675, doi:10.5194/tc-15-1663-2021, 2021.

Chollet, F.: Deep learning with Python, Manning Publications Co, Shelter Island, New York., 2017.

Christoffersen, P., O'Leary, M., Van Angelen, J. H. and Van Den Broeke, M.: Partitioning effects from ocean and atmosphere on the calving stability of Kangerdlugssuaq Glacier, East Greenland, Ann. Glaciol., 53(60), 249–256, doi:10.3189/2012AoG60A087, 2012.

Cook, A. J., Copland, L., Noël, B. P. Y., Stokes, C. R., Bentley, M. J., Sharp, M. J., Bingham, R. G. and Broeke, M. R. van den: Atmospheric forcing of rapid marine-terminating glacier retreat in the Canadian Arctic Archipelago, Science Advances, 5(3), doi:10.1126/sciadv.aau8507, 2019.

Csatho, B. M., Schenk, A. F., van der Veen, C. J., Babonis, G., Duncan, K., Rezvanbehbahani, S., van den Broeke, M. R., Simonsen, S. B., Nagarajan, S. and van Angelen, J. H.: Laser altimetry reveals complex pattern of Greenland Ice Sheet dynamics, Proc Natl Acad Sci U S A, 111(52), 18478–18483, doi:10.1073/pnas.1411680112, 2014.

Enderlin, E. M., Howat, I. M., Jeong, S., Noh, M.-J., Angelen, J. H. van and Broeke, M. R. van den: An improved mass budget for the Greenland ice sheet, Geophysical Research Letters, 41(3), 866–872, doi:10.1002/2013GL059010, 2014.

Everett, A., Kohler, J., Sundfjord, A., Kovacs, K. M., Torsvik, T., Pramanik, A., Boehme, L. and Lydersen, C.: Subglacial discharge plume behaviour revealed by CTD-instrumented ringed seals, Scientific Reports, 8(1), 13467, doi:10.1038/s41598-018-31875-8, 2018.

Foga, S., Stearns, L. A. and van der Veen, C. J.: Application of satellite remote sensing techniques to quantify terminus and ice mélange behavior at Helheim Glacier, East Greenland, Marine Technology Society Journal, 48(5), 81–91, doi:10.4031/MTSJ.48.5.3, 2014.

Frey, H., Paul, F. and Strozzi, T.: Compilation of a glacier inventory for the western Himalayas from satellite data: methods, challenges, and results, Remote Sensing of Environment, 124, 832–843, doi:10.1016/j.rse.2012.06.020, 2012.

Gerrish, Laura: The coastline of Kalaallit Nunaat/ Greenland available as a shapefile and geopackage, covering the main land and islands, with glacier fronts updated as of 2017., 2 files, 5.26 MB, doi:10.5285/8CECDE06-8474-4B58-A9CB-B820FA4C9429, 2020.

Goodfellow, I., Bengio, Y. and Courville, A.: Deep Learning, MIT Press. [online] Available from: https://www.deeplearningbook.org/ (Accessed 22 July 2020), 2016.

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

Guo, W., Liu, S., Xu, J., Wu, L., Shangguan, D., Yao, X., Wei, J., Bao, W., Yu, P., Liu, Q. and Jiang, Z.: The second Chinese glacier inventory: data, methods and results, Journal of Glaciology, 61(226), 357–372, doi:10.3189/2015JoG14J209, 2015.

Hill, E. A., Carr, J. R. and Stokes, C. R.: A review of recent changes in major marine-terminating outlet glaciers in Northern Greenland, Front. Earth Sci., 4, doi:10.3389/feart.2016.00111, 2017.

Hochreuther, P., Neckel, N., Reimann, N., Humbert, A. and Braun, M.: Fully automated detection of supraglacial lake area for Northeast Greenland using Sentinel-2 time-series, Remote Sensing, 13(2), 205, doi:10.3390/rs13020205, 2021.

Hoeser, T., Bachofer, F. and Kuenzer, C.: Object detection and image segmentation with deep learning on earth observation data: a review—part ii: applications, Remote Sensing, 12(18), 3053, doi:10.3390/rs12183053, 2020.

How, P., Benn, D. I., Hulton, N. R. J., Hubbard, B., Luckman, A., Sevestre, H., van Pelt, W. J. J., Lindbäck, K., Kohler, J. and Boot, W.: Rapidly changing subglacial hydrological pathways at a tidewater glacier revealed through simultaneous observations of water pressure, supraglacial lakes, meltwater plumes and surface velocities, The Cryosphere, 11(6), 2691–2710, doi:10.5194/tc-11-2691-2017, 2017.

Howat, I. M., Joughin, I. and Scambos, T. A.: Rapid changes in ice discharge from Greenland outlet glaciers, Science, 315(5818), 1559–1561, doi:10.1126/science.1138478, 2007.

Howat, I. M., Ahn, Y., Joughin, I., Broeke, M. R. van den, Lenaerts, J. T. M. and Smith, B.: Mass balance of Greenland's three largest outlet glaciers, 2000–2010, Geophysical Research Letters, 38(12), doi:10.1029/2011GL047565, 2011.

Johnson, J. M. and Khoshgoftaar, T. M.: Survey on deep learning with class imbalance, J Big Data, 6(1), 27, doi:10.1186/s40537-019-0192-5, 2019.

Joughin, I., Howat, I., Alley, R. B., Ekstrom, G., Fahnestock, M., Moon, T., Nettles, M., Truffer, M. and Tsai, V. C.: Ice-front variation and tidewater behavior on Helheim and Kangerdlugssuaq Glaciers, Greenland, Journal of Geophysical Research: Earth Surface, 113(F1), doi:10.1029/2007JF000837, 2008a.

Joughin, I., Howat, I. M., Fahnestock, M., Smith, B., Krabill, W., Alley, R. B., Stern, H. and Truffer, M.: Continued evolution of Jakobshavn Isbrae following its rapid speedup, Journal of Geophysical Research: Earth Surface, 113(F4), doi:10.1029/2008JF001023, 2008b.

Juan, J. de, Elósegui, P., Nettles, M., Larsen, T. B., Davis, J. L., Hamilton, G. S., Stearns, L. A., Andersen, M. L., Ekström, G., Ahlstrøm, A. P., Stenseng, L., Khan, S. A. and Forsberg, R.: Sudden increase in tidal response linked to calving and acceleration at a large Greenland outlet glacier, Geophysical Research Letters, 37(12), doi:10.1029/2010GL043289, 2010.

King, M. D., Howat, I. M., Jeong, S., Noh, M. J., Wouters, B., Noël, B. and Broeke, M. R. van den: Seasonal to decadal variability in ice discharge from the Greenland Ice Sheet, The Cryosphere, 12(12), 3813–3825, doi:10.5194/tc-12-3813-2018, 2018.

King, M. D., Howat, I. M., Candela, S. G., Noh, M. J., Jeong, S., Noël, B. P. Y., van den Broeke, M. R., Wouters, B. and Negrete, A.: Dynamic ice loss from the Greenland Ice Sheet driven by sustained glacier retreat, Communications Earth & Environment, 1(1), 1–7, doi:10.1038/s43247-020-0001-2, 2020.

Kingma, D. P. and Ba, J.: Adam: A method for Stochastic Optimization, arXiv:1412.6980 [cs] [online] Available from: http://arxiv.org/abs/1412.6980 (Accessed 23 August 2020), 2017.

Krieger, L. and Floricioiu, D.: Automatic glacier calving front delineation on TerraSAR-X and Sentinel-1 SAR imagery, in 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 2817–2820., 2017.

Lea, J. M.: The Google Earth Engine Digitisation Tool (GEEDiT) and the Margin change Quantification Tool (MaQiT) – simple tools for the rapid mapping and quantification of changing Earth surface margins, Earth Surf. Dynam., 6(3), 551–561, doi:10.5194/esurf-6-551-2018, 2018.

LeCun, Y., Bengio, Y. and Hinton, G.: Deep learning, Nature, 521(7553), 436–444, doi:10.1038/nature14539, 2015.

Li, X., Myint, S. W., Zhang, Y., Galletti, C., Zhang, X. and Turner, B. L.: Object-based land-cover classification for metropolitan Phoenix, Arizona, using aerial photography, International Journal of Applied Earth Observation and Geoinformation, 33, 321–330, doi:10.1016/j.jag.2014.04.018, 2014.

Lillesand, T. M., and Kiefer, R.W.: Remote sensing and image interpretation, 3rd ed., Wiley & Sons, New York., 1994.

Liu, H. and Jezek, K. C.: A complete high-resolution coastline of Antarctica extracted from orthorectified Radarsat SAR imagery, Photogramm Eng Remote Sensing, 70(5), 605–616, doi:10.14358/PERS.70.5.605, 2004.

Liu, X., Deng, Z. and Yang, Y.: Recent progress in semantic image segmentation, Artif Intell Rev, 52(2), 1089–1106, doi:10.1007/s10462-018-9641-3, 2019.

Maggiori, E., Tarabalka, Y., Charpiat, G. and Alliez, P.: Fully convolutional neural networks for remote sensing image classification, in 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 5071–5074., 2016.

Marochov, M., and Carbonneau, P.: Image classification of marine-terminating outlet glaciers using deep learning methods: pre-trained models [dataset], doi:10.5281/zenodo.4081095, 2020.

Miles, B. W. J., Stokes, C. R. and Jamieson, S. S. R.: Pan–ice-sheet glacier terminus change in East Antarctica reveals sensitivity of Wilkes Land to sea-ice changes, Science Advances, 2(5), e1501350, doi:10.1126/sciadv.1501350, 2016.

Miles, B. W. J., Stokes, C. R. and Jamieson, S. S. R.: Velocity increases at Cook Glacier, East Antarctica, linked to ice shelf loss and a subglacial flood event, The Cryosphere, 12(10), 3123–3136, doi:10.5194/tc-12-3123-2018, 2018.

Mohajerani, Y., Wood, M., Velicogna, I. and Rignot, E.: Detection of glacier calving margins with Convolutional Neural Networks: A case study, Remote Sensing, 11(1), 74, doi:10.3390/rs11010074, 2019.

Moon, T. and Joughin, I.: Changes in ice front position on Greenland's outlet glaciers from 1992 to 2007, Journal of Geophysical Research: Earth Surface, 113(F2), doi:10.1029/2007JF000927, 2008.

Mouginot, J., Rignot, E., Bjørk, A. A., Broeke, M. van den, Millan, R., Morlighem, M., Noël, B., Scheuchl, B. and Wood, M.: Forty-six years of Greenland Ice Sheet mass balance from 1972 to 2018, PNAS, 116(19), 9239–9244, doi:10.1073/pnas.1904242116, 2019.

Nijhawan, R., Das, J. and Raman, B.: A hybrid of deep learning and hand-crafted features based approach for snow cover mapping, International Journal of Remote Sensing, 40(2), 759–773, doi:10.1080/01431161.2018.1519277, 2019.

Noël, B., Berg, W. J. van de, Lhermitte, S. and Broeke, M. R. van den: Rapid ablation zone expansion amplifies north Greenland mass loss, Science Advances, 5(9), doi:10.1126/sciadv.aaw0123, 2019.

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

Rastner, P., Bolch, T., Mölg, N., Machguth, H., Le Bris, R. and Paul, F.: The first complete inventory of the local glaciers and ice caps on Greenland, The Cryosphere, 6(6), 1483–1495, doi:10.5194/tc-6-1483-2012, 2012.

Rignot, E. and Kanagaratnam, P.: Changes in the velocity structure of the Greenland Ice Sheet., Science, 311(5763), 986–990, doi:10.1126/science.1121381, 2006.

Robson, B. A., Nuth, C., Dahl, S. O., Hölbling, D., Strozzi, T. and Nielsen, P. R.: Automated classification of debris-covered glaciers combining optical, SAR and topographic data in an object-based environment, Remote Sensing of Environment, 170, 372–387, doi:10.1016/j.rse.2015.10.001, 2015.

Robson, B. A., Bolch, T., MacDonell, S., Hölbling, D., Rastner, P. and Schaffer, N.: Automated detection of rock glaciers using deep learning and object-based image analysis, Remote Sensing of Environment, 250, 112033, doi:10.1016/j.rse.2020.112033, 2020.

Ronneberger, O., Fischer, P. and Brox, T.: U-Net: Convolutional Networks for biomedical image segmentation, in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, edited by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, pp. 234–241, Springer International Publishing, Cham., 2015.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: Learning internal representations by error propagation, 23, 1986.

Samarth, G. C., Bhowmik, N. and Breckon, T. P.: Experimental exploration of Compact Convolutional Neural Network architectures for non-temporal real-time fire detection, arXiv:1911.09010 [cs, eess] [online] Available from: http://arxiv.org/abs/1911.09010 (Accessed 23 August 2020), 2019.

Seale, A., Christoffersen, P., Mugford, R. I. and O'Leary, M.: Ocean forcing of the Greenland Ice Sheet: calving fronts and patterns of retreat identified by automatic satellite monitoring of eastern outlet glaciers, Journal of Geophysical Research: Earth Surface, 116(F3), doi:10.1029/2010JF001847, 2011.

Sharma, A., Liu, X., Yang, X. and Shi, D.: A patch-based convolutional neural network for remote sensing image classification, Neural Networks, 95, 19–28, doi:10.1016/j.neunet.2017.07.017, 2017.

Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv:1409.1556 [cs] [online] Available from: http://arxiv.org/abs/1409.1556 (Accessed 21 July 2020), 2015.

Sohn, H.-G. and Jezek, K. C.: Mapping ice sheet margins from ERS-1 SAR and SPOT imagery, International Journal of Remote Sensing, 20(15–16), 3201–3216, doi:10.1080/014311699211705, 1999.

Stokes, C. R., Andreassen, L. M., Champion, M. R. and Corner, G. D.: Widespread and accelerating glacier retreat on the Lyngen Peninsula, northern Norway, since their 'Little Ice Age' maximum, Journal of Glaciology, 64(243), 100–118, doi:10.1017/jog.2018.3, 2018.

Straneo, F., Curry, R. G., Sutherland, D. A., Hamilton, G. S., Cenedese, C., Våge, K. and Stearns, L. A.: Impact of fjord dynamics and glacial runoff on the circulation near Helheim Glacier, Nature Geosci, 4(5), 322–327, doi:10.1038/ngeo1109, 2011.

Straneo, F., Hamilton, G. S., Stearns, L. A. and Sutherland, D. A.: Connecting the Greenland Ice Sheet and the Ocean: a case study of Helheim Glacier and Sermilik fjord, Oceanography, 29(4), 34–45, 2016.

Sutherland, D. A., Jackson, R. H., Kienholz, C., Amundson, J. M., Dryer, W. P., Duncan, D., Eidam, E. F., Motyka, R. J. and Nash, J. D.: Direct observations of submarine melt and subsurface geometry at a tidewater glacier, Science, 365(6451), 369–374, doi:10.1126/science.aax3528, 2019.

Tuckett, P. A., Ely, J. C., Sole, A. J., Livingstone, S. J., Davison, B. J., Melchior van Wessem, J. and Howard, J.: Rapid accelerations of Antarctic Peninsula outlet glaciers driven by surface melt, Nature Communications, 10(1), 4311, doi:10.1038/s41467-019-12039-2, 2019.

Vaughan, D. G., Comiso, J. C., Allison, I., Carrasco, J., Kaser, G., Kwok, R., Mote, P., Murray, T., Paul, F., Ren, J., Rignot, E., Solomina, O., Zhang, T., Arendt, A. A., Bahr, D. B., Cogley, J. G., Gardner, A. S., Gerland, S., Gruber, S., Haas, C., Hagen, J. O., Hock, R., Holland, D., Huss, M., Markus, T., Marzeion, B., Massom, R., Moholdt, G., Overduin, P. P., Payne, A., Pfeffer, W. T., Prowse, T., Radić, V., Robinson, D., Sharp, M., Shiklomanov, N., Stammerjohn, S., Velicogna, I., Wadhams, P., Worby, A., Zhao, L., Bamber, J., Huybrechts, P. and Lemke, P.: 4 Observations: Cryosphere, 66, 2013.

Wood, M., Rignot, E., Fenty, I., Menemenlis, D., Millan, R., Morlighem, M., Mouginot, J. and Seroussi, H.: Ocean-induced melt triggers glacier retreat in Northwest Greenland, Geophysical Research Letters, 45(16), 8334–8342, doi:10.1029/2018GL078024, 2018.

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

Yu, Y., Zhang, Z., Shokr, M., Hui, F., Cheng, X., Chi, Z., Heil, P. and Chen, Z.: automatically extracted Antarctic coastline using remotely-sensed data: an update, Remote Sensing, 11(16), 1844, doi:10.3390/rs11161844, 2019.

Yuan, J., Chi, Z., Cheng, X., Zhang, T., Li, T. and Chen, Z.: Automatic extraction of supraglacial lakes in Southwest Greenland during the 2014–2018 melt seasons based on Convolutional Neural Network, Water, 12(3), 891, doi:10.3390/w12030891, 2020.

Zhang, E., Liu, L. and Huang, L.: Automatically delineating the calving front of Jakobshavn Isbræ from multitemporal TerraSAR-X images: a deep learning approach, The Cryosphere, 13(6), 1729–1741, doi:10.5194/tc-13-1729-2019, 2019.